Department of Interdisciplinary Life Sciences
University of Veterinary Medicine Vienna

Konrad Lorenz Institute of Ethology
(Head: Prof Dr Leonida Fusani)

Modulation and consistency of ultrasonic vocalizations in the Eastern European House Mouse (*Mus musculus musculus*): functions and translational value

Ph.D. thesis submitted for the fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY (PhD)

University of Veterinary Medicine Vienna

submitted by
Maria Adelaide Marconi

Vienna, November 2023

Ph.D. ADVISORY COMMITTEE

1st PhD advisor
Priv. Doz. Dr. Dustin J Penn
Konrad Lorenz Institute of Ethology
Department of Interdisciplinary Life Sciences
University of Veterinary Medicine of Vienna
Dustin.Penn@vetmeduni.ac.at

2nd PhD advisor
Priv. Doz. Dr. Peter Balazs
Acoustics Research Institute (ARI)
Austrian Academy of Science
Peter.Balazs@oeaw.ac.at

LIST OF ARTICLES INCLUDED IN THE Ph.D. THESIS:

1. Luzynski, Kenneth C.; Nicolakis, Doris; Marconi, Maria Adelaide; Zala, Sarah M.; Kwak, Jae; Penn, Dustin J., Pheromones that correlate with reproductive success in competitive conditions.

2. Zala, Sarah M.; Nicolakis, Doris; Marconi, Maria Adelaide; Noll, Anton; Ruf Thomas; Balazs, Peter; Penn, Dustin J.
Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female.

3. Nicolakis, Doris; Marconi, Maria Adelaide; Zala, Sarah, M.; Penn, Dustin J.
Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success.

4. Marconi, Maria Adelaide; Nicolakis, Doris; Abbasi, Reyhaneh; Zala, Sarah, M.; Penn, Dustin J.
Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures.

5. Caruso, Angela; Marconi, Maria Adelaide; Scattoni, Maria Luisa; Ricceri, Laura
Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences.

6. Abbasi, Reyhaneh; Balazs, Peter; Marconi, Maria Adelaide; Nicolakis, Doris; Zala, Sarah M.; Penn, Dustin J.Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap).

TABLE OF CONTENTS

1. ACKNOWLEDGEMENTS 1
2. CONTRIBUTIONS BY CHAPTER 2
3. DECLARATION 3
4. SUMMARY 4
5. GENERAL INTRODUCTION 9
5.1 Wild Mus musculus and laboratory mice 9
5.2 Chemical signals 11
5.3 Acoustic signals 12
5.4 Research objectives 15
References 21
CHAPTER 1: Pheromones that correlate with reproductive success in competitive conditions 28
CHAPTER 2: Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female 45
CHAPTER 3: Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success 74
CHAPTER 4: Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures 94
CHAPTER 5: Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences 124
CHAPTER 6: Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap) 141
6. DISCUSSION AND CONCLUSIONS 173
References 181

1. ACKNOWLEDGEMENTS

My experience in Vienna and this PhD helped me to grow as a person, to find outstanding colleagues and supervisors, and to feel at home even though I was far away.

I would like to thank my supervisors, Professor Dustin Penn and Dr. Sarah Zala, who have guided me on this journey and shown me how science can broaden horizons, it can create friendships and collaborations through exciting research environments and also throw up difficult challenges to overcome. I thank Professor Peter Balazs for supporting me as a second advisor and Dr. Anton Noll from the Acoustic Research Institute for giving valuable insights on the study of the apparatuses to analyze ultrasounds and the complexity of their software, and the FWF fund for the material and financial support. I thank Reyhaneh Abbasi for patiently and repeatedly explaining to me all the steps necessary to apply complicated algorithms.

A special thanks to the collaborators who gave important contributions to our studies: Dr. Jae Kwak, Bogi Balint, Dr. Steve Smith, Dr. Angela Caruso, Dr. Laura Ricceri, and Dr. Maria Luisa Scattoni, my outstanding Italian mentor from the beginning of the academic studies.

I am grateful to my colleague and friend Doris Nicolakis, who has been a key reference point for me in my adventure in Vienna. Doris has been my peer colleague in this academic and life experience and I feel truly blessed to have shared this part of my human and academic journey with her. I would also like to thank our TAs and friends Bettina Wernisch and Inka (our office dog), Teresa Klaus, and Eva Straßer with whom we collaborated on many projects, and spent lovely evenings after work.

Special thanks also to our senior colleagues who gave valuable suggestions when I had doubts and encouraged me in difficult moments: Dr. Kerstin Auer, Dr. Michaela Thoß, and Dr. Ken Luzynski. KLIVV and FIWI have been important places to exchange ideas, feedback, and time with friends and outstanding colleagues.

I thank all the “Glasshouse” people and collaborators who were involved in the PhD experience: Heidi and Roland Sasse, Renate Hengsberger, Tanja Szabo, Dr. Amélie Desvars-Larrive, Dr. Filipa Cunha Saraiva, Dr. Sarah Lado, Dr. Franz Hölzl, Dr. Jessica Cornils, Corinna Schüssele, Diana Gliga, Manuela Grijincu, Sophie Dumfart, Natalie Fuhrmann, Anita Pocsai, Bora Kim, Dr. Sylvain Giroud, Dr. Ivan Maggini, Dr. Sara Lupi, Dr, Andrea Ferretti, Dr. Judith Janisch, and Dr. Elisa Perinot, and many other students and researchers that I met.

I would like to thank my fiancé Dario who has been patiently waiting for me over these years with frequent visits to Vienna and my family who supported me during these long years of study always spurring me to learn and improve.

2. CONTRIBUTIONS BY CHAPTER

1. Pheromones that correlate with reproductive success in competitive conditions.
Luzynski, Kenneth C.; Nicolakis, Doris; Marconi, Maria Adelaide; Zala, Sarah M.; Kwak,Jae; Penn, Dustin J.
Scientific Reports, impact factor: 4.6; published

Experimental design	Experiments	Data analyses and software	Writing of the paper	Contribution to materials
KCL; DN; MAM; SMZ; JK; DJP	KCL; JK; DN; MAM	KCL; JK	KCL 1st Author; DJP	KCL; DN; MAM; JK; DJP

2. Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female.

Zala, Sarah M.; Nicolakis, Doris; Marconi, Maria Adelaide; Noll, Anton; Ruf Thomas; Balazs, Peter; Penn, Dustin J.
PLoS ONE, impact factor: 3.7; **published**

Experimental design	Experiments	Data analyses and software	Writing of the paper	Contribution to materials
SMZ, DN	DN	SMZ; DN; MAM; TR; AN; PB	SMZ; DJP	SMZ; DJP; DN; MAM; TR; PB; AN

3. Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success.
Nicolakis, Doris; Marconi, Maria Adelaide; Zala, Sarah, M.; Penn, Dustin J.
Frontiers in Zoology: Impact factor: 3.3; **published**

Experimental design	Experiments	Data analysis and software	Writing of the paper	Contribution to materials
DN; MAM; SMZ; DJP	DN, MAM	DN, MAM	DN, DJP; SMZ	DN; MAM; SMZ; DJP

4. Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures.
Marconi,Maria Adelaide; Nicolakis, Doris; Abbasi, Reyhaneh; Zala, Sarah, M.; Penn, Dustin J.
Animal Behaviour, impact factor: 3.04; **published**

Experimental design	Experiments	Data analyses and software	Writing of the paper	Contribution to materials
MAM; DN; SMZ; DJP	MAM; DN	MAM; DN; RA	MAM 1st Author; DJP; SMZ	MAM; DN; RA; SMZ; DJP

5. Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences.
Caruso, Angela; Marconi, Maria Adelaide; Scattoni, Maria Luisa; Ricceri, Laura
Genes, Brain & Behavior, impact factor: 3.7; **published**

Experimental design	Experiments	Data analyses and software	Writing of the paper	Contribution to materials
LR; MLS	LR; MLS; MAM	AC; MAM; LR	AC; LR; MLS	AC; MAM; MLS; LR

6. Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap).
Abbasi, Reyhaneh; Balazs, Peter; Marconi, Maria Adelaide; Nicolakis, Doris; Zala, Sarah M.; Penn, Dustin J.
PLoS Computational Biology, impact factor: 4.8; **published**

Experimental design	Experiments	Data analyses and software	Writing of the paper	Contribution to materials
RA; PB; DJP	RA; PB; MAM; DN	RA; MAM; DN; SMZ; DJP	RA; PB; SMZ; DJP	RA; PB; MAM; DN; SMZ; DJP

3. DECLARATION

I, Maria Adelaide Marconi, declare that all the experiments and studies were performed according to the rules of good Scientific Practice as indicated in the faculty guidelines and have been followed in all aspects.

4. SUMMARY

Research on the communication of house mice has largely focused on olfactory signals, whereas their acoustic signals have received relatively little attention until recent years. I investigated the expression and functions of chemosensory and acoustic signals of Eastern European house mice (*Mus musculus* subsp. *musculus*), and particularly the functions of male courtship vocalizations. Male mice emit ultrasonic vocalizations (USVs) during courtship, which potentially influence their mating and reproductive success; however, evidence for this hypothesis has been lacking. My aims were to determine the effect of male pheromones and courtship vocalizations on male reproductive success, and I focused my research on wild-derived *Mus musculus* (henceforth "wild mice"). To determine what kind of information is contained in male USVs (signaling functions), which females could use to assess the state and identity of potential mates, I also investigated whether USVs reliably signal males' sexual arousal and individual identity. Due to the enormous time and effort required for analyzing mouse vocalizations, I also collaborated on research aimed at developing better methods for automatic USV detection and classification. Below, I summarize the main aims and results from each of the chapters in my dissertation.

The first chapter provides results from a study that we conducted to test the effect of chemical signals on reproductive success of wild house mice living in seminatural conditions. We found that males increased the excretion of certain volatile and involatile pheromones only once they became socially dominant and acquired a territory, and dominant males had higher reproductive success than subordinates. Females discriminated and showed a preferential attraction towards the urinary scent of dominant versus subordinate males, but they showed no preferences for male urine with higher concentrations of involatile major urinary proteins

(MUPs). Our study provides the first results to show that pheromone production correlates with male reproductive success for any mammal. They also raise important caveats for studying the regulation and functions of the pheromones of male mice in laboratory conditions.

The second chapter presents the results from an experiment that we conducted to test whether males alter the emission of USVs after exposure to a female (i.e., socio-sexual priming). Early studies on male laboratory mice suggested that exposure to a female, or female odor enhances the rate of male USV emission, and that USVs therefore provide a reliable indicator of sexual arousal and interest in mating; however, the evidence for this hypothesis has been mixed. In this study, we implemented an improved version of a method that we developed to improve detection of USVs (Automated Mouse Ultrasound Detector or A-MUD), which is the first step during data processing. Our results show that the socio-sexual priming increased the number and the types of USVs that males emitted, and the effects were long-lasting. This study provides the first evidence for socio-sexual priming of USV emission in wild house mice and the first to show that priming enhances the diversity as well as the rate of male USV emission.

The third chapter reports the results of an experiment in which we aimed to test whether mice modulate USV production depending upon the genetic relatedness of a potential mate, and whether courtship USV emission correlates with a breeding pair's subsequent reproductive success. We found that courtship USVs are differently modulated in different consecutive social contexts (indirect and subsequent direct social contact) and that males increased the number and spectral complexity of vocalizations during direct interactions. In this latter context, USV number and modulation correlated with reproductive success and relatedness to the female partner: mating pairs of genetically unrelated individuals emitted USVs with higher duration

and spectral complexity and showed a higher reproductive success compared to pairs that were closely related, whose USVs were shorter and spectrally simple and had a lower reproductive success. Our results provide the first evidence that mice modulate USV emission during courtship depending upon genetic relatedness of a potential mate and that USV emission (rate and types) correlate with their subsequent reproductive success.

In the fourth chapter, our goal was to test the hypothesis that male courtship USVs provide cues of individual identity (individual vocal signatures). We first tested whether USVs change once males are provided with a female odor stimulus; and then we evaluated the variation in the quantity and quality of USVs within individuals versus among individuals over three consecutive weeks. The males rarely vocalized until presented with female odor, and then individual males consistently increased the number and spectral complexity of USVs. After the presentation of female odor, we found that the spectrotemporal features of male USVs differed between individuals and mice could be classified according to their "vocalizing style" with a striking difference between low versus high vocalizers. We found that some males have a distinct vocal repertoire that is repeatable over time and that some spectrotemporal features of their USV structure could potentially be used by conspecifics for individual recognition. This study provides the first evidence that male mice increase the numbers and types of USVs upon investigating female odor, and also the first evidence that male USVs contain individual signatures that are consistent over time.

In the fifth chapter, we aimed to cross-validate findings from different strains of laboratory mice, and we analyzed the USV emission and the social behavior of three strains, commonly used in studies on neurodevelopmental disorders (NDDs). Our results showed strain differences in the number and duration of USVs, the onset of the peak in the calling rate during

neonatal age, and also sex and strain differences in the number of vocalizations and repertoire composition during adulthood during same sex (i.e., female-female) and opposite-sex interactions. For the first time, we showed that: (1) there is a large phenotypic variation among strains in their social and communicative profile (i.e., number and duration of social investigations, vocalizations and spectrotemporal features of USVs); (2) strain C57/B6 emitted the lowest number of USVs that were spectrally simple and performed the lowest number of partner investigations; and (3) CD1 outbred strain showed an intermediate socio-communicative profile compared to the two inbred strains. Our results highlight the differences in the behavior and vocalizations among strains of laboratory mice and importance of choosing the appropriate strain to conduct experiments on socio-communicative profiles in mice.

In the sixth chapter, we developed a new method for automatically classifying different types of USVs, and compared its performance with other methods. Manual classification of USVs requires a tremendous amount of time, even for trained researchers, and therefore, we aimed to develop a new automatic USV classification method that is more reliable and generalizable than the current methods that are available. Our new method combines bootstrapping and ensemble machine learning. We found that it can accurately classify USVs into 12 different types (though accuracy was improved by pooling more types), and that it outperformed all other methods. This is the first study to systematically compare different automatic USV detection and classification methods and evaluate their performance, and we also provide a new method that now provides the state-of-the-art tool for USV classification.

In summary, the main goals of my research were to investigate the adaptive functions of chemical and acoustic signals of house mice in terms of their effects on reproductive success. The results of these studies provide novel evidence that male pheromones and vocalizations

influence reproductive success. To better understand why females are attracted to male USVs, and what kind of information they could provide to potential mates, I also tested hypotheses about their signaling functions. The results of these studies provide novel evidence that courtship USVs provide reliable information about a male's individual identity (individual vocal signatures) and their state (sexual arousal). We presented a novel non-invasive tool to profile animal models used for NDDs and implemented a novel detection system and classification method for analyzing USVs in both, wild and domesticated mice. Results from studies on laboratory mice have provided many important and interesting findings, but as we have shown, the results from one strain cannot necessarily be extrapolated to other strains or their wild counter parts. Therefore, in the future, more studies on wild house mice are necessary to address questions about the ecological relevance and adaptive significance of their complex chemical and acoustic signals.

5. GENERAL INTRODUCTION

My research focused on the courtship behavior of house mice (*Mus musculus),* and mainly their ultrasonic acoustic communication. Laboratory mice are the single most widely studied model species in the biomedical sciences, whereas wild house mice have received surprisingly little attention. As I focused my research on wild house mice rather than any of the standard laboratory mouse models, below I first summarize the differences in between these rodents. In the following sections, I provide background information for each chapter of my dissertation.

5.1 Wild *Mus musculus* and laboratory mice

For more than 100 years, domesticated laboratory mice have been used as animal models in biological and biomedical sciences, and they have provided the model organisms used by many scientists who have won Nobel Prizes (1). However, there are many differences between laboratory mice from their wild counterparts, including genetics, physiology, and behavior, and limitations as model organisms that are increasingly becoming appreciated.

The ancestors of modern laboratory and wild house mice lived in the Iranian plateau from where different populations became geographically and genetically distinct. This process fostered the emergence of the modern subspecies that diverged around 350,000 to 500,000 years ago, the same time as the earliest *H. sapiens* arose (2, 3). Mice spread around the planet with human migration, they were artificially selected to be companion animals from Japan to Europe, and for the past 100 years, as animal models in the laboratory (1, 4, 5). Thus, laboratory mice are hybrids; genetic mosaics of different *Mus* subspecies artificially crossed over centuries, and some have therefore proposed that they should be classified as a distinct species, *Mus laboratorius* or *Mus gemischus*, since they do not exist in nature and they show striking

differences to wild house mice (4-6). Artificial selection and inbreeding led to a reduction in genetic diversity and increased homozygosity, as inbred strains are maintained by mating siblings, and even so-called "outbred" stocks, which are bred haphazardly, lack the genetic diversity of wild populations (7). Laboratory mice suffer from a variety of problems, including obesity, early arthritis, early vision and hearing loss, reduced microbiota diversity, and reduced immune responses (8). Strains of laboratory mice are different in their fur color, body mass, and litter sizes, and they also can differ dramatically in their behavior, as some are less aggressive, some are tamer and easier to handle, some are more acrophobic than other strains (8) (personal observations).

In comparison, wild house mice have high levels of genetic diversity and individual heterozygosity, more diverse microbiota, and a more efficient immune systems (9). In the wild and seminatural conditions, house mice usually live in groups composed of a single dominant male and one or more females, or a dominant pair and subordinate females sharing the same territories. Dominant territorial males sometimes allow the presence of subordinate males, though other dominant males chase all other males away and control the resources on their territory. In captivity, wild-caught and wild-derived house mice are very active, jumpy, neophobic, and usually try to flee or bite when handled. After puberty, males became highly territorial, even with their cage-mate siblings and thus they are usually singly housed to prevent fighting. Thus, due to the many differences between laboratory mice and wild house mice, results from studies on laboratory mice should not be generalized or extrapolated to wild house mice (or even other strains of laboratory mice), and vice versa.

5.2 Chemical signals

Studies on olfaction and chemical signals have been mainly performed on laboratory strains and on the subspecies *Mus musculus domesticus.* The scent-marking behavior is part of the signaling system, which advertises the presence of a dominant adult male, or breeding pair holding a territory. There are a wide variety of assays used to study social status in laboratory mice, such as the resident-intruder paradigm in a cage that is the “gold standard" method to study the aggressive behavior (10). However, more ecologically relevant setups are needed to study both chemical communication and social hierarchy establishment, and whether urine compounds can signal the social status of individuals over time and in the presence of other interacting mice.

Several studies have concluded that the major urinary proteins (MUPs) of house mice urine control individual recognition (11). However, more recent studies on another subspecies, *Mus musculus musculus,* did not detect evidence for this hypothesis (12-14). Exposure to male urine influences the estrous cycles of females (Whitten effect) and MUPs and volatile organic compounds (VOCs) in male urine have been proposed to explain these effects (15-18) including the acceleration of puberty, promote the vaginal opening, and female olfactory preference (19, 20). Although most studies focused on laboratory strains, one recent study on wild mice showed that females accelerate the cycling when repeatedly exposed to male odor (i.e., soiled bedding) since the urinary proteins might influence the estrous cycle as previously documented in laboratory mice (21). A similar effect was found in wild males which increased faster the body mass when repeatedly exposed to female odor compared to control males (22). However, no previous studies tested whether chemical signals affect the reproductive success of both sexes.

5.3 Acoustic signals

As with olfactory studies, acoustic communication in mice has been investigated for more than 60 years in laboratory strains (23). Both sexes emit USVs and these vocalizations require close proximity to be heard by conspecifics since they are short-ranged (9, 24). Mice can emit both sonic and ultrasonic vocalizations and different mechanisms have been proposed for the sound production: an airflow passing through the larynx generates a whistle and produces a USV, whereas sonic vocalizations are produced when the vocal folds vibrate. However, sounds that present the fundamental frequency of around 20 kHz (the human hearing range), can be produced in both modalities (reviewed in 25). Together with chemical communication, USV emission is influenced by neuroendocrine control in both sexes and the neural pathways involved in the USV emission and signal processing are still under investigation (26, 27).

Biomedical studies usually record mouse USVs in specific contexts and at different ages: 1) wriggling calls and retrieving calls during pups' development; 2) “play” calls during social behaviors (28); 3) USVs during social interactions between same-sex adult females, and less frequently males; 4) courtship USVs (reviewed in 29) and are often used as non-invasive tools to investigate the genetic basis of neurodevelopmental disorders (e.g., autism, schizophrenia and speech disorders) (24, 30, 31). USVs are spectrally complex and are emitted in discrete elements defined as “calls” or “syllables”, *sensu lato*. Each syllable can be classified into 4-16 different classes according to the number, duration and different spectro-temporal features (32-35), including short temporal intervals of “silence” between each element within a sequence, and the longer intervals of “silence” between sequences (35-37). Thus, USVs have the features of bird songs whose sequences are still under investigation (32, 34, 35, 38).

Most studies on USVs have been performed on inbred and outbred strains (reviewed in 39), but only a few research groups have investigated the functions of ultrasonic vocalizations in wild house mice and mainly during courtship in *Mus musculus musculus* and *M. musculus domesticus* (34, 40-46). These studies provide the following results: 1) females preferred playback USVs from unfamiliar and unrelated males compared to familiar brothers (40), and show assortative preferences for USVs from conspecific (42); 2) males' courtship USVs contain distinctive individual vocal signatures and kinship signatures and are detectable during a single recording session (41); 3) mice show a distinct USV patterns depending on the sex of the sender and receiver, in *Mus musculus musculus* (44) and on the population of origin when comparing different populations of *Mus musculus domesticus* (34); 4) females can modulate their USVs when interacting with another female through a partition over time with different proposed functions: hierarchy establishment, affiliative behavior and familiarization over time (45), as previously suggested in outbred females (47, 48); 5) USV modulation and MUPs production (i.e., Darcin) can act as multimodal signals to reveal the health status of males (46).

Courtship vocalizations of laboratory mice have been extensively studied and yet research on their adaptive functions in wild house mice has barely begun. Courtship songs are elicited by female odor (soiled bedding, urine), or her presence (35) and USVs are often recorded when males and females are separated by a perforated partition (44, 49), or during direct interactions (33, 40, 50). However, more studies are needed to investigate whether socio-sexual experience might affect males' USVs and thus, enhance the chance to sire more offspring, and whether USVs can contain information on the age, sex, population (or strain), and identity of the vocalizing individual.

A few studies have found suggestive evidence for individual vocal signatures (IVSs), distinct vocal features used by conspecifics for individual recognition, in some USV parameters in laboratory mice. For example, the number of USVs and the repertoire size (i.e., number of classes of USVs) (38, 51); the sequence complexity in some strains (32, 38), but not in others (52); and in wild mice, only some spectrotemporal features over one 90-min recording (41). Thus, more studies are needed to detect which candidates can be reliable parameters for vocal individual recognition.

To analyze the composition of USVs in more detail, hundreds of USV recordings must be processed, but this is extremely time consuming and requires trained researchers. Recently, machine learning methods have been implemented to automatically detect and classify USVs (53-55). Since there is still no consensus on the correct number of USV classes for a biological relevant classification, three different approaches have been used to solve this dilemma: 1) a human-based USV classification method based on the shape of the vocalizations and according to measurable spectrotemporal features from the spectrogram (33, 34, 56, 57); 2) the use of trained mice to recognize specific USV human-based categories. Recent studies found that mice can distinguish USVs that are spectrally simple compared to complex USVs (i.e., USVs with frequency jumps and harmonics) (58). However, more spectrally similar USVs are more difficult to discriminate (59), and socially housed mice could perform better when compared to socially isolated individuals (60); 3) a fuzzy USV automatic classification method where all USVs are part of a classification continuum (61). Although some programs to classify USVs are designed to detect, cluster, and classify vocalizations, researchers should provide the golden standard to validate the automatic classification. These methods should also be generalizable to different contexts and mouse groups.

5.4 Research objectives

The first chapter includes our study on the modulation of chemosensory signals in seminatural enclosures and effects on reproductive success. In chapter two and three, we investigate possible functions related to the modulation of USVs after socio-sexual priming, in different social contexts, correlated with the reproductive success. In chapter four, we tested whether house mice (*M. m. musculus*), show individual vocal consistency and interindividual variation for individual recognition. Chapter five shows a cross-validation study in three laboratory strains: two inbred strains (C57/B6J and FVB) and one outbred strain (CD1). Chapter six explains the structure, validation and performance of the new version of an automatic USV detector and a new classification software applicable to both wild and laboratory mice recordings.

In the first chapter, we tested whether chemosensory signals indicate the sex and social status of individuals according to the production and regulation of pheromones and proteins in the mouse urine in seminatural enclosures. Our second aim was to test whether the regulation of MUPs and VOCs according to the mice' social status was related to the reproductive success in both sexes in *Mus musculus musculus*. We collected data from four replicates of the enclosures with the same sex ratio and we sampled mouse urine before, during (pooled over different time points), and after the enclosure over four months. We also compared our data with mice of the same population kept in the cage during the enclosure experiment. Behavioral information on the social status and reproductive success of mice were analyzed and correlated with the concentration of MUPs and VOC excretion from the sampled urine at different time points. A previous study on laboratory inbred mice reported that females showed a preference towards the urine of males that repeatedly won male-male agonistic encounters in the cage (15)

and where the winners defined as “dominant” males increased the concentration of VOCs that are known to affect sexual maturity and promote the Whitten effect in females (15-18). However, no previous studies showed a direct correlation between the social status of males, the Whitten effect and the reproductive success in mice.

In the second chapter, we aimed to test whether sexual priming affected the modulation of males’ USVs and signal males’ arousal and motivation to mate. Many studies on laboratory mice include a shorter or longer direct contact between males and females before recording to enhance USV emission in mice (35, 50, 62), but without data to support this method. Thus, our first aim was to experimentally test whether sexual priming of male wild mice that experienced a brief encounter with a female (i.e., sexual priming of five minutes duration) affects their USV emission. We compared four groups of males that were exposed to females either one, 10, 20, or 30 days after priming and compared their vocalizations with those from control mice that were sexually naïve, and never exposed to a female or her scent. Our second goal was to analyze in detail how the USV emission of socio-sexually unprimed males differed from the modulation of sexually primed males recorded at different time points after priming, and whether this effect was long-lasting. The extraction of the USV spectrotemporal features was performed with an improved version of the semi-automatic detection method Automatic Mouse Ultrasound Detector (A-MUD3.2) and we presented its performance. The threshold of detection in the new version could be adjusted according to the background noise of the recordings, and the software presents a built-in quality evaluation score for each putative segment scored as USV (43). For the qualitative analyses of vocalizations, we manually classified both sonic and ultrasonic vocalization. Each USV was classified according to 15 different categories and included in the

analyses of the repertoire size and composition. We also performed the semi-automatic USV detection and manual classification for the next two chapters.

In the third chapter, we analyzed the modulation of USVs during sexual encounters between either unrelated or related pairs and possible effects of inbreeding avoidance on their reproductive success. Our first aim was to test whether USV signal male's sexual arousal and changed accordingly during two consecutive social contexts. The social encounters included a first phase with a partition that separated the cage partners and then a second phase where male and female could directly interact. Our second aim was to test whether USV modulation might change according to the genetic relatedness of the partners, favoring disassortative mating. Previous studies on wild mice showed that females prefer spending time close to the playback of unrelated and unfamiliar males compared to familiar siblings (40). Our third aim focused on testing whether the relatedness of the breeding pairs and the USV emission during the social interactions correlated with their reproductive success. The same pairs were included in the breeding regime after being recorded for both social tests and their reproductive success measured. Two previous studies on laboratory mice correlated sociosexual experience with the reproductive success of the vocalizing individuals, however, males were first paired with a group of females and their reproductive success was measured, and then recorded with a different group of females, leading to unclear results (63, 64).

In the fourth chapter, we tested whether mice could change the USV emission upon female odor stimulation and whether USVs might contain distinctive individual vocal signatures over time. Our first aim was to test whether USV emission changed when the mouse was alone in the clean cage with no stimuli and compared it after stimulation with female odor (i.e., a pool of female urine). Since the male was alone in the cage, we could also test whether

mice show individual distinctiveness in the USV emission to signal individual's identity over three consecutive weeks. No previous study measured in detail individual distinctiveness in males' USV emission over time. A few studies on laboratory mice showed the consistency in USV emission as part of other research questions. The main findings included a high inter-individual variation in the USV number and repertoire size (i.e., number of different syllables), but not in other spectrotemporal parameters (38, 51); and the consistency in the spectral complexity of some USV categories in inbred laboratory mice (32, 38). Thus, our first aim was to investigate in detail the modulation of USVs including the number and duration of USVs, the vocal repertoire and the spectrotemporal features upon female stimulation. Our second aim was to test whether mice showed a higher interindividual variation over time, and can be grouped according to their vocalizing style. Our third aim was to assess whether USV emission could signal vocal personality and we aimed to identify candidates as individual vocal signatures (IVSs). IVS candidates should fulfill two criteria: 1) to show high interindividual and low intraindividual variation; and 2) to be consistent over time (i.e., repeatable within the same individual). We also applied multivariate statistics and machine learning methods to predict whether the candidates as IVSs might be reliable features used for individual recognition.

In the fifth chapter, we aimed to cross-validate the results on the same behavioral tests and perform USV recordings of three laboratory strains: two inbred and one outbred laboratory strains, commonly used for the study of neurodevelopmental disorders (NDDs) (33, 65-68). The first aim of the study was to test whether the three strains differed in their socio-communicative profile at each time point and social context. Our second goal was to present such differences and make suggestions on how to choose the best animal model for specific

research questions. USV parameters were inspected in detail and used as biomarkers to detect qualitative alterations in the vocal features of mouse models in studies on NDDs at different ages (i.e., pups and adults). Most studies mainly focused on the early phase of the individuals' development during the first weeks of life or during adolescence (28, 69, 70). Thus, we recorded different groups of mice from both sexes and at different ages: 1) as pups, while briefly isolated from their mothers; 2) as adults, during direct social interactions (i.e., between same and different sex mice). We analyzed in detail the USVs according to ten different syllable classes to draw a profile according to sex, age, and strain in isolated pups and interacting adults. Behavioral and vocal profiles were correlated in adult mice and compared between groups (33).

In the sixth chapter, we aimed to compare the performance of our improved version of the semi-automatic USV detection method (i.e., A-MUD3.2) with other state-of-the-art and available detection methods. Detected segments were used as putative USVs and classified using Bootsnap, an ensemble machine learning method with minimum intervention of the researcher. Our second aim was to validate Bootsnap performance by comparing it with the "ground truth" obtained by trained researchers who manually classified recordings from both wild mice and laboratory mice also reporting the inter-observer reliability for each USV category. In the last few years, many USV classification models have been implemented. However, each software has some limitations: some software is not user-friendly, some has a high signal-to-noise ratio preventing a good analysis on files with a lot of background noise, whereas others are designed to work on platforms that are not free for scientific use (53, 71-73). Thus, our third goal was to develop a classification method that aimed at the generalizability of the datasets used for the train and test sessions (i.e., recordings from mice of different sex, at different ages and social contexts), at the best classification accuracy also with

low quality recordings (i.e., low signal-to-noise-ratio), with a minimum external intervention, and free for scientific use. Bootsnap was trained on wild mice recordings and tested for both wild and laboratory mice recordings in different social contexts (i.e., male-female interaction through a partition, direct interaction between males and females and males exposed to female odor) and its performance was compared to the best state-of-the-art USV classifier.

References

1. Steensma DP, Kyle RA, Shampo MA. Abbie Lathrop, the "Mouse Woman of Granby": rodent fancier and accidental genetics pioneer. Mayo Clinic Proceedings. 2010;85:e83. Epub November 2010. doi: 10.4065/mcp.2010.0647.
2. Ullrich KK, Tautz D. Population genomics of the house mouse and the brown rat. Methods Mol Biol. 2020;2090:435-52. Epub 2020/01/25. doi: 10.1007/978-1-0716-0199-0_18. PubMed PMID: 31975178.
3. Mounier A, Mirazón Lahr M. Deciphering African late middle Pleistocene hominin diversity and the origin of our species. Nat Commun. 2019;10(1):3406. Epub 2019/09/12. doi: 10.1038/s41467-019-11213-w. PubMed PMID: 31506422; PubMed Central PMCID: PMCPMC6736881.
4. Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 2007;448(7157):1050-3. doi: 10.1038/nature06067.
5. Guénet JL, Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet. 2003;19(1):24-31. Epub 2002/12/21. doi: 10.1016/s0168-9525(02)00007-0. PubMed PMID: 12493245.
6. Didion JP, de Villena FP. Deconstructing Mus gemischus: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. Mamm Genome. 2013;24(1-2):1-20. Epub 2012/12/12. doi: 10.1007/s00335-012-9441-z. PubMed PMID: 23223940; PubMed Central PMCID: PMCPMC4034049.
7. Tuttle AH, Philip VM, Chesler EJ, Mogil JS. Comparing phenotypic variation between inbred and outbred mice. Nat Methods. 2018;15(12):994-6. Epub 2018/12/07. doi: 10.1038/s41592-018-0224-7. PubMed PMID: 30504873; PubMed Central PMCID: PMCPMC6518396.
8. Tam WY, Cheung K-K. Phenotypic characteristics of commonly used inbred mouse strains. Journal of Molecular Medicine. 2020;98(9):1215-34. doi: 10.1007/s00109-020-01953-4.
9. Musolf K, Penn DJ. Ultrasonic vocalizations in house mice: a cryptic mode of acoustic communication. In: Macholan M, Baird SJE, Munclinger P, Pialek J, editors. Evolution of the house mouse: Cambridge University Press; 2012. p. 253-77.
10. Koolhaas JM, Coppens CM, de Boer SF, Buwalda B, Meerlo P, Timmermans PJ. The resident-intruder paradigm: a standardized test for aggression, violence and social stress. J Vis Exp. 2013(77):e4367. Epub 2013/07/16. doi: 10.3791/4367. PubMed PMID: 23852258; PubMed Central PMCID: PMCPMC3731199.
11. Roberts SA, Prescott MC, Davidson AJ, McLean L, Beynon RJ, Hurst JL. Individual odour signatures that mice learn are shaped by involatile major urinary proteins (MUPs). BMC Biology. 2018;16(1):48. doi: 10.1186/s12915-018-0512-9.

12. Thoß M, Luzynski K, Ante M, Miller I, Penn D. Major urinary protein (MUP) profiles show dynamic changes rather than individual “barcode” signatures. Frontiers in Ecology and Evolution. 2015;3. doi: 10.3389/fevo.2015.00071.

13. Thoß M, Enk V, Yu H, Miller I, Luzynski KC, Balint B, et al. Diversity of major urinary proteins (MUPs) in wild house mice. Scientific Reports. 2016;6(1):38378. doi: 10.1038/srep38378.

14. Kwak J, Strasser E, Luzynski K, Thoß M, Penn DJ. Are MUPs a toxic waste disposal system? PLOS ONE. 2016;11(3):e0151474. doi: 10.1371/journal.pone.0151474.

15. Fang Q, Zhang Y-H, Shi Y-L, Zhang J-H, Zhang J-X. Individuality and transgenerational inheritance of social dominance and sex pheromones in isogenic male mice. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2016;326(4):225-36. doi: https://doi.org/10.1002/jez.b.22681.

16. Mucignat-Caretta C, Caretta A, Cavaggioni A. Acceleration of puberty onset in female mice by male urinary proteins. J Physiol. 1995;486 (Pt 2)(Pt 2):517-22. Epub 1995/07/15. doi: 10.1113/jphysiol.1995.sp020830. PubMed PMID: 7473215; PubMed Central PMCID: PMCPMC1156539.

17. Jemiolo B, Alberts J, Sochinski-Wiggins S, Harvey S, Novotny M. Behavioural and endocrine responses of female mice to synthetic analogues of volatile compounds in male urine. Animal Behaviour. 1985;33(4):1114-8. doi: https://doi.org/10.1016/S0003-3472(85)80170-6.

18. Jemiolo B, Harvey S, Novotny M. Promotion of the Whitten effect in female mice by synthetic analogs of male urinary constituents. Proc Natl Acad Sci U S A. 1986;83(12):4576-9. Epub 1986/06/01. doi: 10.1073/pnas.83.12.4576. PubMed PMID: 3459193; PubMed Central PMCID: PMCPMC323778.

19. Novotny MV, Ma W, Wiesler D, Zídek L. Positive identification of the puberty-accelerating pheromone of the house mouse: the volatile ligands associating with the major urinary protein. Proc Biol Sci. 1999;266(1432):2017-22. Epub 1999/12/10. doi: 10.1098/rspb.1999.0880. PubMed PMID: 10584338; PubMed Central PMCID: PMCPMC1690305.

20. Jouhanneau M, Goudet C, Moussu C, Tashiro T, Buatois B, Mori K, et al. Peripubertal exposure to male chemosignals accelerates vaginal opening and induces male-directed odor preference in female mice. Frontiers in Ecology and Evolution. 2015;3. doi: 10.3389/fevo.2015.00034.

21. Wölfl S, Zala SM, Penn DJ. Male scent but not courtship vocalizations induce estrus in wild female house mice. Physiology & Behavior. 2023;259:114053. doi: https://doi.org/10.1016/j.physbeh.2022.114053.

22. Zala SM, Church B, Potts WK, Knauer F, Penn DJ. Female scent accelerates growth of juvenile male mice. Sci Rep. 2023;13(1):7371. Epub 2023/05/06. doi: 10.1038/s41598-023-34548-3. PubMed PMID: 37147391; PubMed Central PMCID: PMCPMC10163255.

23. Zippelius H-M, Schleidt WM. Ultraschall-Laute bei jungen Mäusen. Naturwissenschaften. 1956;43(21):502-. doi: 10.1007/BF00632534.

24. Ehret G. Chapter 18 - Characteristics of vocalization in adult mice. In: Brudzynski SM, editor. Handbook of Behavioral Neuroscience. 25: Elsevier; 2018. p. 187-95.

25. Fernández-Vargas M, Riede T, Pasch B. Mechanisms and constraints underlying acoustic variation in rodents. Animal Behaviour. 2022;184:135-47. doi: https://doi.org/10.1016/j.anbehav.2021.07.011.

26. Asaba A, Hattori T, Mogi K, Kikusui T. Sexual attractiveness of male chemicals and vocalizations in mice. Frontiers in Neuroscience. 2014;8. doi: 10.3389/fnins.2014.00231.

27. Hoglen NEG, Manoli DS. Cupid's quiver: Integrating sensory cues in rodent mating systems. Front Neural Circuits. 2022;16:944895. Epub 2022/08/13. doi: 10.3389/fncir.2022.944895. PubMed PMID: 35958042; PubMed Central PMCID: PMCPMC9358210.

28. Panksepp JB, Jochman KA, Kim JU, Koy JJ, Wilson ED, Chen Q, et al. Affiliative behavior, ultrasonic communication and social reward are influenced by genetic variation in adolescent mice. PLOS ONE. 2007;2(4):e351. doi: 10.1371/journal.pone.0000351.

29. Premoli M, Pietropaolo S, Wöhr M, Simola N, Bonini SA. Mouse and rat ultrasonic vocalizations in neuroscience and neuropharmacology: state of the art and future applications. European Journal of Neuroscience. 2023;57(12):2062-96. doi: https://doi.org/10.1111/ejn.15957.

30. Heckman J, McGuinness B, Celikel T, Englitz B. Determinants of the mouse ultrasonic vocal structure and repertoire. Neurosci Biobehav Rev. 2016;65:313-25. Epub 2016/04/10. doi: 10.1016/j.neubiorev.2016.03.029. PubMed PMID: 27060755.

31. Brudzynski SM, Brudzynski S. Handbook of ultrasonic vocalization a window into the emotional brain. London, United Kingdom; Academic Press; 2018.

32. Holy TE, Guo Z. Ultrasonic songs of male mice. PLoS Biol. 2005;3(12):e386. Epub 2005/10/27. doi: 10.1371/journal.pbio.0030386. PubMed PMID: 16248680; PubMed Central PMCID: PMCPMC1275525.

33. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes Brain Behav. 2011;10(1):44-56. Epub 2010/07/14. doi: 10.1111/j.1601-183X.2010.00623.x. PubMed PMID: 20618443; PubMed Central PMCID: PMCPMC2972364.

34. von Merten S, Hoier S, Pfeifle C, Tautz D. A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (Mus musculus domesticus). PLOS ONE. 2014;9(5):e97244. doi: 10.1371/journal.pone.0097244.

35. Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Frontiers in Behavioral Neuroscience. 2015;9. doi: 10.3389/fnbeh.2015.00076.

36. Castellucci GA, McGinley MJ, McCormick DA. Knockout of Foxp2 disrupts vocal development in mice. Sci Rep. 2016;6:23305. Epub 2016/03/17. doi: 10.1038/srep23305. PubMed PMID: 26980647; PubMed Central PMCID: PMCPMC4793191.

37. Castellucci GA, Calbick D, McCormick D. The temporal organization of mouse ultrasonic vocalizations. PLoS One. 2018;13(10):e0199929. Epub 2018/10/31. doi: 10.1371/journal.pone.0199929. PubMed PMID: 30376572; PubMed Central PMCID: PMCPMC6207298.

38. Melotti L, Siestrup S, Peng M, Vitali V, Dowling D, von Kortzfleisch VT, et al. Individuality, as well as genetic background, affects syntactical features of courtship songs in male mice. Animal Behaviour. 2021;180:179-96. doi: https://doi.org/10.1016/j.anbehav.2021.08.003.

39. Sterling ML, Teunisse R, Englitz B. Rodent ultrasonic vocal interaction resolved with millimeter precision using hybrid beamforming. eLife. 2023;12:e86126. doi: 10.7554/eLife.86126.

40. Musolf K, Hoffmann F, Penn DJ. Ultrasonic courtship vocalizations in wild house mice, Mus musculus musculus. Animal Behaviour. 2010;79(3):757-64. doi: https://doi.org/10.1016/j.anbehav.2009.12.034.

41. Hoffmann F, Musolf K, Penn DJ. Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. Physiol Behav. 2012;105(3):766-71. Epub 2011/11/01. doi: 10.1016/j.physbeh.2011.10.011. PubMed PMID: 22037196.

42. Musolf K, Meindl S, Larsen AL, Kalcounis-Rueppell MC, Penn DJ. Ultrasonic vocalizations of male mice differ among species and females show assortative preferences for male calls. PLoS One. 2015;10(8):e0134123. Epub 2015/08/27. doi: 10.1371/journal.pone.0134123. PubMed PMID: 26309246; PubMed Central PMCID: PMCPMC4550448.

43. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Automatic mouse ultrasound detector (A-MUD): a new tool for processing rodent vocalizations. PLoS One. 2017;12(7):e0181200. Epub 2017/07/21. doi: 10.1371/journal.pone.0181200. PubMed PMID: 28727808; PubMed Central PMCID: PMCPMC5519055.

44. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (Mus musculus musculus). PLoS One. 2017;12(12):e0188647. Epub 2017/12/14. doi: 10.1371/journal.pone.0188647. PubMed PMID: 29236704; PubMed Central PMCID: PMCPMC5728457.

45. Hoier S, Pfeifle C, von Merten S, Linnenbrink M. Communication at the garden fence - context dependent vocalization in female house mice. PloS one. 2016;11(3):e0152255-e. doi: 10.1371/journal.pone.0152255. PubMed PMID: 27022749.

46. Lopes PC, König B. Choosing a healthy mate: sexually attractive traits as reliable indicators of current disease status in house mice. Animal Behaviour. 2016;111:119-26. doi: https://doi.org/10.1016/j.anbehav.2015.10.011.

47. D'Amato FR, Moles A. Ultrasonic vocalizations as an index of social memory in female mice. Behav Neurosci. 2001;115(4):834-40. Epub 2001/08/18. doi: 10.1037//0735-7044.115.4.834. PubMed PMID: 11508722.

48. Moles A, Costantini F, Garbugino L, Zanettini C, D'Amato FR. Ultrasonic vocalizations emitted during dyadic interactions in female mice: a possible index of sociability? Behavioural Brain Research. 2007;182(2):223-30. doi: https://doi.org/10.1016/j.bbr.2007.01.020.

49. Burke K, Screven LA, Dent ML. CBA/CaJ mouse ultrasonic vocalizations depend on prior social experience. PLoS One. 2018;13(6):e0197774. Epub 2018/06/07. doi: 10.1371/journal.pone.0197774. PubMed PMID: 29874248; PubMed Central PMCID: PMCPMC5991354.

50. Hanson JL, Hurley LM. Female presence and estrous state influence mouse ultrasonic courtship vocalizations. PLOS ONE. 2012;7(7):e40782. doi: 10.1371/journal.pone.0040782.

51. Rieger MA, Dougherty JD. Analysis of within subjects variability in mouse ultrasonic vocalization: pups exhibit inconsistent, state-like patterns of call production. Front Behav Neurosci. 2016;10:182. Epub 2016/10/14. doi: 10.3389/fnbeh.2016.00182. PubMed PMID: 27733819; PubMed Central PMCID: PMCPMC5039195.

52. Mahrt EJ, Perkel DJ, Tong L, Rubel EW, Portfors CV. Engineered deafness reveals that mouse courtship vocalizations do not require auditory experience. J Neurosci. 2013;33(13):5573-83. Epub 2013/03/29. doi: 10.1523/jneurosci.5054-12.2013. PubMed PMID: 23536072; PubMed Central PMCID: PMCPMC3691057.

53. Coffey KR, Marx RE, Neumaier JF. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology. 2019;44(5):859-68. doi: 10.1038/s41386-018-0303-6.

54. Premoli M, Baggi D, Bianchetti M, Gnutti A, Bondaschi M, Mastinu A, et al. Automatic classification of mice vocalizations using machine learning techniques and convolutional neural networks. PLoS One. 2021;16(1):e0244636. Epub 2021/01/20. doi: 10.1371/journal.pone.0244636. PubMed PMID: 33465075; PubMed Central PMCID: PMCPMC7815145.

55. Pessoa D, Petrella L, Martins P, Castelo-Branco M, Teixeira C. Automatic segmentation and classification of mice ultrasonic vocalizations. J Acoust Soc Am. 2022;152(1):266. Epub 2022/08/06. doi: 10.1121/10.0012350. PubMed PMID: 35931540.

56. Grimsley JMS, Monaghan JJM, Wenstrup JJ. Development of Social Vocalizations in Mice. PLOS ONE. 2011;6(3):e17460. doi: 10.1371/journal.pone.0017460.

57. Zala SM, Nicolakis D, Marconi MA, Noll A, Ruf T, Balazs P, et al. Primed to vocalize: wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female. PLoS One. 2020;15(12):e0242959. Epub 2020/12/10. doi: 10.1371/journal.pone.0242959. PubMed PMID: 33296411; PubMed Central PMCID: PMCPMC7725367.

58. Neilans EG, Holfoth DP, Radziwon KE, Portfors CV, Dent ML. Discrimination of ultrasonic vocalizations by CBA/CaJ mice (Mus musculus) is related to spectrotemporal dissimilarity of vocalizations. PLOS ONE. 2014;9(1):e85405. doi: 10.1371/journal.pone.0085405.

59. Screven LA, Dent ML. Discrimination of frequency modulated sweeps by mice. J Acoust Soc Am. 2016;140(3):1481. Epub 2016/12/05. doi: 10.1121/1.4962223. PubMed PMID: 27914389; PubMed Central PMCID: PMCPMC6910002.

60. Screven L, A., Dent ML. Perception of ultrasonic vocalizations by socially housed and isolated mice. eneuro. 2019;6(5):ENEURO.0049-19.2019. doi: 10.1523/ENEURO.0049-19.2019.

61. Goffinet J, Brudner S, Mooney R, Pearson J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. eLife. 2021;10:e67855. doi: 10.7554/eLife.67855.

62. Arriaga G, Zhou EP, Jarvis ED. Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds. PLOS ONE. 2012;7(10):e46610. doi: 10.1371/journal.pone.0046610.

63. Asaba A, Osakada T, Touhara K, Kato M, Mogi K, Kikusui T. Male mice ultrasonic vocalizations enhance female sexual approach and hypothalamic kisspeptin neuron activity. Hormones and Behavior. 2017;94:53-60. doi: https://doi.org/10.1016/j.yhbeh.2017.06.006.

64. Kanno K, Kikusui T. Effect of Sociosexual Experience and Aging on Number of Courtship Ultrasonic Vocalizations in Male Mice. Zoolog Sci. 2018;35(3):208-14. Epub 2018/06/09. doi: 10.2108/zs170175. PubMed PMID: 29882498.

65. Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism. PLoS One. 2008;3(8):e3067. Epub 2008/08/30. doi: 10.1371/journal.pone.0003067. PubMed PMID: 18728777; PubMed Central PMCID: PMCPMC2516927.

66. Gaub S, Fisher SE, Ehret G. Ultrasonic vocalizations of adult male Foxp2-mutant mice: behavioral contexts of arousal and emotion. Genes Brain Behav. 2016;15(2):243-59. Epub 2015/11/15. doi: 10.1111/gbb.12274. PubMed PMID: 26566793.

67. Simola N, Granon S. Ultrasonic vocalizations as a tool in studying emotional states in rodent models of social behavior and brain disease. Neuropharmacology. 2019;159:107420. Epub 2018/11/18. doi: 10.1016/j.neuropharm.2018.11.008. PubMed PMID: 30445100.

68. Premoli M, Memo M, Bonini SA. Ultrasonic vocalizations in mice: relevance for ethologic and neurodevelopmental disorders studies. Neural Regen Res. 2021;16(6):1158-67. Epub 2020/12/04. doi: 10.4103/1673-5374.300340. PubMed PMID: 33269765; PubMed Central PMCID: PMCPMC8224126.

69. Peleh T, Eltokhi A, Pitzer C. Longitudinal analysis of ultrasonic vocalizations in mice from infancy to adolescence: insights into the vocal repertoire of three wild-type strains in two different social contexts. PLOS ONE. 2019;14(7):e0220238. doi: 10.1371/journal.pone.0220238.

70. Binder MS, Pranske ZJ, Lugo JN. Evaluating the DeepSqueak and Mouse Song Analyzer vocalization analysis systems in C57BL/6J, FVB.129, and FVB neonates. J Neurosci

Methods. 2021;364:109356. Epub 2021/09/12. doi: 10.1016/j.jneumeth.2021.109356. PubMed PMID: 34508783.

71. Chabout J, Jones-Macopson J, Jarvis ED. Eliciting and analyzing male mouse ultrasonic vocalization (USV) songs. J Vis Exp. 2017(123). Epub 2017/05/19. doi: 10.3791/54137. PubMed PMID: 28518074; PubMed Central PMCID: PMCPMC5607930.

72. Van Segbroeck M, Knoll AT, Levitt P, Narayanan S. MUPET—Mouse Ultrasonic Profile Extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. Neuron. 2017;94(3):465-85.e5. doi: https://doi.org/10.1016/j.neuron.2017.04.005.

73. Tachibana RO, Kanno K, Okabe S, Kobayasi KI, Okanoya K. USVSEG: a robust method for segmentation of ultrasonic vocalizations in rodents. PLOS ONE. 2020;15(2):e0228907. doi: 10.1371/journal.pone.0228907.

CHAPTER 1: Pheromones that correlate with reproductive success in competitive conditions

Luzynski Kenneth C.; Nicolakis, Doris; Marconi, Maria Adelaide;
Zala, Sarah M.; Kwak, Jae; Penn, Dustin J.

Scientific Reports, impact factor: 4.6; **published**

scientific reports

Check for updates

OPEN

Pheromones that correlate with reproductive success in competitive conditions

Kenneth C. Luzynski[1]✉, Doris Nicolakis[1], Maria Adelaide Marconi[1], Sarah M. Zala[1], Jae Kwak[2,3] & Dustin J. Penn[1]✉

The major urinary proteins (MUPs) of house mice (*Mus musculus*) bind and stabilize the release of pheromones and other volatile organic compounds (VOCs) from urinary scent marks, which mediate chemical communication. Social status influences MUP and VOC excretion, and the urinary scent of dominant males is attractive to females. Urinary pheromones influence the sexual behavior and physiology of conspecifics, and yet it is not known whether they also affect reproductive success. We monitored the excretion of urinary protein and VOCs of wild-derived house mice living in large seminatural enclosures to compare the sexes and to test how these compounds correlate with reproductive success. Among males, urinary protein concentration and VOC expression correlated with reproductive success and social status. Territorial dominance also correlated with reproductive success in both sexes; but among females, no urinary compounds were found to correlate with social status or reproductive success. We found several differences in the urinary protein and volatile pheromones of mice in standard cages versus seminatural enclosures, which raises caveats for conventional laboratory studies. These findings provide novel evidence for chemical signals that correlate with male reproductive success of house mice living in competitive conditions.

Male house mice scent-mark their territories with urine and they excrete several compounds often proposed to enhance mating and reproductive success[1–4]. Males produce large quantities of protein in their urine, mainly composed of major urinary proteins (MUPs)[5,6]. MUPs bind and stabilize the release of volatile organic compounds (VOCs) from urinary scent marks[7,8]. These VOCs include the male pheromones, α- and β-farnesene, 2-s-butyl-4, 5-dihydrothiazole (SBT), 3,4-dehydro-exo-brevicomin (DHB), and 6-hydroxy-6-methyl-3-heptanone (HMH), which trigger changes in female sexual development, physiology and behavior[1,9,10]. Trimethylamine (TMA) is a sexually dimorphic VOC that is highly expressed in males. Interestingly, it is attractive to mice at normal levels, but aversive at high levels in urine[11]. MUP proteoforms can act as pheromones as well as transporters, and MUP20 (darcin) is a predominantly male-expressed urinary protein that elicits place preferences and spatial learning in female mice[12]. Females detect MUPs in male urine by upregulating VNO receptor expression during estrus[13], and exposure to a combination of volatile male pheromones (SBT, DHB, and HMH) induces female olfactory preferences for these compounds and accelerates puberty[1,14] (but see[15]). Most mammalian studies on female responses to male pheromones examined domesticated mouse strains in artificial laboratory conditions, and focused on female sexual development, estrous cycling, lordosis, or pregnancy block. It is still not known whether male pheromones influence reproductive success. Our first goal was to test whether these pheromones influence the reproductive success of wild-derived male house mice in seminatural conditions.

One way that chemical signals are often suggested to influence reproductive success is by providing a reliable indicator of social status[16]. Wild house mice are highly territorial, and dominant, territorial males have higher reproductive success than non-territorial subordinates[17]. Males that are socially aggressive have higher urinary protein concentrations[18,19] (but see[20]), and produce higher intensities of particular volatile pheromones (DHB, HMH, SBT, and α/β-farnesene) than submissive males[10,21]. Social defeat can result in decreased expression[21]. Estrous females prefer the urinary scent of aggressive 'dominant' males over submissive 'subordinate' males[22]. An important caveat to these studies is that most were conducted with laboratory mice, and male social status was assessed using the outcomes of brief, dyadic agonistic interactions in the laboratory; a proxy that does not

[1]Department of Interdisciplinary Life Sciences, Konrad Lorenz Institute of Ethology, University of Veterinary Medicine Vienna, Savoyenstraße 1, Vienna, Austria. [2]Department of Interdisciplinary Life Sciences, Research Institute of Wildlife Ecology, University of Veterinary Medicine Vienna, Savoyenstraße 1, Vienna, Austria. [3]International Flavors & Fragrances Inc., 1515 State Route 36, Union Beach, NJ, USA. ✉email: luzyn1kl@gmail.com; Dustin.Penn@vetmeduni.ac.at

Model predictor	Estimate (SEM)	*z*-value	p-value	Sum of weights	No. containing models	Variance inflation factor (VIF)
(Intercept)	− 3.57 (5.88)	0.59	0.55	–	–	–
Total protein	*1.47E−3 (5.78E−4)*	*2.54*	*0.011*	*0.75*	*16*	*1.32*
Social status	*− 1.10 (0.51)*	*2.05*	*0.04*	*0.69*	*16*	*1.69*
Age	1.40E−2 (8.70E−3)	1.51	0.13	0.39	16	1.49
Creatinine	− 3.30E−3 (2.57E3)	1.21	0.22	0.32	15	1.33
Mass	− 1.71E−1 (1.76E−1)	0.91	0.36	0.23	16	1.69

Table 1. Male reproductive success in response to urinary protein and social status. LME conditional model average for effects of male urine composition, age, mass, and social status during enclosures on RS (natural log transformed (1 + no. of offspring)). Table rows are ordered by predictor importance based on the sum of weights. Significant predictors are italicized. PC ratio omitted due to high VIF value suggesting collinearity (see Supplementary Table S1 for full model).

necessarily predict social status in more natural social conditions. Indeed, one study found that this proxy did not correlate with social status of group-housed male mice[23]. A recent study on wild-derived house mice living in seminatural conditions found that once males acquired a territory and became socially dominant, they increased the production of some (MUP20 and HMH), but not other pheromones (e.g. SBT, DHB, farnesene), whereas males did not reduce pheromone excretion after they became subordinates[24]. Estrous females were more attracted to the urinary scent of dominant, territorial males than subordinates, and variation in protein concentration of male urine had no effect on female preferences when male social status was controlled. This study confirmed that male pheromone expression is context- and status-dependent and that estrous females are more attracted to the scent of dominant than subordinate males, but unlike studies on social defeat in the laboratory, subordinate males did not reduce pheromone excretion in naturalistic conditions.

Studies are also needed to compare the expression of chemical signals between the sexes in more natural social contexts, and test for compounds that influence female reproductive success. Putative pheromones in female urine include 2-heptanone[25], 2,5-dimethylpyrazine[25], and isobutylamine (IBA)[26], which have been reported to signal estrus[27,28], attract males[28], and delay puberty in juvenile females[29,30] (though the opposite effect is reported for IBA[26]). Yet the pheromones of female mice remain under-investigated, and there have been no studies on female VOCs in seminatural conditions to our knowledge. Two studies investigated female MUP excretion in seminatural conditions, and one found that female MUP excretion was positively correlated with aggressive behaviors[31], whereas another found no such relationship[24]. Both studies found that the large sex difference in urinary protein concentration reported in standard laboratory conditions was significantly lower when mice live in seminatural conditions due to increased female urinary protein excretion[24,31]. Therefore, our second aim was to compare the production of volatile and non-volatile urinary compounds between the sexes, and test whether these compounds are regulated depending upon social status or correlate with reproductive success.

We conducted our study on wild-derived house mice (F3 from wild-trapped *M. musculus musculus*) in seminatural conditions. We released mice into four indoor enclosures (9 m × 4 m each; Supplementary Fig. S1) for 16 weeks and recorded their social behavior. Urine was collected at 4-week intervals throughout the experiment. We measured urinary protein and used gas chromatography coupled with mass spectrometry (GC–MS) to quantify VOCs at multiple time points, allowing us to compare pheromone expression before and during the seminatural enclosure phase. We expected that upregulation and excretion of MUPs and volatile pheromones would influence male reproductive success, and that the regulation of these compounds would depend upon their social status[24]. Furthermore, we expected that the degree of sexual dimorphism in urinary compounds would change in competitive conditions due to socially-mediated effects on the chemical signals of both sexes[24,31–34], but we did not expect females to regulate protein excretion depending upon their social status[24] or to correlate with reproductive success[31].

Results

Reproductive success. The production of urinary pheromones correlated with male but not female reproductive success (RS; defined in "Materials and methods" section). The most important predictors of male RS were total urinary protein concentration (75%) and social status (69%; Table 1; based on conditional model average sum of weights). The relative importance of age, creatinine, and mass ranged from 23 to 39%; PC ratio (protein:creatinine concentration) was excluded from the model due to collinearity (VIF = 6.97). Total urinary protein concentration during the enclosure phase was positively correlated with RS for males (Spearman $R = 0.52$, $p = 0.01$; Fig. 1a), but not females (Fig. 1b). This correlation is explained by the low protein concentration in the urine of non-reproductive males, as it is no longer significant after removing these males from the analysis ($R = 0.12$, $p = 0.62$; Supplementary Fig. S2). The median total urinary protein concentration was 5512 µg mL^{-1} and 5028 µg mL^{-1} for reproductive and non-reproductive males, respectively (Wilcoxon rank sum test $W = 5$, $p < 0.001$; Supplementary Fig. S2).

The most important predictors of female RS were mean body mass (89%) and social status (76%), whereas age, PC ratio, and total protein and creatinine concentration ranged from 14 to 20% (Supplementary Table S1). Female mean body mass during the enclosure was positively correlated to RS ($R = 0.57$, $p = 0.004$). When mean body mass during the enclosure is replaced with initial body mass as a model predictor, the relative influence of

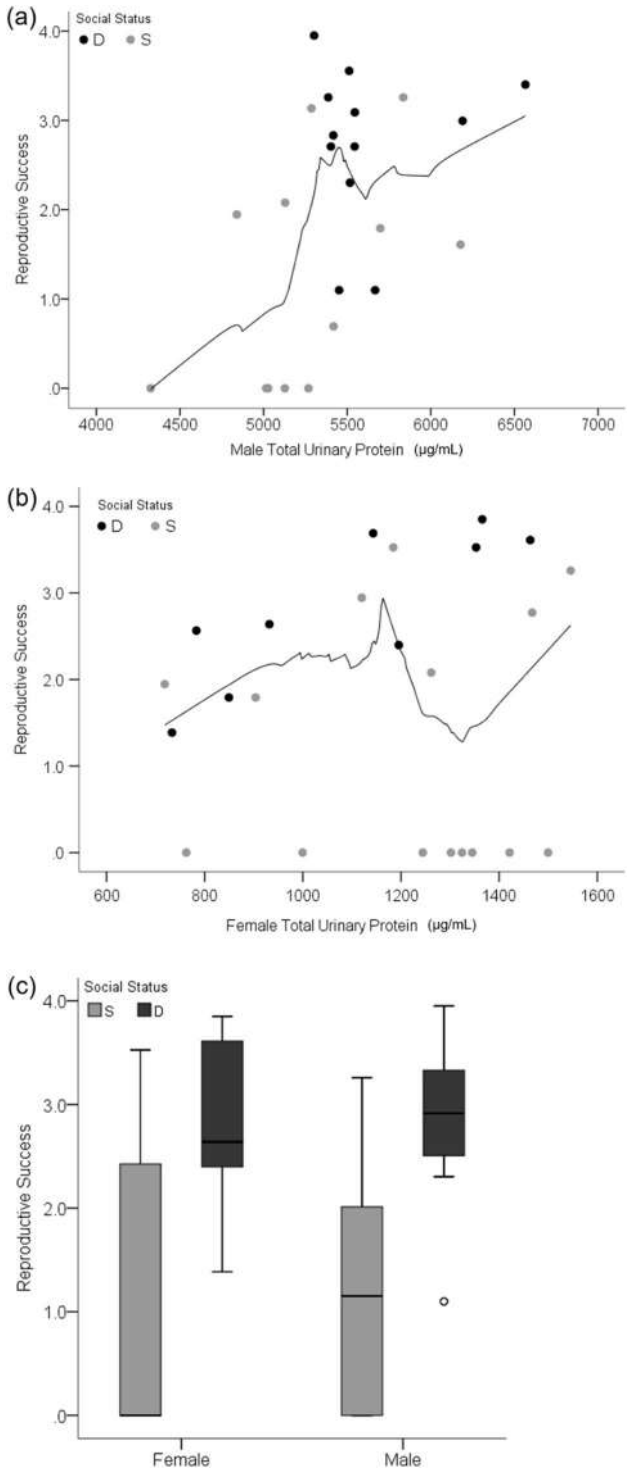

Figure 1. Reproductive success in relation to urinary protein and social status. Scatterplots show the total urinary protein concentration of males (**a**) and females (**b**) in relation to reproductive success. The boxplot (**c**) shows female and male social status in relation to reproductive success. Light gray coloration of data points and boxes indicate subordinate (S) social status during the enclosures. Black data points and dark gray boxes indicate dominant (D) social status. The black trend line in the scatterplots shows the loess (local regression) fit for non-parametric data (50% of data points to fit Epanechnikov kernel).

social status on female RS is 94%; all other variables ranged from 14 to 34% with initial body mass at 23% (Supplementary Table S1). For both sexes, dominant individuals (male = 12; female = 9) accounted for the majority

of reproduction compared to subordinates (male = 12; female = 15; Welch's t-test post hoc male $p = 0.006$, female $p = 0.01$; Fig. 1c). Reproduction in the enclosures resulted in 306 offspring from 51 litters (multiple paternity in 69%; Supplementary Table S2). Mate fidelity was 29% and 8% for males and females, respectively. The non-reproductive mice were all subordinates (male = 5; female = 8).

Male urinary VOC expression during the enclosure phase also correlated with male reproductive success. The explained variance (R^2Y) and cross validation score (Q^2) of orthogonal partial least-squares (OPLS) models showed a significant correlation between RS and VOC expression of denatured and intact urine (Fig. 2a; denatured: $R^2Y = 0.54$, $Q^2 = 0.46$; intact: $R^2Y = 0.51$, $Q^2 = 0.39$). Two specific urinary volatiles, HMH and TMA, correlated with male RS. In intact urine, peaks corresponding to HMH expression during enclosures were positively correlated to RS (Fig. 2b; $R = 0.63$, $p_{adj} < 0.004$), but this correlation is weak in denatured urine ($R = 0.47$, $p_{adj} = 0.02$ (n.s.)). We also confirmed that minor ions of HMH in intact urine were correlated with male RS (8 HMH peaks: $R > 0.61$, $p_{adj} < 0.004$). TMA was negatively correlated with RS during the enclosure phase, regardless of protein conformation (Fig. 2c; intact: $R = -0.59$, $p_{adj} < 0.004$; denatured: $R = -0.55$, $p_{adj} < 0.008$). After omitting non-reproductive males, the correlations between reproductive male RS and both HMH and TMA expression were not significant (HMH: $R = 0.23$, $p = 0.36$; TMA: $R = -0.12$, $p = 0.62$; Supplementary Fig. S2). Significant differences in HMH and TMA expression were observed when comparing reproductive and non-reproductive males (Wilcoxon rank sum test $p < 0.003$ for both VOCs; Supplementary Fig. S2).

Male RS was correlated with both pheromone excretion and social status, and therefore, we examined dominants and subordinates separately and re-ran OPLS models to isolate the effect of VOC expression on reproduction. VOC expression and RS did not correlate among dominant males (OPLS model $Q^2 < 0$, $p > 0.05$), whereas the VOC expression of intact urine from subordinate males was strongly correlated with RS, and to a lesser degree in denatured urine (intact: $R^2Y = 0.75$, $Q^2 = 0.64$; denatured: $R^2Y = 0.59$, $Q^2 = 0.49$). We found a correlation of subordinate male RS with HMH expression ($R = 0.71$, $p = 0.01$), and a negative correlation with TMA ($R = -0.70$, $p = 0.01$), though neither were significant after Bonferroni adjustment for multiple comparisons (refer to "Materials and methods").

Among females, we found no significant associations between VOC expression during the enclosure phase and RS (OPLS models: R^2Y and Q^2 $p > 0.05$; Supplementary Table S3). We also examined whether VOC expression before enclosure phase could predict RS, but OPLS models based on the female and both male MS-datasets showed no significant correlations (R^2Y and Q^2 $p > 0.05$).

Male urinary proteins. Male urinary protein excretion in seminatural conditions depended upon social status. Urinary PC ratio (ln transformed) of dominant males significantly increased over time and became higher during the enclosure phase than before (pairwise Tukey post hoc $p < 0.04$; Supplementary Table S4; Fig. 3a). In contrast, the PC ratio of subordinate males did not vary throughout the experiment (post hoc $p > 0.41$). Linear mixed effects (LME) modelling reveals that the factors of social status ($F_{1,76} = 4.3$, $p = 0.04$), time point ($F_{4,76} = 5.3$, $p < 0.001$), and their interaction ($F_{4,76} = 3.3$, $p = 0.01$) all had a significant effect on PC ratio. Age had a marginal effect on PC ratio ($F_{1,76} = 3.26$, $p = 0.07$), but not body mass ($F_{1,76} = 0.5$, $p = 0.47$). Male urinary creatinine concentration (ln transformed), as with PC ratio, changed after release into the enclosures, depending upon social status. Creatinine concentration significantly decreased in dominant male urine during the enclosure phase compared to before (post hoc $p < 0.02$; Supplementary Table S4; Fig. 3b), whereas PC ratio increased. Urinary creatinine concentration of subordinate males did not vary significantly throughout the experiment (post hoc $p > 0.45$). The factors of social status ($F_{1,76} = 5.4$, $p = 0.02$), time point ($F_{4,76} = 4.3$, $p = 0.004$), and their interaction ($F_{4,76} = 3.1$, $p = 0.02$) all had significant effects on urinary creatinine concentration, but not age or body mass. The LME model of total urinary protein concentration showed a significant increase over time for both social status groups ($F_{4,76} = 15.0$, $p < 0.001$; Supplementary Table S4; Fig. 3c), but was not associated with social status, age, or body mass (all $p > 0.12$).

We indirectly measured urinary MUP20 production based on liver RNA transcription 14 days after the enclosure phase; however, the LME model average of hepatic *Mup20* gene expression showed no association with social status, RS, or total urinary protein or creatinine concentration in male mice. Predictor importance ranged from 29 to 14%, suggesting a weak, non-significant correlation between *Mup20* transcription and age (29%, $R = 0.21$, $p = 0.35$), as well as RS (27%, $R = -0.25$, $p = 0.28$; Supplementary Table S1). Social status was the least important predictor of *Mup20* transcription (14%). A similar pattern was observed when the response variable was absolute hepatic *Mup20* transcription. Predictor importance ranged from 25 to 14% with age and RS as the most important (both 25%; Supplementary Table S1) and social status the least.

Male urinary VOCs. We used OPLS models to examine correlations between protein concentration and VOC expression in male urine. Total protein in denatured urine during the enclosures showed a stronger correlation with VOC expression than intact urine both before (denatured: $R^2Y = 0.68$, $Q^2 = 0.63$; intact: $R^2Y = 0.40$, $Q^2 = 0.22$; Fig. 4a) and during the enclosures (denatured: $R^2Y = 0.89$, $Q^2 = 0.62$; intact: $R^2Y = 0.38$, $Q^2 = 0.15$; Fig. 4b). Regardless of urinary protein conformation, HMH peaks correlate with protein concentration of urine collected before the enclosures (intact: Pearson $R = 0.67$, $p_{adj} < 3.8E{-}3$; denatured: $R = 0.77$, $p_{adj} < 0.005$). Other pre-enclosure correlations between VOCs and urinary protein concentration depended on conformation, including SBT from denatured urine ($R = 0.74$, $p_{adj} < 0.005$) and TMA from intact urine ($R = 0.21$, $p_{adj} < 3.8E{-}3$). No peaks correlated with total protein concentration of intact or denatured urine during the enclosures ($p_{adj} > 0.003$).

We tested whether the expression of VOCs in standard conditions predicted male social status during the enclosure phase. The discriminant analysis (OPLS-DA) of VOC expression in denatured urine collected before the enclosure phase did not reliably discriminate males that became dominant during the enclosure phase (Fig. 5a; full MS-data: $R^2Y = 0.5$, $Q^2 = -0.121$, misclassification rate (mcr) = 0.17; candidate MS-data: $R^2Y = 0.311$,

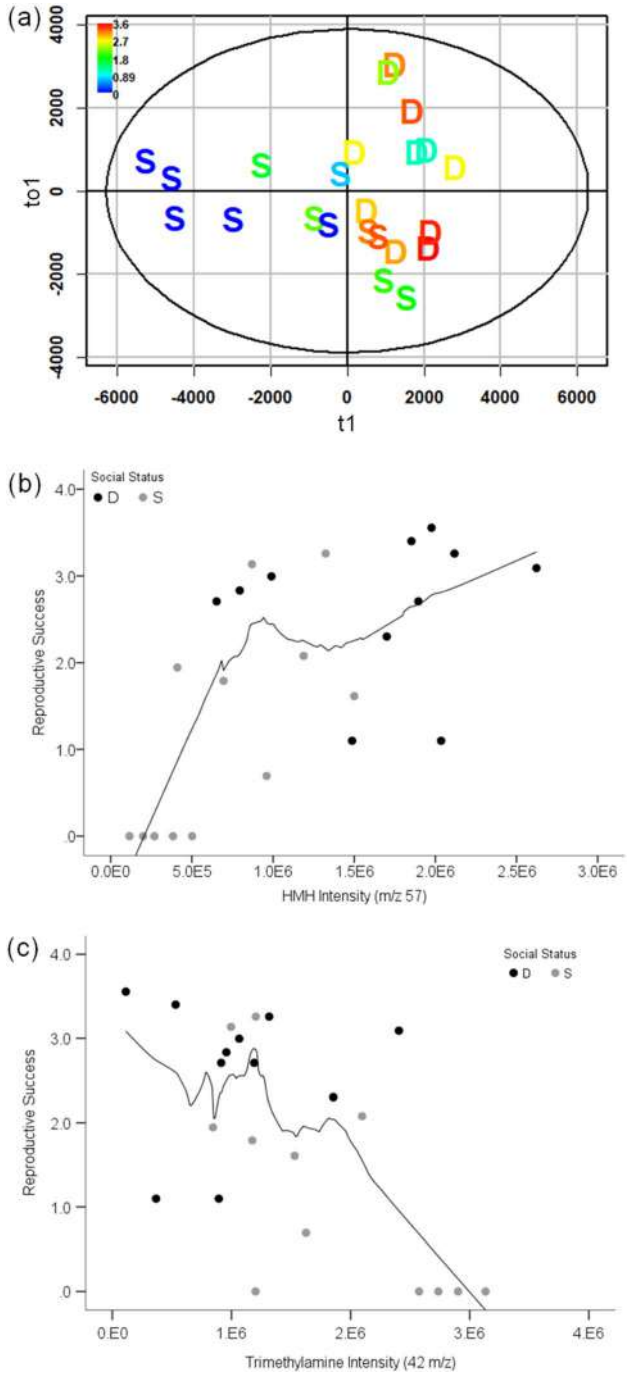

Figure 2. Male reproductive success in relation to VOC expression. OPLS scores plot of reproductive success based on candidate MS-data derived from denatured male urine collected during the enclosure phase (**a**). The x-axis of the scores plot is the predictive component (t1) of the RS response variable. The y-axis is the first orthogonal component (to1). Data points for dominant and subordinates are labeled D and S, respectively. Coloration of the data points indicates the range of RS measured for males; high and low RS range from red to blue, respectively. The Spearman rank correlation of HMH and Trimethylamine expression (**b** and **c**, respectively) with RS shown for intact urine from dominant (black) and subordinate (gray) males during the enclosure phase. The black trend line in the scatterplots shows the loess fit for non-parametric data (50% of data points to fit Epanechnikov kernel).

$Q^2 < -0.01$, mcr = 0.26; Fig. 5b). Furthermore, the VOC peak expression and total ion chromatogram (TIC) intensity of pre-enclosure urine did not significantly differ based on the social status the individual obtained during the enclosure phase (Welch's t-test of TIC: full MS-data p = 0.54; candidate MS-data p = 0.55).

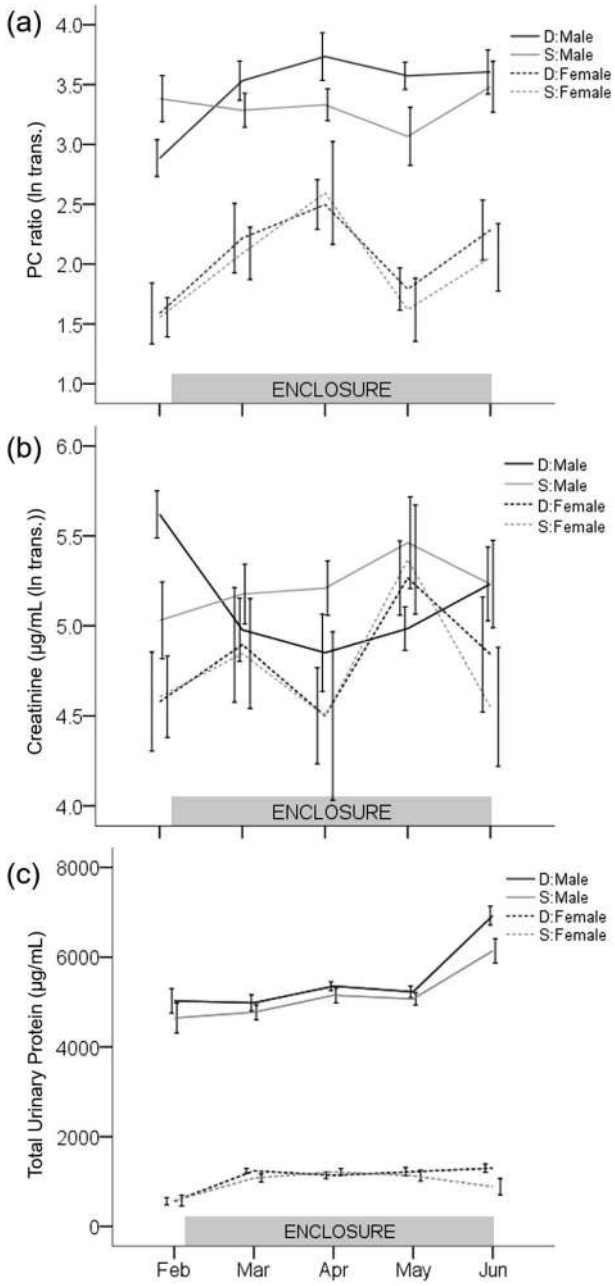

Figure 3. Expression of urinary protein in relation to enclosure phase. Line graphs of PC ratio (ln transformed, **a**), urinary creatinine concentration (µg mL^{-1} (ln transformed, **b**)), and total urinary protein concentration (µg mL^{-1}, **c**). Solid and dashed lines indicate males and females, respectively. Black and gray color indicate dominant (D) and subordinate (S) social status, respectively. Note, February is the before enclosure phase measurement; all other time points were during. Error bars are ± 1 SEM.

There was a strong association between male social status and urinary VOC expression during the enclosure phase. The OPLS-DA of full MS-data showed robust separation of dominant and subordinate males based on VOC expression of denatured urine collected during the enclosures ($R^2Y = 0.79$, $Q^2 = 0.65$, mcr = 0.04; Fig. 5c). The denatured urine model of candidate MS-data also showed separation by social status but to a lesser degree ($R^2Y = 0.62$, $Q^2 = 0.51$, mcr = 0.13; Fig. 5d). The models of intact urine VOC expression also discriminate social status but to a lesser degree than the denatured urine models (intact:full: $R^2Y = 0.69$, $Q^2 = 0.41$, mcr = 0.17; intact:candidate: $R^2Y = 0.55$, $Q^2 = 0.51$, mcr = 0.17; Supplementary Table S3). In models of full MS-data, one peak in intact urine and 88 peaks in denatured urine were upregulated in dominant males. The peaks correspond to HMH in denatured urine (mean difference = 1.2E6, Wilcoxon rank-sum post hoc $p_{adj} < 5.5E{-}4$; Fig. 5e), and 4-methyl-6-hepten-3-one in both urinary protein conformations (denatured: mean difference = 4.6E4, $p_{adj} < 3.5E{-}4$; intact: mean difference = 1.0E4, $p_{adj} < 4.4E{-}4$; Fig. 5f). Details for differentiating the spectra of

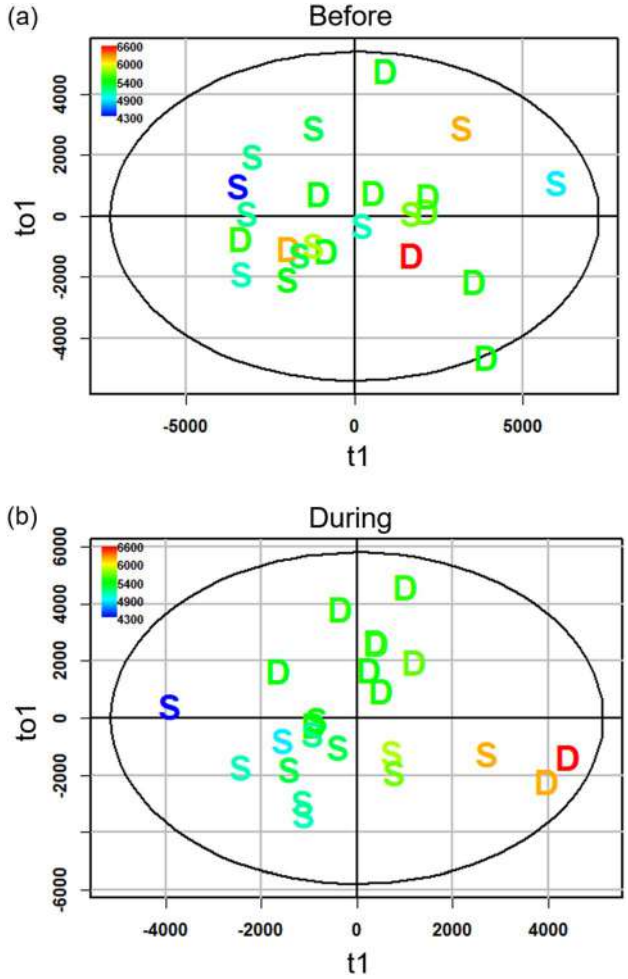

Figure 4. Male VOC expression in relation to urinary protein concentration. OPLS scores plots of total urinary protein concentration based on candidate MS-data derived from denatured male urine (n = 23) collected before (**a**) and during (**b**) the enclosure phase. The x-axis of the scores plot is the predictive component (t1) and the y-axis is the first orthogonal component (to1). Data points for dominant and subordinate males are labelled D and S, respectively. Coloration of the data points indicate the range of urinary protein concentration (μg mL^{-1}); high and low concentration range from red to blue, respectively.

these compounds is in Supplementary Fig. S3. Based on full MS-data, dominant males have a higher TIC intensity than subordinates when comparing denatured urine (mean difference = 1.3E7, p = 0.02), whereas this pattern was not significant for intact urine (mean difference = 7.8E6, p = 0.2). In models of candidate MS-data, peaks that correspond to HMH were upregulated in dominant male intact and denatured urine. The TIC intensity of candidate MS-data did not differ between dominant and subordinate males, regardless of urine conformation (intact urine p = 0.62; denatured urine p = 0.28).

Female urinary proteins. Female mice showed a significant increase in protein excretion (PC ratio) after being released in the enclosures regardless of their social status (Fig. 3a). We observed a significant effect of time point on female PC ratio (LME: PC ratio (ln transformed): $F_{4,75} = 3.3$, p = 0.02; Supplementary Table S4), but not for social status, age, body mass, or status:time point interaction (all p > 0.55). Time point also had a strong effect on the LME model of total urinary protein concentration (Fig. 1c; $F_{4,75} = 9.9$, p < 0.001; Supplementary Table S4). Female mice significantly upregulated total protein concentration and PC ratio during the enclosure phase (Feb-Mar pairwise Tukey post hoc comparison for both D and S p < 0.001; Fig. 3a,c). Age and body mass had a marginal effect on urinary protein concentration in females (age: $F_{1,75} = 3.6$, p = 0.06; mass: $F_{1,75} = 2.8$, p = 0.09), but not social status or status:time point interaction (all p > 0.34). The LME of urinary creatinine concentration (ln transformed) was not significantly affected by the model variables (all p > 0.18; Supplementary Table S4), and although stochastic, mean values did not vary significantly between time points (Fig. 3b).

Female urinary VOCs. Total urinary protein concentration was correlated with VOC expression in denatured female urine, as observed for males but to a lesser extent for female urine (Supplementary Table S3). Total protein concentration of denatured urine collected during the enclosures showed a slightly stronger correlation

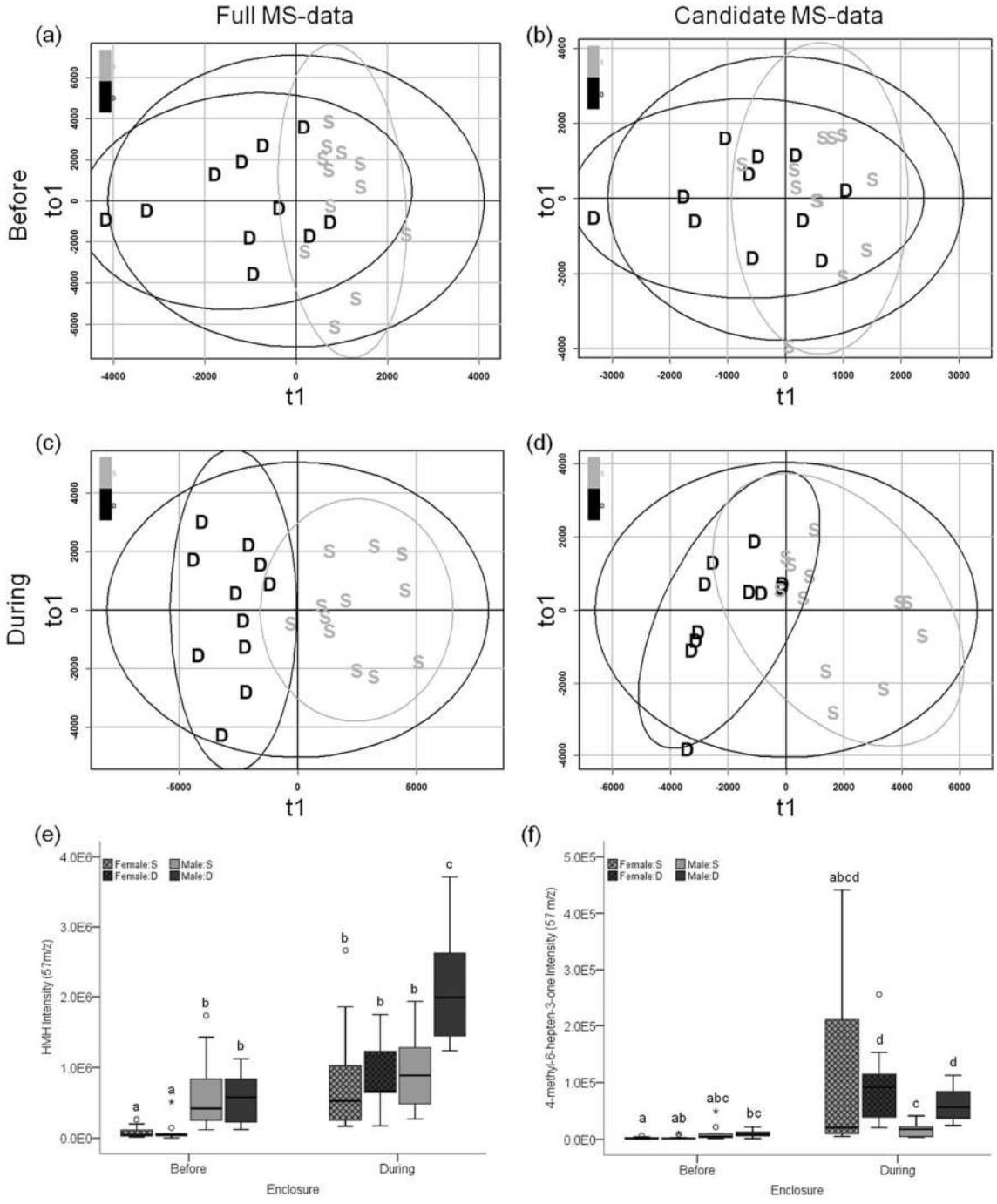

Figure 5. VOC expression in relation to social status. OPLS-DA scores plots of social status before (**a**,**b**) and during the enclosures (**c**,**d**) based on the full MS-data (**a**,**c**) and the candidate MS-data (**b**,**d**) derived from denatured male urine (n = 23).The x-axis of the scores plot is the predictive component (t1) and the y-axis is the first orthogonal component (to1). Data points for dominant (black) and subordinates (light gray) are labelled D and S, respectively. The boxplots show differential expression of HMH and 4-methyl-6-hepten-3-one (**e** and **f**, respectively) in intact urine for dominant (D) and subordinate (S; dark and light gray, respectively) males and females (lattice; n = 24) at both enclosure phases. Different letters above the boxplots denote significant differences.

to VOC expression compared to before the enclosures (before: $R^2Y = 0.68$, $Q^2 = 0.44$ during: $R^2Y = 0.71$, $Q^2 = 0.28$). A positive correlation with total protein concentration was observed for 10 peaks before and 2 peaks during the enclosure phase; the VOC(s) to which the peaks correspond were not identified. The OPLS models of female urine examining intact total protein concentration, or PC ratio and creatinine of both intact and denatured urine did not correlate with VOC expression regardless of enclosure phase ($p > 0.05$; Supplementary Table S3).

Unlike males, VOC expression was not associated with social status in females, regardless of protein conformation and enclosure phase. The OPLS-DA of full MS-data moderately discriminate social status with low predictive ability in denatured female urine ($R^2Y = 0.52$, $Q^2 = 0.37$, mcr = 0.04), and to a lesser extent in intact urine ($R^2Y = 0.47$, $Q^2 = 0.19$, mcr = 0.17). For both intact and denatured urine analyses, there were no significant differences in peak intensity based on social status. The TIC intensity was slightly higher for subordinate females during the enclosures, but this difference was not significant (intact:D mean TIC = 2.5E7, intact:S mean TIC = 3.2E7, p = 0.14; denatured:D mean TIC = 2.3E7, denatured:S mean TIC = 2.8E7, p = 0.36). The OPLS-DA of denatured female urine before enclosure phase was not related to social status ($R^2Y = 0.46$, $Q^2 < 0.01$, mcr = 0.20). There were no expression differences in specific peaks and females that became subordinate during enclosures showed a slightly higher TIC intensity than dominants, though this difference was not significant (before:S mean TIC = 1.3E7, before:D mean TIC = 1.2E7, p = 0.65). With regard to specific female pheromones, the peaks corresponding to 2-heptanone did not correlate with female RS, social status, or urinary protein excretion (R^2Y and Q^2 $p > 0.05$; Supplementary Table S3). Two other female pheromones, IBA and 2,5-dimethylpyrazine, were not detected in any samples.

Sexual dimorphism of chemosensory signals. Total urinary protein concentration and PC ratio increased significantly during the enclosure phase in both sexes (generalized mixed model (GLMM)); protein concentration $X^2 = 28.1$, $\varphi = 0.77$, $p < 0.001$; PC ratio $X^2 = 28.6$, $\varphi = 0.77$, $p < 0.001$; creatinine concentration $X^2 = 4.6$, $\varphi = 0.31$, p = 0.3 (n.s.); Supplementary Table S5). Overall, the mean values of PC ratio and both protein and creatinine concentration were significantly greater for males than females (all $p < 0.001$). There was a significant sex:housing interaction on urinary protein ($X^2 = 43.8$, $\varphi = 0.96$, $p < 0.001$) and creatinine concentration ($X^2 = 9.1$, $\varphi = 0.44$, p = 0.002), and a marginal effect on PC ratio ($X^2 = 3.7$, $\varphi = 0.28$, p = 0.053). The interaction result indicates greater sex differences in protein concentration in standard housing conditions (M:F ratio = 8.5; Supplementary Table S5) compared to seminatural enclosure conditions (M:F ratio = 5). Similarly, the degree of sexual dimorphism in urinary creatinine decreased from before (M:F ratio = 1.7) to during enclosure phase (M:F ratio = 1).

Sexual dimorphism in urinary volatiles was discernible after controlling for protein conformation and enclosure phase. OPLS-DA of intact urine better discriminate the sexes before rather than during enclosures (before: $R^2Y = 0.87$, $Q^2 = 0.62$, mcr = 0.04; during: $R^2Y = 0.82$, $Q^2 = 0.7$, mcr = 0.09; Supplementary Table S3). The expression of 82 peaks representing IT, SBT, TMA, and HMH (Fig. 5e) showed a male bias in pre-enclosure intact urine. During the enclosures, we observed a sex-biased expression of 74 peaks (female:male bias 8:66) in intact urine. Peaks representing TMA and SBT were upregulated in males during the enclosure phase, while females upregulated 4-methyl-6-hepten-3-one, which was also upregulated in the denatured urine of dominant males (Fig. 5f). Male TIC intensity of intact urine was greater than female TIC intensity before (mean difference = 1.7E7, $p < 0.001$) and during the enclosure phase (mean difference = 1.4E7, $p < 0.001$). As observed with urinary protein levels, the sexual dimorphism of intact urine TIC intensity was significantly greater before compared to during the enclosure phase (M:F before = 2.1; M:F during = 1.5; $X^2 = 11.6$, $\varphi = 0.71$, $p < 0.001$; Supplementary Table S5).

OPLS-DA of sexual dimorphism are improved when analyzing denatured versus intact urine. Sex discrimination based on VOC expression of denatured urine is more accurate during than before the enclosure phase (before: $R^2Y = 0.72$, $Q^2 = 0.51$, mcr = 0.04; during: $R^2Y = 0.89$, $Q^2 = 0.84$, mcr = 0.04; Supplementary Table S3). The expression of 88 peaks representing 4-methyl-6-hepten-3-one, HMH, and TMA showed a male bias in pre-enclosure denatured urine. During the enclosures, we observed male-biased expression of 76 peaks, with upregulations of DHB, IT, SBT, and TMA in denatured urine. Male TIC intensity of denatured urine was greater than female TIC intensity before (mean difference = 2.1E7, $p < 0.001$) and during the enclosure phase (mean difference = 2.1E7, $p < 0.001$). Consistent with the intact urine result, the sexual dimorphism of denatured urine TIC intensity significantly decreased during the enclosure phase (M:F before = 2.6; M:F during = 1.8; $X^2 = 7.9$, $\varphi = 0.59$, p = 0.005; Supplementary Table S5).

Discussion

The most important predictor of male reproductive success in the enclosures was urinary protein concentration, which is mainly composed of MUPs[6]. The intensity of HMH, a volatile male pheromone, was also correlated with male RS, and thus, the production of non-volatile and volatile pheromones both correlated with male RS. HMH is unstable during GC analysis and produces several ions with different peaks (Supplementary Fig. S3)[21], but we found that the minor ions of HMH were correlated with male RS as well. The second most important predictor of male RS was social status, and social status was associated with differences in the excretion of both urinary protein and VOCs. Therefore, the effects of pheromone production on RS could have been through direct male-male competition, female mate choice, or both. MUP excretion may have deterred rival males from entering dominant males' territories[22], thereby reducing agonistic interactions and mate-competition. MUP excretion may have attracted females to males' territories, or increased female attraction and sexual receptivity by controlling the release of HMH and other pheromones that influence female reproductive physiology and behavior.

Our findings corroborate results from previous studies on social status in wild-derived mice living in seminatural conditions (i.e., reduced reproductive success of subordinate males[17] and increased urinary protein and HMH pheromone expression in dominant males[24]). Social status did not correlate with body mass, which also confirms results from a previous study in seminatural conditions on mice from this population[24], but contrary to a result on group-housed laboratory strains[35]. Males that obtained a territory substantially increased urinary protein excretion within four weeks after release in the enclosures, whereas subordinate males did not show any changes in protein excretion over time. There were no differences in pheromone production between dominant and subordinate males before their release into the enclosures, confirming that acquisition of dominant social

status influenced pheromone regulation, rather than *vice versa*[24]. The increased protein excretion of dominant males was revealed only after controlling for urine concentration using creatinine levels (PC ratio), and social status had no effect on the total protein concentration. The rate of creatinine production is reportedly consistent for animals of similar body mass[2], yet a considerable drop in creatinine concentration in dominant males was found in the present study and in a study on domesticated male mice in social housing[23]. It is not known whether urinary creatinine is used as a signal of social status. Low creatinine concentrations can indicate that dominant males excrete higher volumes of urine per day[36]; however, we found that males had similar urine volumes regardless of social status.

Although we did not measure daily urine production, dominant males that upregulated the excretion of MUPs and VOCs may have also increased their urinary scent mark deposition in the enclosures. Indeed, previous studies found that dominant males produced more urine[37] and scent marks compared to subordinates[38]; and male scent-marking is correlated with RS when females can select their mates[39]. We also investigated hepatic *Mup20* gene expression of males, as high levels of MUP20 excretion have been found in dominant males[24], but we found no correlations with social status or RS. This negative result is not definitive, however, because males were not sampled until 14 days after terminating the enclosure phase and differences in protein excretion between dominant and subordinate males have been found to disappear after removal from seminatural conditions (≤ 28 days[24]). Nevertheless, this finding supports a previous study showing that that dominant and subordinate males no longer show differences in MUP expression after being removed from competitive conditions[24].

Social status also correlated with the intensity of VOCs in male urine, and analyses of the full MS-data were better at discerning dominants from subordinates than the candidate MS-data. This finding indicates that social status affected the expression of several unidentified VOCs in male urine. Some volatile pheromones (HMH and 4-methyl-6-hepten-3-one) were differentially expressed in the urine of dominant males, but others were not (DHB, SBT, and farnesene). The urine of dominant, territorial males was also found to have higher intensities of HMH than subordinates in a previous study of wild-derived mice in seminatural conditions[24]. HMH is androgen-dependent and a female attractant, but only when combined with DHB and SBT[40]. The expression of DHB, SBT and farnesene were not upregulated in dominant males, and these volatile compounds were excreted by all males (before and during the enclosures). Therefore, it is possible that they help to elicit reproductive receptivity in females when combined with other chemosensory compounds to form a multicomponent pheromone[4]. The signaling functions of 4-methyl-6-hepten-3-one are not well-studied, though it has been found to be upregulated (along with DHB and SBT) in the urine of aged males (15–20 mo), and is preferred by females in olfactory assays over the urine of younger adults (3–8 mo[41]). Furthermore, the VOC expression in male urine does not sufficiently discriminate dominants from subordinates before the mice were released into the enclosure, indicating that social status regulates VOC production, and not vice versa.

Because pheromone production (urinary proteins and VOCs) and social status were both correlated with male RS, we investigated their independent effects. Unexpectedly, we found that male VOC expression in intact urine was correlated with RS of subordinate but not dominant males. This finding is largely influenced by low HMH and high TMA expression from non-reproductive subordinate males, since subordinate sires had expression levels similar to some dominant males. Therefore, subordinate males may increase their ability to attract females via HMH expression, despite being non-territorial. We found that the intensity of TMA was highly elevated in subordinate males and it was negatively correlated with male reproductive success. TMA is abundant in the urine of mice and shows ca. 30-fold higher levels in males than females[42]. This amine is a metabolite of gut microbiota[43], and it is sexually dimorphic because females oxidize it in the liver[11]. Mice show an attraction to urine with normal physiological levels, but an aversion to high TMA concentrations (≥ 1000 mM[11]). TMA is detected by trace amine-associated receptors (TAARs[44,45]), and TAAR5 knockout mice lose their attraction to TMA and to mouse urinary scent[11]. Yet, avoidance of high TMA persists in knockout mice, indicating other receptors are sufficient for aversive responses[46]. TMA has been proposed to influence sex and species-specific recognition in mice[44], and to function as an aversive allomone (*Mus musculus* excrete 1000-fold higher levels than other *Mus* species, and it is repellant to rats[11]). TMA is an indicator of spoilage and putrefaction of dead and decaying animals (it smells like rotten fish to humans), and it is highly aversive to many species[42]. High urinary TMA provides an indicator of parasitic infection (*Schistosoma*[47] and *Cryptosporidium*[48]), which may help explain how females discriminate and prefer the scent of healthy over infected males[49,50]. TMA is also elevated in the urine of distressed mice (under stress restraint[51]). Taken together, our results show that high TMA is associated with subordinate social status and low reproductive success. Studies are now needed to experimentally test whether females avoid males having high levels of urinary TMA.

In contrast to males, there was no correlation between female RS and urinary protein or VOC expression. Furthermore, females showed no correlation between social status and total protein concentration (or PC ratio), confirming one previous study[24] but not another one[31]. There was no correlation between VOC expression and female social status, though we did not detect the urinary pheromones Isobutylamine[28] and 2,5-dimethylpyrazine[1,52]. We confirmed that female RS was correlated with social status[53], and also body mass, but the latter was undoubtedly due to gaining weight during pregnancy (initial mass showed no such correlation and several females were visibly pregnant during urine collections). Based on these findings and our behavioral observations, dominant females may have deterred subordinate females from mating through direct agonistic interactions, rather than pheromonal excretion of estrus-inhibitors or mate-attractants in urine. Notably, we did not definitively measure female reproductive state, and periodic fluctuations of urinary compounds coincide with stage of estrus[27,33] or pregnancy[32,54]. Closely monitoring for such effects in seminatural conditions would increase the frequency of handling the mice, and we opted to minimize disturbances that potentially affect behavior[55].

We confirmed sex differences in urinary protein and volatile excretion of house mice, and also that these well-established sex differences in standard housing are reduced in competitive, seminatural conditions[24]. We confirmed that baseline levels of urinary PC ratio in standard housing conditions applies to subordinate males,

but not to dominant males or females[24]. Furthermore, the degree of sexual dimorphism of VOC expression depended on housing conditions and urinary protein conformation. Our findings suggest that the relatively low variation among males and large sex differences in laboratory studies are artifacts due to artificial conditions.

In contrast to a previous study[24], total protein concentration of intact urine did not correlate with TIC intensity for either sex during the enclosure phase. However, our statistical analyses differed from this previous study, as OPLS models of urinary protein concentration in relation to VOC expression control for sex, housing condition, and protein conformation. Despite our attempts to minimize confounding factors, we observed inconsistent expression of VOCs associated with RS, social status, urinary protein, housing conditions, and sex differences depending on whether the GC–MS data is derived from intact or denatured urine. Disparities in VOC expression are likely due to the affinity of ligands to the MUP binding cavity despite protein denaturation[8]. These results raise additional caveats for results obtained by studying chemical signals of rodents kept in standard housing conditions.

Our results show that the production of specific pheromones correlated with the reproductive success of wild-derived male house mice living in competitive conditions. Males regulated the production of these chemosensory compounds depending upon their social status. To our knowledge, these findings are the first to describe the relationship between pheromone expression and reproductive success in any mammal. Since urinary protein output and HMH intensity correlated with male but not female reproductive success, our findings help to explain the evolution of sexually dimorphic (male-biased) expression of MUPs and HMH in house mice. Female RS was associated with social status based on agonistic interactions, and though we did not find any correlations with specific urinary chemosensory compounds, this does not mean that we can rule them out. More studies are needed to investigate female pheromones and RS. Future studies are also needed to examine the biochemical pathways and neuro-endocrine mechanisms through which males regulate chemosensory signals and experimentally test whether pheromones affect RS under competitive conditions. Furthermore, chemosensory compounds are found in lachrymal, mammary, salivary, and vaginal secretions of mice[56,57], and though it would require invasive sampling, future studies are needed to incorporate more of the emanations that mice use for chemical communication. Our results suggest that the 'normal' or 'baseline' levels of pheromones found in the laboratory are not ecologically relevant and are more indicative of studying mice in cages. Therefore, efforts to understand the mechanisms and functions of chemical signals require studying animals under more natural social contexts.

Materials and methods

Trapping, breeding animals, and standard housing conditions. Experimental animals (N = 48) were virgin F3 offspring of 17 breeding pairs of wild house mice (*Mus musculus musculus*) trapped at seven locations within a 300 m radius in Vienna, Austria (48°13′14″N; 16°17′00″E). Mice were weaned at 21 ± 1 d, separated from siblings at the age of 35 ± 1 d and housed in standard mouse cages (type IIL, 36.5 × 20.5 × 14 cm, Tecniplast, Germany) containing wooden bedding (ABEDD, Austria), a cardboard roll, cotton nestlets© (ABEDD, Austria), and a plastic nest box (Tecniplast, Germany). Water and food (Altromin rodent diet 1324) were provided ad libitum and temperature was maintained at 22 ± 2 °C. Mice were kept on a 12:12 h light:dark cycle with red lights on at 1500. Wild-derived mice in our colony are often aggressive toward same-sex conspecifics when multiply-housed in cages. Thus, all individuals were singly housed from their weaning date until the start of the experiment. We use these descriptions of housing, diet, and light:dark cycle to define standard housing conditions. At weaning, all animals received an ear-punch for individual identification and to obtain tissue for DNA paternity analyses. The Ethical and Animal Welfare Commission at the University of Veterinary Medicine Vienna approved the experimental protocols (permit no. 02/08/97/2013). We confirm that all experiments and animal handling were performed according to the ethical standards and guidelines outlined by the Ethical and Animal Welfare Commission. All reported procedures conformed to the Animal Research Reporting of in vivo Experiments—ARRIVE guidelines[58].

Seminatural enclosure housing conditions. Mice were simultaneously released at the center of one of four indoor seminatural enclosures (Supplementary Fig. S1). Each enclosure consisted of 12 mice (1:1 sex ratio) that were sexually mature at the start of the experiment (mean ± s.d. age of males = 134 d ± 28, females = 133 d ± 26). Mice within an enclosure were non-siblings and males were matched for body mass (maximum difference = 0.9 g). Each enclosure (4 × 9 m) was subdivided into eight compartments (wire-mesh fencing, 40-cm high), which mice could scale, but tended to use as territorial boundaries. Wooden bedding (ABEDD, Austria), plastic nest boxes, water stations, wood wool, and nesting material (paper towels) were provided. Food (Altromin rodent diet 1324) and water were provided ad libitum and temperature was maintained at 22 ± 2 °C. Mice lived in the enclosures for 16 weeks between February–June 2016 with a light:dark cycle, as described for standard housing conditions.

Behavioral observations. To assess social status, behavioral observations of the mice were conducted three to five days per week for 30 min/day between 1500 and 1700 during the entire period (241 h total observation time, 60.3 h mean time per enclosure). Males received unique fur cuts before release, facilitating identification under red light; females were identified by their unique ear punches. Observers monitored the behavior of mice through observation windows from adjacent rooms to avoid disturbance. They recorded interactions (classified as aggressive, submissive, and investigatory behaviors), the location of the interaction, and the individuals involved. A dominance index for each individual was calculated by dividing the number of aggressive and investigatory interactions by the total number of interactions[17] (depicted in Eq. 1).

$$Dominance\ Index = \frac{Aggressive + Investigatory}{Submissive + Aggressive + Investigatory} \quad (1)$$

Mice obtaining a dominance index ≥ 80% within an enclosure compartment were considered to be dominant; otherwise they were considered subordinate. Adult survival was monitored daily and offspring born in the enclosures were removed upon discovery. Offspring tissue was collected for paternity analyses.

Urine sampling. For monitoring urinary protein and volatile pheromone production, we conducted five urine collection events from each enclosure population over 16 weeks, at 4-week intervals (outlined in Supplementary Fig. S1). The first collection event was conducted immediately prior to the release of mice into the enclosures, while they were still kept in standard housing conditions ('before enclosure phase' sample). Four additional collection events occurred while mice were living in seminatural conditions ('during enclosure phase'). Urine was collected from mice in metabolic cages (Techniplast, Germany), which minimizes handling stress and fecal contamination. All collections were conducted under red light at the beginning of the dark cycle in the enclosures. Upon excreting > 70 µL of urine, mice were put into individual cages and then simultaneously released back into the enclosure (60 min duration for the entire collection event). Only 2 out of 227 urine collections provided an insufficient volume during the sampling periods. Urine and feces were transferred to separate Eppendorf tubes, immediately frozen, and stored at − 80 °C; handling was the same for each sample to avoid possible freezing and storage biases. An aliquot of urine was collected in a glass GC vial for GC–MS analyses and stored as described.

Urinary protein measurements. As previously described[24], total urinary protein concentration (µg mL^{-1}) was measured in triplicates using a standard Bradford assay on a 96-well microplate[59]. Triplet values not within a ± 10% range were repeated. We adjusted total urinary protein concentration with creatinine concentration to calculate urinary protein excretion (PC ratio), as this value is expected to account for renal activity and urine dilution (creatinine measured by InVitro: Labor für Veterinärmedizinische Diagnostik & Hygiene GmbH, Vienna, Austria). Some studies report 99% of urinary protein is composed of MUPs[60], in our wild-derived mice, c.85% of the total protein in urine consists of MUPs[6], yet proteoform expression varies depending on age[61] and social conditions[24]. In accordance with other studies on urinary protein[19,23,62], we report PC ratio, total protein concentration (unadjusted values), and creatinine concentration (µg mL^{-1}) as response variables in separate models.

GC–MS analysis of volatile molecules. Urine samples of 23 males (11 dominant, 12 subordinate) and 24 females (9 dominant, 15 subordinate) were obtained before and during the enclosure phase and analyzed with GC–MS as previously described[63]. The 'during enclosure' sample was a pool containing an individual's urine from all collection events while living in the enclosures (outlined in Supplementary Fig. S1). We examined intact urine and denatured urine (15 µL each) because VOC expression in the headspace has been shown to change depending on urinary protein conformation[64]. Pheromonal MUP ligands are elevated after denaturing protein in the urine, as expected, and we denatured urine protein with guanadine hydrochloride (GdmCl) to assess the intensity of these MUP-bound pheromones[63,64]. We are not aware of any alternative approaches for investigating VOCs bound to MUPs (and we are not aware of any volatile artifacts derived from GdmCl). An aliquot of intact urine was denatured with 20 mg of GdmCl (product # G3272, Sigma-Aldrich, Vienna, Austria) in a 4 mL glass vial. The vial was submerged in a water bath at 37 °C and was equilibrated for 10–15 min. The VOCs in the headspace of the sample were extracted by a 2 cm three-component solid phase microextraction (SPME) fiber (30 µm carboxen, 50 µm divinyl-benzene, polydimethylsiloxane; Supelco Corp., Bellefonte, PA, USA) for 15 min at 37 °C. The urine sample in the vial was agitated using a magnetic stirrer during the equilibration period, but not in the extraction period. The SPME fiber containing the adsorbed compounds was then inserted into the injection port of the GC–MS instrument and desorbed for 1 min at 240 °C. A Supelcowax 10 GC column (30 m × 0.25 mm with 0.50 µm film thickness; Sigma-Aldrich, Vienna, Austria) coupled with a Shimadzu GC–MS QP2010 Plus (Duisburg, Germany) were used to analyze the VOCs collected via SPME. The oven temperature was held at 40 °C for 1 min, then programmed at 6 °C/min to 220 °C with a 9-min hold at this final temperature. The carrier gas was helium at a 38.9 cm/sec linear velocity. The transfer line temperature between GC and MS was 250 °C. Operating parameters for the mass spectrometer were as follows: ion source temperature at 200 °C; electron impact ionization (70 eV); and scanning frequency was 4/s from m/z 41 to m/z 300.

Raw GC–MS spectral files were converted to a registry of "peaks" defined as a single ion (mass/charge or m/z) at a specific GC retention time as well as the intensity of that ion for each sample as previously described[64]. Statistical analyses of differential compound expression were performed on the peak registry dataset (N = 1079 peaks), henceforth referred to as the "full MS-data". Second, we took a candidate approach by focusing on peaks associated with the following male signaling pheromones: (1) 3,4-dehydro-exo-brevicomin (DHB), (2) farnesene, (3) 4-methyl-6-hepten-3-one, (4) 2-isopropyl-4,5-dihydrothiazole (IT), (5) 2-s-butyl-4,5-dihydrothiazole (SBT), (6) trimethylamine (TMA), and (7) 6-hydroxy-6-methyl-3-heptanone (HMH). Peaks were identified after comparison to a mass spectral library (NIST08) combined with manual interpretation. Our GC–MS method accounts for co-elution of HMH with other volatiles, an issue that has been observed for nonpolar SGE columns[65]. The candidate approach yields a peak registry dataset (N = 39 peaks) referred to by the authors as the "candidate MS-data". The seven listed pheromones correspond to 39 peaks because the fragmentation of a single molecule during mass spectrometry creates multiple ions (peaks), which are then quantified (e.g. 3 peaks correspond to TMA). The analyses of full and candidate MS-data were largely congruent for male urine. We report the results of candidate MS-data in the results section unless otherwise noted; analyses of full MS-data are detailed in

Supplementary Table S3. Notably, there is no method for determining the actual total amounts of MUP ligands (or relative intensities), as some unknown portion still remains bound to denatured MUPs even after repeated extractions[8]. Consequently, negative results for MUP ligands should be treated with caution, especially for urine samples with high MUP concentration.

***Mup20* gene expression and genetic paternity analyses.** We used ddPCR to quantify *Mup20* gene expression in the hepatic tissue (henceforth, italicized *Mup20* refers to nucleic acid molecules whereas MUP20 refers to protein). Due to logistical issues, 14 days elapsed between termination of the enclosure phase and euthanization of the mice. Upon euthanization, the liver was removed and immediately immersed in RNAlater (Qiagen) for 24 h before storage at − 80 °C. The RNA from hepatic tissue (c.25 mg) was extracted with RNeasy Mini Kit (Qiagen) following the manufacturer's instructions, and concentration was measured with a NanoDrop (Thermo Scientific). Between 1 and 1.5 mg of RNA was reverse-transcribed to cDNA using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). The QX200™ Droplet Digital™ PCR system (Bio-Rad) was used to quantify *Mup20* transcripts (247 bp) relative to a non-variable copy number reference gene (*c-myc*, 82 bp, Accession no. NM_001177354). The 20 µL ddPCR mixture was composed of: (1) 10 µL ddPCR supermix for probes (Bio-Rad), (2) 50 µM *Mup20* forward and reverse primers, (3) 10 µM *Mup20* probe, (4) 50 µM *c-myc* forward and reverse primers, (5) 10 µM *c-myc* probe (see Supplementary Table S6 for primer and probe sequences), and 6) 2 µL of cDNA sample. The mixture and 70 µL of droplet generation oil were placed into the DG8 cartridge and inserted to the droplet generator (Bio-Rad) for droplet formation. Droplets were then transferred to a 96-well PCR plate (Eppendorf). The thermal profile for PCR amplification was an initial denaturation at 95 °C for 10 min, followed by 40 cycles of 94 °C for 30 s and 57 °C for 60 s, 1 cycle of 98 °C for 10 min, and ending at 4 °C; reaction performed by C1000 thermal cycler (Bio-Rad). After amplification, the plate was loaded on the droplet reader (Bio-Rad) and quantified. A no-template reaction was included to control for possible reagent contamination. The ratio of *Mup20* positive droplets to *c-myc* positive droplets was used to calculate *Mup20* gene expression. We also measured absolute gene expression by dividing *Mup20* copy number by amount of RNA (ng) used for reverse-transcription. We did not conduct genetic analyses to assess variation in MUPs because a previous survey on this population of mice detected no individual variation in *Mup* gene sequences and low microsatellite diversity throughout the entire *Mup* gene cluster[66].

For genetic paternity analyses, DNA was extracted from ear punch samples using a proteinase K/isopropanol protocol[67] and individuals were genotyped at a minimum of nine and a maximum of 14 microsatellite loci (see Supplementary Table S6 for primer sequences). Amplification by PCR and scoring were done as previously reported[39]. Paternity results were confirmed with a ≥ 95% trio confidence (dam-sire-offspring relationship) using the program CERVUS 3.0.3[68].

Statistical analyses. Statistical analyses were performed using the R statistical package[69] (R Development Core Team version 4.0) and the assumptions of the methods used were first verified.

Reproductive success (RS). The relationship between urinary protein excretion and RS was analyzed with a LME model for each sex (*lme* function, package *nlme*[70]). RS was calculated using a natural log transformation of (1 + the number of offspring produced by an individual). Using RS instead of total number of offspring as a measure of fitness provides a Gaussian distribution of the model residuals and accounts for individuals with zero values (no offspring). Social status and the mean values of total urinary protein concentration, creatinine, PC ratio, age, and body mass during the enclosure phase were included as variables in the model with enclosure as a random factor. Due to collinearity in the models, we calculated the VIF for each variable to determine the level of inter-correlation between them. We removed variables sequentially from the model and recalculated VIFs until all values were < 3[71]. Post hoc analyses were performed with Welch's t-test for categorical variables, or Spearman rank correlation test for continuous variables.

Due to the number of model variables and uncertainty in model selection, a multi-model inference procedure[72] was used to estimate the relative explanatory importance of each variable. Models were ranked based on Akaike's Information Criterion value corrected for small sample sizes (AICc) and coefficients for each variable were calculated based on weighted estimates of retained models (the condition for retaining a model was ΔAICc < 10). The relative importance of a variable was established by summing the Akaike weights of each model in which it was a predictor (i.e. variables with high importance are included in models with low AICc). This multi-model inference procedure was applied to LME models in which RS and *Mup20* gene expression are response variables, allowing us to simultaneously test their multiple potential relationships with age, size, social status, and chemical signaling. We refer to this procedure as model averaging.

Urinary protein excretion. Three LME models for each sex were used to examine the relationship between urinary protein excretion (PC ratio, total protein concentration, and creatinine concentration) and social status. Model averaging was not performed because fixed factors were inferred from a previous study[24]. PC ratios and creatinine values were natural log transformed to obtain a Gaussian curve of the model residuals. Social status, time point (urine collection event), and the interaction of social status:time point were used as fixed factors; age and body mass were covariates in each model. Since we repeatedly sampled individuals, we include a random factor of individual ID nested in enclosure. We used the *varIdent* function to account for heteroscedasticity in the social status factor and fit the LME model using the maximum likelihood method. The *anova.lme* function produced F and p values for testing significance of model variables. Pairwise comparison post hoc tests were performed for the interaction social status:time point in each model using the *glht* function with Tukey correction for multiple testing (package *multcomp*[73]).

Sex differences in urinary protein excretion were examined with GLMMs (*glmer* function, package *lme4*[74]) due to non-parametric data comprising repeated measures of sexually dimorphic traits. Sex, housing (standard or enclosure), and the sex:housing interaction as fixed effects on the response variables of PC ratio, total protein concentration, and creatinine concentration. A previous study found a correlation between urinary protein concentration and TIC intensity[24]. Therefore, we included the TIC intensity of intact and denatured urine samples as response variables for a total of five GLMMs. Individual ID nested in enclosure were included as random factors. An inverse link function was used to account for the inverse Gaussian distribution of the urinary protein variables. Wald chi-square tests were used to determine significance of the model effects (function *anova*, package *stats*[69]).

Gene expression of Mup20. LME models were used to examine the relationship between hepatic expression of *Mup20* and urinary protein excretion, social status, and reproduction. Social status, RS, and the mean values of total urinary protein concentration, creatinine, PC ratio, age, and mass during the enclosure phase were included as variables in the model with enclosure as a random factor. We used *Mup20* gene expression and absolute values (based on initial RNA amount) as response variables in separate models. Corrections for collinearity, model averaging, and post hoc analyses were performed as described for RS.

VOC expression and identification of differentiating compounds. We conducted OPLS-DA models on the full MS-data with either sex, enclosure phase, or social status as the categorical covariate. The continuous covariates of reproductive success, total urinary protein concentration, creatinine (ln transformed), and PC ratio (ln transformed) were analyzed with OPLS models. Additional models using candidate MS-data were performed on male GC–MS data only. The relationship between MS-data and a given covariate was analyzed using *opls* function in the package *ropls*[75]. The supervised OPLS-DA performs a sevenfold cross validation based on the latent components. This allows us to visualize groups of mice based on the regression of latent components to the covariate. The variance explained by the model (R^2Y) and the predictive ability of the model based on cross validation (Q^2) describe the relationship of the covariate to the MS-data. This method also calculates a value of importance to the projection (VIP) for each peak. A large VIP indicates a strong association between a peak and the model covariate. For OPLS-DA, Wilcoxon rank-sum tests with a Bonferroni adjusted p-value due to multiple comparisons (alpha = 0.05) were used to compare the peaks important for the discriminant analysis (VIP > 1) between the two classes of the covariate. Pearson or Spearman rank (non-Gaussian) correlation tests with Bonferroni adjusted p-values were used to compare the association between the continuous covariate with important VOC peaks (VIP > 1) derived from the OPLS model. Since multiple peaks can relate to a single VOC, we provide the results of the greatest VIP peak associated with the male pheromones in the candidate MS-data. The R^2Y and Q^2 model coefficients reported in the results section have a corresponding alpha value of 0.05 unless otherwise noted.

Data availability

Upon publication, datasets will be available on Mendeley Data repository or by contacting the corresponding authors.

Received: 1 April 2021; Accepted: 9 September 2021
Published online: 09 November 2021

References

1. Novotny, M. V., Ma, W., Žídek, L. & Daev, E. Recent biochemical insights into puberty acceleration, estrus induction, and puberty delay in the house mouse. *Adv. Chem. Signals Vertebr.* https://doi.org/10.1007/978-1-4615-4733-4_7 (1999).
2. Beynon, R. J. & Hurst, J. L. Urinary proteins and the modulation of chemical scents in mice and rats. *Peptides* **25**, 1553–1563 (2004).
3. Stopka, P., Stopková, R. & Janotová, K. Mechanisms of chemical communication. In *Evolution of the House Mouse* (eds Macholán, M. *et al.*) 191–220 (Cambridge University Press, 2012).
4. Wyatt, T. *Pheromones and Animal Behavior: Chemical Signals and Signatures* (Cambridge University Press, 2014).
5. Mucignat-Caretta, C. & Caretta, A. Message in a bottle: Major urinary proteins and their multiple roles in mouse intraspecific chemical communication. *Anim. Behav.* **97**, 255–263 (2014).
6. Enk, V. M. *et al.* Regulation of highly homologous major urinary proteins in house mice quantified with label-free proteomic methods. *Mol. Biosyst.* **12**, 3005–3016 (2016).
7. Hurst, J. L., Robertson, D. H. L., Tolladay, U. & Beynon, R. J. Proteins in urine scent marks of male house mice extend the longevity of olfactory signals. *Anim. Behav.* **55**, 1289–1297 (1998).
8. Kwak, J. *et al.* Changes in volatile compounds of mouse urine as it ages: Their interactions with water and urinary proteins. *Physiol. Behav.* **120**, 211–219 (2013).
9. Jemioło, B., Alberts, J., Sochinski-Wiggins, S., Harvey, S. & Novotny, M. Behavioural and endocrine responses of female mice to synthetic analogues of volatile compounds in male urine. *Anim. Behav.* **33**, 1114–1118 (1985).
10. Novotny, M. V., Harvey, S. & Jemioło, B. Chemistry of male dominance in the house mouse, *Mus domesticus*. *Experientia* **46**, 109–113 (1990).
11. Li, Q. *et al.* Synchronous evolution of an odor biosynthesis pathway and behavioral response. *Curr. Biol.* **23**, 11–20 (2013).
12. Roberts, S. A., Davidson, A. J., McLean, L., Beynon, R. J. & Hurst, J. L. Pheromonal induction of spatial learning in mice. *Science* **338**, 1462–1465 (2012).
13. Dey, S. *et al.* Cyclic regulation of sensory perception by a female hormone alters behavior. *Cell* **161**, 1334–1344 (2015).
14. Jouhanneau, M. & Keller, M. Acceleration of puberty onset in female mice by male urinary odours: chemosensory, physiological and neuroendocrine mechanisms. In *Chemical Signals in Vertebrates*, Vol. 12, 17–36 (Springer New York, 2013). https://doi.org/10.1007/978-1-4614-5927-9_2.
15. Flanagan, K. A., Webb, W. & Stowers, L. Analysis of male pheromones that accelerate female reproductive organ development. *PLoS ONE* **6**, e16660 (2011).

16. Hurst, J. L. *et al.* Information in scent signals of competitive social status: the interface between behaviour and chemistry. In *Chemical Signals in Vertebrates*, Vol. 9, 43–52 (Springer US, 2001). https://doi.org/10.1007/978-1-4615-0671-3_6.
17. Meagher, S., Penn, D. J. & Potts, W. K. Male–male competition magnifies inbreeding depression in wild house mice. *Proc. Natl. Acad. Sci.* **97**, 3324–3329 (2000).
18. Guo, H., Fang, Q., Huo, Y., Zhang, Y. & Zhang, J. Social dominance-related major urinary proteins and the regulatory mechanism in mice. *Integr. Zool.* **10**, 543–554 (2015).
19. Janotová, K. & Stopka, P. The level of major urinary proteins is socially regulated in wild *Mus musculus* musculus. *J. Chem. Ecol.* **37**, 647–656 (2011).
20. Nevison, C. M., Barnard, C. J., Beynon, R. J. & Hurst, J. L. The consequences of inbreeding for recognizing competitors. *Proc. R. Soc. B Biol. Sci.* **267**, 687–694 (2000).
21. Harvey, S., Jemioło, B. & Novotny, M. Pattern of volatile compounds in dominant and subordinate male mouse urine. *J. Chem. Ecol.* **15**, 2061–2072 (1989).
22. Jones, R. B. & Nowell, N. W. A comparison of the aversive and female attractant properties of urine from dominant and subordinate male mice. *Anim. Learn. Behav.* **2**, 141–144 (1974).
23. Lee, W., Khan, A. & Curley, J. P. Major urinary protein levels are associated with social status and context in mouse social hierarchies. *Proc. R. Soc. B Biol. Sci.* **284**, 20171570 (2017).
24. Thoß, M. *et al.* Regulation of volatile and non-volatile pheromone attractants depends upon male social status. *Sci. Rep.* **9**, 489 (2019).
25. Perez-Miller, S., Zou, Q., Novotny, M. V. & Hurley, T. D. High resolution X-ray structures of mouse major urinary protein nasal isoform in complex with pheromones. *Protein Sci.* **19**, 1469–1479 (2010).
26. Nishimura, K., Utsumi, K., Yuhara, M., Fujitani, Y. & Iritani, A. Identification of puberty-accelerating pheromones in male mouse urine. *J. Exp. Zool.* **251**, 300–305 (1989).
27. Schwende, F. J., Wiesler, D. & Novotny, M. Volatile compounds associated with estrus in mouse urine: Potential pheromones. *Experientia* **40**, 213–215 (1984).
28. Harmeier, A. *et al.* How female mice attract males: A urinary volatile amine activates a trace amine-associated receptor that induces male sexual interest. *Front. Pharmacol.* **9**, 924 (2018).
29. Jemioło, B. & Novotny, M. Long-term effect of a urinary chemosignal on reproductive fitness in female mice. *Biol. Reprod.* **48**, 926–929 (1993).
30. Novotny, M., Jemioło, B., Harvey, S., Wiesler, D. & Marchlewska-Koj, A. Adrenal-mediated endogenous metabolites inhibit puberty in female mice. *Science* **231**, 722–725 (1986).
31. Stockley, P., Bottell, L. & Hurst, J. L. Wake up and smell the conflict: Odour signals in female competition. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20130082 (2013).
32. Fujii, Y., Nagumo, A., Hasegawa, M. & Nagasawa, H. Change of urinary component level during pregnancy and lactation in mice. *J. Reprod. Dev.* **39**, 187–191 (1993).
33. Stopka, P., Janotova, K. & Heyrovsky, D. The advertisement role of major urinary proteins in mice. *Physiol. Behav.* **91**, 667–670 (2007).
34. Whitten, W. K. Modification of the oestrous cycle of the mouse by external stimuli associated with the male. *J. Endocrinol.* **17**, 307–313 (1958).
35. DeFries, J. C. & McClearn, G. E. Social dominance and darwinian fitness in the laboratory mouse. *Am. Nat.* **104**, 408–411 (1970).
36. Drickamer, L. C. Rates of urine excretion by house mouse (*Mus domesticus*): Differences by age, sex, social status, and reproductive condition. *J. Chem. Ecol.* **21**, 1481–1493 (1995).
37. Drickamer, L. C. Urine marking and social dominance in male house mice (*Mus musculus domesticus*). *Behav. Processes* **53**, 113–120 (2001).
38. Desjardins, C., Maruniak, J. A. & Bronson, F. H. Social rank in house mice: Differentiation revealed by ultraviolet visualization of urinary marking patterns. *Science* **182**, 939–941 (1973).
39. Thonhauser, K. E., Raveh, S., Hettyey, A., Beissmann, H. & Penn, D. J. Scent marking increases male reproductive success in wild house mice. *Anim. Behav.* **86**, 1013–1021 (2013).
40. Jouhanneau, M. *et al.* Peripubertal exposure to male chemosignals accelerates vaginal opening and induces male-directed odor preference in female mice. *Front. Ecol. Evol.* **3**, 34 (2015).
41. Osada, K., Tashiro, T., Mori, K. & Izumi, H. The identification of attractive volatiles in aged male mouse urine. *Chem. Senses* **33**, 815–823 (2008).
42. Dewan, A. Olfactory signaling via trace amine-associated receptors. *Cell Tissue Res.* **383**, 1–13 (2020).
43. Al-Waiz, M., Mikov, M., Mitchell, S. C. & Smith, R. L. The exogenous origin of trimethylamine in the mouse. *Metabolism* **41**, 135–136 (1992).
44. Liberles, S. D. & Buck, L. B. A second class of chemosensory receptors in the olfactory epithelium. *Nature* **442**, 645–650 (2006).
45. Spehr, M. & Munger, S. D. Olfactory receptors: G protein-coupled receptors and beyond. *J. Neurochem.* **109**, 1570–1583 (2009).
46. Liberles, S. D. Trace amine-associated receptors: Ligands, neural circuits, and behaviors. *Curr. Opin. Neurobiol.* **34**, 1–7 (2015).
47. Wang, Y. *et al.* Metabonomic investigations in mice infected with Schistosoma mansoni: An approach for biomarker identification. *Proc. Natl. Acad. Sci.* **101**, 12676–12681 (2004).
48. Bolick, D. T. *et al.* Increased urinary trimethylamine N-oxide following cryptosporidium infection and protein malnutrition independent of microbiome effects. *J. Infect. Dis.* **216**, 64–71 (2017).
49. Penn, D., Schneider, G., White, K., Slev, P. & Potts, W. Influenza infection neutralizes the attractiveness of male odour to female mice (*Mus musculus*). *Ethology* **104**, 685–694 (1998).
50. Zala, S. M. Scent-marking displays provide honest signals of health and infection. *Behav. Ecol.* **15**, 338–344 (2004).
51. Schaefer, M. L. *et al.* Mouse urinary biomarkers provide signatures of maturation, diet, stress level, and diurnal rhythm. *Chem. Senses* **35**, 459–471 (2010).
52. Andreolini, F., Jemioło, B. & Novotny, M. Dynamics of excretion of urinary chemosignals in the house mouse (*Mus musclus*) during the natural estrous cycle. *Experientia* **43**, 998–1002 (1987).
53. Hurst, J. L. Behavioural variation in wild house mice *Mus domesticus* Rutty: A quantitative assessment of female social organization. *Anim. Behav.* **35**, 1846–1857 (1987).
54. Jemioło, B., Andreolini, F., Wiesler, D. & Novotny, M. Variations in mouse (*Mus musculus*) urinary volatiles during different periods of pregnancy and lactation. *J. Chem. Ecol.* **13**, 1941–1956 (1987).
55. Gariépy, J. L., Rodriguiz, R. M. & Jones, B. C. Handling, genetic and housing effects on the mouse stress system, dopamine function, and behavior. *Pharmacol. Biochem. Behav.* **73**, 7–17 (2002).
56. Shahan, K., Denaro, M., Gilmartin, M., Shi, Y. & Derman, E. Expression of six mouse major urinary protein genes in the mammary, parotid, sublingual, submaxillary, and lachrymal glands and in the liver. *Mol. Cell. Biol.* **7**, 1947–1954 (1987).
57. Černá, M., Kuntová, B., Talacko, P., Stopková, R. & Stopka, P. Differential regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the house mouse. *Sci. Rep.* **7**, 11674 (2017).
58. du Sert, N. P. *et al.* Reporting animal research: Explanation and elaboration for the arrive guidelines 2.0. *PLoS Biol.* **18**, e3000411 (2020).

59. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).
60. Humphries, R. E., Robertson, D. H. L., Beynon, R. J. & Hurst, J. L. Unravelling the chemical basis of competitive scent marking in house mice. *Anim. Behav.* **58**, 1177–1190 (1999).
61. Thoß, M., Luzynski, K. C., Ante, M., Miller, I. & Penn, D. J. Major urinary protein (MUP) profiles show dynamic changes rather than individual 'barcode' signatures. *Front. Ecol. Evol.* **3**, 71 (2015).
62. Nelson, A. C., Cunningham, C. B., Ruff, J. S. & Potts, W. K. Protein pheromone expression levels predict and respond to the formation of social dominance networks. *J. Evol. Biol.* **28**, 1213–1224 (2015).
63. Kwak, J., Strasser, E., Luzynski, K., Thoß, M. & Penn, D. J. Are MUPs a toxic waste disposal system?. *PLoS ONE* **11**, e0151474 (2016).
64. Kwak, J. *et al.* Differential binding between volatile ligands and major urinary proteins due to genetic variation in mice. *Physiol. Behav.* **107**, 112–120 (2012).
65. Lin, D. Y., Zhang, S.-Z., Block, E. & Katz, L. C. Encoding social signals in the mouse main olfactory bulb. *Nature* **434**, 470–477 (2005).
66. Thoß, M. *et al.* Diversity of major urinary proteins (MUPs) in wild house mice. *Sci. Rep.* **6**, 1–9 (2016).
67. Sambrook, J., Fritsch, E. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (1989).
68. Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7**, 639–655 (1998).
69. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
70. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models* (2019).
71. Zuur, A. F., Ieno, E. N. & Elphick, C. S. A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1**, 3–14 (2010).
72. Burnham, K. & Anderson, D. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2002).
73. Hothorn, T., Bretz, F., Westfall, P. & Heiberger, R. M. Multcomp: simultaneous inference for general linear hypotheses. *R Packag. version* 0–1 (2008).
74. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–51 (2015).
75. Thévenot, E. A. T., Roux, A., Xu, Y., Ezan, E. & Junot, C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *ACS Publ.* **14**, 3322–3335 (2015).

Acknowledgements

We are extremely grateful to T. Klaus, E. Strasser, B. Wernisch, K. Auer, A. Desvars-Larrive, and M. Grijincu for excellent support in conducting experiments. We thank R. Sasse and A. Sasse for enclosure construction and animal care, and M. Thoß for colony management. I. Miller, M. Hämmerle, and S. Smith provided assistance with molecular analyses. A. Arthyukin and T. Wyatt provided helpful comments on the manuscript, and we thank the editor and two anonymous reviewers for their comments. This work was funded by the Austrian Science Fund (FWF) to DP: P24711-B21 and P28141-B25. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

D.P. conceived the idea for the study; D.P., K.L., D.N., M.A.M., S.Z., and J.K. designed or established the methods; K.L., D.N., and M.A.M. collected the data; K.L. and J.K. analyzed the data; K.L. and D.P. wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-01507-9.

Correspondence and requests for materials should be addressed to K.C.L. or D.J.P.

CHAPTER 2: Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female

Zala, Sarah M.; Nicolakis, Doris; Marconi, Maria Adelaide; Noll, Anton; Ruf Thomas; Balazs, Peter; Penn, Dustin J.

PLoS ONE, impact factor: 3.7; **published**

PLOS ONE

RESEARCH ARTICLE

Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female

Sarah M. Zala[1]*, Doris Nicolakis[1], Maria Adelaide Marconi[1], Anton Noll[2], Thomas Ruf[3], Peter Balazs[2], Dustin J. Penn[1]

1 Konrad Lorenz Institute of Ethology, University of Veterinary Medicine, Vienna, Austria, **2** Acoustic Research Institute, Austrian Academy of Sciences, Vienna, Austria, **3** Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, Austria

* sarah.zala@vetmeduni.ac.at

OPEN ACCESS

Citation: Zala SM, Nicolakis D, Marconi MA, Noll A, Ruf T, Balazs P, et al. (2020) Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female. PLoS ONE 15(12): e0242959. https://doi.org/10.1371/journal.pone.0242959

Editor: Claudia Mettke-Hofmann, Liverpool John Moores University, UNITED KINGDOM

Received: July 8, 2019

Accepted: November 13, 2020

Published: December 9, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: This research was supported by a grant (FWF P 28141-B25) of the Austrian Science Foundation (http://www.fwf.ac.at) to DJP and SMZ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Males in a wide variety of taxa, including insects, birds and mammals, produce vocalizations to attract females. Male house mice emit ultrasonic vocalizations (USVs), especially during courtship and mating, which are surprising complex. It is often suggested that male mice vocalize at higher rates after interacting with a female, but the evidence is mixed depending upon the strain of mice. We conducted a study with wild-derived house mice (*Mus musculus musculus*) to test whether male courtship vocalizations (i.e., vocalizations emitted in a sexual context) are influenced by a prior direct interaction with a female, and if so, determine how long the effect lasts. We allowed sexually naïve males to directly interact with a female for five minutes (sexual priming), and then we recorded males'vocalizations either 1, 10, 20, or 30 days later when presented with an unfamiliar female (separated by a perforated partition) and female scent. We automatically detected USVs and processed recordings using the Automatic Mouse Ultrasound Detector (A-MUD version 3.2), and we describe our improved version of this tool and tests of its performance. We measured vocalization rate and spectro-temporal features and we manually classified USVs into 15 types to investigate priming effects on vocal repertoire diversity and composition. After sexual priming, males emitted nearly three times as many USVs, they had a larger repertoire diversity, and their vocalizations had different spectro-temporal features (USV length, slope and variability in USV frequency) compared to unprimed controls. Unprimed control males had the most distinctive repertoire composition compared to the primed groups. Most of the effects were found when comparing unprimed to all primed males (treatment models), irrespective of the time since priming. Timepoint models showed that USV length increased 1 day after priming, that repertoire diversity increased 1 and 20 days after priming, and that the variability of USV frequencies was lower 20 and 30 days after priming. Our results show that wild-derived male mice increased the number and diversity of courtship vocalizations if they previously interacted with a female. Thus, the USVs of house mice are not only context-dependent, they depend upon previous social experience and perhaps the contexts of these experiences. The effect of sexual priming on male courtship vocalizations is likely mediated by

Competing interests: The authors have declared that no competing interests exist.

neuro-endocrine-mechanisms, which may function to advertise males' sexual arousal and facilitate social recognition.

Introduction

Males in many species produce complex courtship vocalizations to attract females, which can provide information about their quality and compatibility to potential mates [1–4]. In some taxa, such as insects, amphibians, rodents, and bats, individuals communicate through ultrasonic vocalizations (USVs) (>20 kHz) [5–8]. Male house mice (*Mus musculus)* produce surprisingly complex USVs, which show features of bird song [9] (reviewed in [10]). Males mainly emit USVs upon encountering females or their scent [11–14], and their vocalizations become more complex during courtship and mating [15]. Vocalizations emitted by male mice during sexual contexts are widely referred to as *courtship ultrasonic vocalizations* (cUSVs) [8, 12, 16, 17], though these vocalizations may have other functions. Both sexes vocalize [18], but males emit most of the USVs during direct opposite-sex interactions [19]. Males vocalize at high rates during anogenital sniffing and copulation [13, 16, 20–22], whereas they cease vocalizing abruptly after ejaculation [13, 20, 23, 24]. Females are attracted to playbacks of male USVs [14, 25–27], and these vocalizations can enhance mating and reproductive success [17, 28, 29]. Females are more attracted to vocalizations of males of their own versus other *Mus* species [30] and to the USVs of unrelated males over siblings [14]. The USVs of house mice are innate in the sense that they do not require vocal learning [31]. Nevertheless, the number and types of vocalizations that male mice produce depends upon their internal state and social and sexual contexts [32], which may be influenced by previous experience and perception of potential mating opportunities. USVs may provide indices of an individual's emotional state [33], and may signal a male's sexual arousal and interest in a potential mate.

Numerous studies on laboratory mice have suggested that male cUSVs are influenced by previous socio-sexual experience, and a variety of different approaches have been used to investigate this hypothesis, from a brief exposure to mouse scent to direct long-term interactions and copulation. Several early studies reported that the rate of male cUSV emission is increased after a previous encounter with a female or female scent, and that even a brief experience may have long-lasting effects (persisting >1d) [reviewed in 12, 34–39]. These early findings are intriguing, and they also suggested that male cUSVs provide a reliable index of sexual arousal. However, the results were mixed and varied depending on the strain of mice, the sex and type of stimuli (direct social interactions vs urine odour, and fresh vs aged urine), and only vocalizations at 70 kHz were recorded due to technical limitations. One previous study concluded that socio-sexual experience is only necessary to elicit male USVs when males are presented with aged female urine as a stimulus [38]. Since then, it has been anecdotally suggested that exposing male mice to a sexually mature female for several days or more before recording increases their motivational state to emit cUSVs [40, 41], but this hypothesis was not tested. Another recent study exposed individual mice (strain CBA/CaJ) to another mouse indirectly (separated with a metal mesh divider) for one hour, and then vocalizations were immediately recorded in a novel environment and without any stimulus (non-sociosexual context) [42]. Males (but not females) showed increased USV emission after an experience with females (or even males) compared to isolated stimulus-deprived controls, which had no prior social exposure. This result shows that males increased

their number of *spontaneous* USVs directly following a socio-sexual experience. An additional study recorded USVs of laboratory mice (strain C57BL/6J) during direct male-female interactions [17]. The study found increased rates of USV emission among pairs if the male had been previously co-housed with a female for two weeks, whereas pairs with individually housed males did not show such changes. The increased number in USVs might have been due to socio-sexual experience (either prior copulation or long-term co-housing with a female), female vocalizations (pairs where recorded while directly interacting), or both. Thus, studies are still needed to determine whether a previous direct interaction with a female is sufficient to influence the number and types of male vocalizations in a sexual context, and if so, then how long such effects last.

These previous studies were conducted with laboratory strains (*Mus laboratorius*), and it is unclear whether the findings generalize to wild house mice or to other strains. Laboratory mice have been artificially selected for rapid breeding, and males quickly initiate courtship and mating behaviour upon perceiving another mouse, even another male (see Discussion). Wild-derived males vocalize during interactions with females or their scent [14] and both sexes vocalize at a higher rate during opposite- compared to same-sex interactions [18]. They show enormous inter-individual variation in USV emission [18]; however, unlike laboratory mice, wild-derived male mice vocalize very little, if at all, when they are alone [43], and require direct contact with another mouse or mouse scent [14]. Only one study to our knowledge has investigated socio-sexual experience and USV emission in wild-derived house mice. Males were exposed to a stimulus mouse (separated by a perforated divider) and then recorded at least 5 days later in a sexual context (presentation of female odour) [14]. The number of male cUSVs emitted were not altered by previous exposure to a male or female conspecific; however, the socio-sexual experience regime was limited in this study and it may not have been sufficient for priming, as the authors acknowledged.

We conducted a study on the courtship vocalizations of wild-derived male mice *(Mus musculus musculus)* and our aims were to test whether previous exposure to a female mouse in direct but brief interactions (sexual priming) influences the number of courtship vocalizations (sonic and ultrasonic), the types of USVs, and the spectro-temporal features of USVs compared to control males not previously exposed to a female. To address how long the effects of sexual priming last, we compared a null model, a treatment model (controls vs all primed males, irrespective of the time since priming) and a timepoints model including all 5 groups: before priming and day 1, day 10, day 20 and day 30 after priming. We introduced a female into a male's cage and allowed the mice to interact for 5 min (sexual priming). The mice never copulated during this time. We subsequently recorded the male courtship vocalizations either 1, 10, 20 or 30 d after sexual priming. We predicted that sexual priming would increase male USVs emission compared to unexposed controls, and that this effect might decline over time. We detected USVs and processed the recordings using an improved version of the Automatic Mouse Ultrasound Detector (A-MUD) [44]. We describe the improvements of A-MUD (version 3.2), which enable users to adjust the detection threshold and to automatically assign a quality evaluation score to each putative vocalization detected (i.e., A-MUD elements), and we provide the results of our evaluation of A-MUD's performance. Sexual experience might influence the types of different USVs and their diversity, as well as the rate of vocalizations that males emit. Therefore, we also manually classified USVs into 15 different vocalization types ('syllables'), and investigated changes in vocal repertoire (repertoire diversity and composition of USVs). These questions are relevant to understanding the proximate mechanisms and adaptive functions of male USV emission, and they are also of practical interest, as sexual priming has become a common procedure for eliciting vocalizations from male mice [45].

Materials and methods

Subjects and housing

Our experiment was conducted with wild-derived house mice (*Mus musculus musculus*), which were F1 offspring of wild house mice caught and bred at the Konrad Lorenz Institute of Ethology in Vienna, Austria (48˚12'38"N, 16˚16'54"E) [see more details in 18]. We systematically bred mice that we trapped from different locations and made crosses among them (mean ±s.d. distance between locations 85 ± 71 m). The mice were housed in standard Type IIL cages (36.5 x 20 x 14 cm, with stainless steel cover, 1 cm mesh width, Tecniplast, Germany). The F1 offspring were weaned at 21 d of age and subsequently housed with their siblings in mixed-sex groups for another two weeks (maximum of four mice per cage). At five weeks of age, the sexes were separated. Males were individually housed to prevent fighting and females were housed in sister pairs when possible. All cages were provided with wood shavings (ABEDD, Austria), one nest box (Tecniplast, Germany), nesting material (Nestlet, Ehret, Austria) and one cardboard paper roll as environmental enrichment. Food (rodent diet 1324, Altromin, Germany) and water were provided *ad libitum*. Mice were kept in standard conditions (mean ±s.d. room temperate: 22±2˚C, in a 12:12 h light:dark cycle, lights off at 15:00). Red light was used instead of a complete dark period to be able to conduct experiments when the mice are active. We used 100 adult mice (equal sex ratio; mean±s.d. age: 264±22 d).

Recording apparatus

To record male vocalizations, we used a Plexiglas cage (36.5 x 21 x 15 cm), which had a perforated Plexiglas divider (0.5 cm diameter holes) in the middle to create two equal sides ("caller" and "stimulus" compartments; for details see [18]). The caller compartment (used for the males) had a metal cage lid (1 cm width mesh), and the stimulus compartment (for the females) was covered with a Plexiglas lid, which prevented recording female vocalizations during the experiment. We used USV playbacks from an ultrasound speaker (Avisoft Bioacoustics, Germany) positioned into the stimulus compartment to confirm that the Plexiglass cover was very effective at blocking USVs [18]. The stimulus compartment was also provided with bedding and 2–3 food pellets. For recording, the Plexiglas cage was placed into a recording chamber, lined with acoustic foam, as described in [30]. A condenser ultrasound microphone (Avisoft Bioacoustics/CM16/CMPA, frequency range from 2 to 200 kHz) was mounted inside the recording chamber, 10 cm above the caller compartment, and connected to an UltraSoundGate 116–200 (Avisoft Bioacoustics, Germany). Mice were recorded using the RECORDER USGH-software and with the following settings: 300 kHz sampling rate, 16 bit format and 256 Hz FFT size.

Socio-sexual priming and recording procedures

For our priming treatment, we introduced an unfamiliar adult female (n = 40) into a male's home cage (n = 40) for 5 min. Wild mice never copulate during such a brief period of time—unlike laboratory mice, it usually takes wild-derived mice days to copulate [46], rather than minutes or hours. We subsequently recorded 10 of these males 1 day later in a sexual context, while presenting them with a novel stimulus female (separated by a perforated partition) and female scent (see below). To investigate whether priming effects are long-lasting, the rest of the males were recorded either 10, 20 or 30 d after priming, using 10 males per time point. We also recorded a control group composed of 10 unprimed males. In total we compared five groups: unprimed males (0d), males primed 1 d prior to recordings (1d), males primed 10 d before recordings (10d), males primed 20 d prior to recordings (20d) and males primed 30 d

before recordings (30d). Each group contained 10 males and no male was recorded on more than one day. To prevent the data from being confounded by time and sequence effects during testing, we primed and recorded the males during two months by priming and respectively recording the males in the following order: 1d, 30d, 20d, 10d (priming), and 1d, 0d 10d, 20d and 30d (recording).

We recorded males' vocalizations using the following procedure: first, a female stimulus was placed into the stimulus compartment of the arena for habituation, 5–10 min before introducing the male. To standardize any potential oestrus status effects of this stimulus female, we included an additional olfactory stimulus (5 μl of female urine on a 4 x 4 cm filter paper) to the male compartment. This urine stimulus was previously collected in metabolic cages (Techniplast, 600M021) from four wild-caught adult females, equally aliquoted and mixed in Eppendorf tubes and stored at -20˚C until the recordings. Second, we placed the male subject into the caller compartment and the entire cage was placed into the recording chamber. After 30 s, we began recording and we recorded males for 10 min. After each recording, the arena was cleaned with ethanol before reusing. Each male subject was unfamiliar and unrelated to the females that he encountered in the experiment. Females were used once for priming and once in the stimulus compartment, but never for the same male subject. This experiment was part of a larger study aimed to test whether USV modulation is sex-dependent [18], and 10 males in the current study (i.e., males tested 1 day after sexual priming) were the same as the 'male focal subjects presented with female stimuli' in our previous study.

Detecting vocalizations and processing sound files

To detect ultrasonic vocalizations and process the sound files, we implemented the Automatic Mouse Ultrasound Detector (A-MUD), which detects elements (i.e., putative vocalizations detected by A-MUD) and quantifies spectro-temporal features such as the frequency, amplitude and time parameters of the elements [18, 44]. This tool is implemented as a script in STx (requiring at least S_TOOLS-STx version 4.3), a software from the Acoustic Research Institute (Austria) that is free for scientific use, and is useful for processing large quantities of data in a timely fashion, such as for speech analysis [47, 48], noise evaluation [49, 50], and psychoacoustics [51]. We developed an improved version of A-MUD (version 3.2) and evaluated its performance (see S1 Methods in S1 File). In brief, A-MUD 1.0 has a *detection threshold* at 10 ms because sounds below this threshold are often background noise, and this threshold can now be adjusted by users. For the present study, we lowered the detection threshold to 5 ms to reduce false negative error rates in element detection, despite that this modification increases the risk of false positives. This trade-off was acceptable for the current study because, after the automatic detection, we manually classified all vocalizations in each 10 min file (see below), and thus we were able to correct the output as necessary. A-MUD 3.2 also includes a *quality evaluation score* for each detected A-MUD element, which provides an estimate in the confidence of a true positive, and enables users to remove segments below a certain criterion from the data (the score varies from 0 to 9 with segments $\geq$ 5 being of good quality). A well-established method to judge the performance of a detection or classification model is the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR). The area under the ROC curve (AUC) is one value for the quality of the method (see S1 File for details). We used a ROC curve to evaluate A-MUD's performance using the default settings and the original 14 files used to develop and evaluate A-MUD 1.0 [44], and the AUC value was 0.989 (values $>$ 0.9 are considered to be excellent; see S1 to S4 Figs in S1 File).

After automatic detection, we manually classified each ultrasonic vocalization, assigning it to one of the 15 different previously described types (see Fig 1). We classified and analysed all

USV Type	Abbreviations	Spectrogram example	Definition	References
ultra short	us		Vocalizations below 91 kHz and shorter than 5 ms (short dots)	Modified from Scattoni et al. 2008 [54]
short	s		Vocalizations below 91 kHz and shorter than 10 ms	Modified from Hanson and Hurley 2012 [16]
flat	f		Vocalizations below 91 kHz and with less than 5 kHz frequency modulation	Modified from Hanson and Hurley 2012 [16]
down	d		Vocalizations below 91 kHz and decrease in frequency for more than 5 kHz	Modified from Hanson and Hurley 2012 [16]
up	up		Vocalizations below 91 kHz and increase in frequency for more than 5 kHz	Modified from Hanson and Hurley 2012 [16]
u-shaped	u		Vocalizations below 91 kHz first decrease and then increase in frequency for more than 5 kHz each	Modified from Hanson and Hurley 2012 [16]
u-shaped inverted	ui		Vocalizations below 91 kHz first increase and then decrease in frequency for more than 5 kHz each	Modified from Hanson and Hurley 2012 [16]
complex	c		Vocalizations below 91 kHz and that contain two or more directional changes in frequency and more than 5 kHz modulation of frequency	Modified from Hanson and Hurley 2012 [16]
complex 2	c2		Vocalizations below 91 kHz, which consists of two components separated by one frequency jump but without time separation	Defined as "1-Frequency-Step" in Musolf et al. 2015 [30]
complex 3	c3		Vocalizations below 91 kHz, which consists of three components separated by two frequency jumps but without time separation	Defined as "2-Frequency-Step" in Musolf et al. 2015 [30]
complex 4	c4		Vocalizations below 91 kHz, which consists of four components separated by three frequency jumps but without time separation	Added category for our classification
complex 5	c5		Vocalizations below 91 kHz, which consists of five components separated by four or more frequency jumps but without time separation	Added category for our classification
harmonic	h		Vocalizations below 91 kHz, which have a harmonic component	Modified from Hanson and Hurley 2012 [16]
ultra high	uh	any shape	All vocalizations above 91 kHz	Defined as "high-frequency calls" in Hoffman et al.2012 [55]
unclassified or unclassifiable	uc		Vocalizations that cannot be classified according to any other of the 14 categories due to background noise or not defined and measurable spectrographic features (shape). These vocalizations were *all* confirmed by auditory inspection	Defined as "unstructured vocalizations" in Scattoni et al. 2011 [56] and as "noisy vocalizations" in Grimsley et al. 2016 [52]

Fig 1. Classification of USVs: USV types, their abbreviations, a spectrogram's example and definitions following the classification of [28, 54–56].

https://doi.org/10.1371/journal.pone.0242959.g001

vocalizations in all 10 min files. This work is very time-consuming, but it allowed us to test for priming effects in the ultrasonic vocal repertoire (ultrasonic repertoire diversity and composition, see below), as well as to correct errors of the automatic detection (eliminate false positives, or manually label false negatives and adjust the vocalization length) and then recalculate the USV spectro-temporal parameters in A-MUD. Manual classification also allowed us to additionaly identify and count *sonic vocalizations* (all vocalizations < 20 kHz). We classified sonic vocalizations into two main categories: *low-frequency vocalizations* (LFV), which are similar to USVs but are at frequencies < 20 kHz (adapted from [52]) and *low-frequency harmonic vocalizations* or *squeaks*, which are qualitatively distinct and are vocalizations showing > 1 harmonic component, starting at the sonic range and often reaching the ultrasonic range [32, 52]. The rationale was to test whether the number of ultrasonic vocalizations, which are arbitrarily defined based on human auditory perception, correlate with the number of sonic vocalizations, as previously most studies have focused only on USVs. Mice can discriminate simple versus complex USVs [53], but it is still unknown whether they can discriminate among the various other types of USVs.

Statistical analyses

Statistical analyses were conducted using IBM SPSS Statistics 25, R 3.5.0 and R 3.6.2 and we provide means and s.e.m., unless stated otherwise. We used following working definitions and analyses:

Vocalization rate. The total number of vocalizations (sonic or/and ultrasonic) counted in each 10 min file. All *sonic* vocalizations were manually labelled in each 10 min file. All analysed vocalizations were confirmed by visual and/or acoustic inspection during manual classification.

Spectro-temporal features of USVs. Vocal spectro-temporal features were automatically calculated by A-MUD. All unclassified vocalizations were excluded from these analyses due to their noisy spectrographic features, even though they could be confirmed as USVs by auditory inspections. One mouse in the 10d group did not emit any vocalization and one mouse in the 30d group only emitted one unclassified vocalization. Thus, no vocal spectro-temporal parameters were calculated for these two mice. We analysed *length* (ms), *mean frequency* (kHz), *mean amplitude* (dB) and *slope* (kHz/ms) of each USV during each 10 min trial. The slope was automatically calculated by linear regression using all points in the detected frequency track (kHz/ms). The slope of the resulting regression line is a simplified approximation of the frequency evolution over time and needs to be interpreted with caution (see S1 File).

The statistical analyses of vocalization rates and spectro-temporal features of USVs were computed using R 3.6.2 [57]. Vocalization rate of all vocalizations and USVs showed a negative binomial distribution and were analysed using function glm.nb from package MASS [58]. There were no signs of overdispersion (dispersion parameter 0.95 and 0.72, respectively). The distribution of all model residuals was visually inspected using package fitdistrplus [59]. Residuals from models for the response variable "latency to vocalize" were approximately normally distributed, and therefore, latency was analysed using function lm.

Data on spectro-temporal features of USVs included repeated measurements, and were analysed using generalized linear mixed effects (GLMM) models (package lme4) [60]. We always used animal identity to compute random intercepts. Applied distribution families were inverse gaussian (USV length, slope and frequency), and gaussian (amplitude). For all response variables, we computed three models: First, a null model with an intercept only (describing the data by their mean). Second, a "treatment" model, comparing controls with all primed animals, irrespective of the time since priming. Third, a "timepoints" model with 5 groups, i.e., before priming and day 1, day 10, day 20 and day 30 after priming. We compared these models using AICc, Akaike's Information Criterion corrected for small sample sizes, computed using package MuMIn [61]. To directly assess the support for each candidate model, we also computed relative model likelihoods from the differences in AICc to the best model (ΔAICc) as relative likelihood = exp(-0.5 * ΔAICc) [62]. We subsequently provide p-values for the effects from the best model, following Zuur et al (2009) [63]. We always used so-called "treatment contrasts" by placing the control group on the intercept, and then comparing all other groups to these unprimed males. We did not compute post-hoc comparisons. The letters in the figures refer to the GLM/GLMM regression coefficients, i.e., the significance comparing differences between day 0 versus any later day (in the case of "timepoints" models), and to differences between "unprimed" and "primed" (in the case of "treatment" models). In one case (variable mean frequency), the variability in the data noticeably decreased after priming. To analyse these changes in variability among USV frequencies, we computed the absolute deviation of frequency measurements from their median as the response variable, analogous the Levene's test, and computed a GLMM, again with random intercepts per animal.

Vocal repertoire of USVs. Vocal repertoire was assessed using two measures. First, repertoire *diversity*, which is total number of vocalization types per sound-file (Fig 1). This number is a rough estimate of diversity and ranges from 0 (no vocalization) to 15 (maximal amount of diversity). Variables "repertoire diversity" and "repertoire diversity without unclassified vocalizations" were Poisson distributed and analysed with glm. There were no signs of overdisperison (dispersion parameters 0.872 and 0.878, respectively). Second, to investigate the number

of vocalizations per each type, i.e., vocalization type occurrence, we calculated repertoire *composition*. We analysed repertoire composition using a multivariate approach by running a non-parametric analysis of similarities (ANOSIM) in R (V 3.5.0) [64] (package "vegan", functions "vegdist" and "anosim") [65]. The ANOSIM statistic compares the mean of ranked dissimilarities between groups to the mean of ranked dissimilarities within groups. The generated R value lies between -1 and +1, with a value of 0 representing the Null hypothesis (indistinguishable groups), an R close to 1 indicates that dissimilarity between groups are greater than within groups, while an R values < 0 indicate that dissimilarities within groups are greater than between groups. The test was run with 999 permutations and using the Bray-Curtis dissimilarity matrix. We also ran a permutational multivariate analysis of variance (PERMANOVA) with 999 permutations in R (package "vegan", function "adonis2") [65]. The calculated pseudo F-ratio compares the total sum of squared dissimilarities among vocalizations of different groups to vocalizations within the same group. Larger F-ratios indicate pronounced group separation. The one mouse in the 10d group, which did not emit any vocalizations was excluded from these analyses. To visualize the results, we used a non-metric multidimensional scaling (nMDS) approach based on the Bray-Curtis dissimilarity index (package "vegan", function "metaMDS"). A stress coefficient of <0.05 indicates an excellent visualization of data, whereas a stress coefficient of >0.3 indicates an almost arbitrary position of data on the graph [66]. Our stress values were calculated as the mean of the 21 iterations we ran. The similarity between groups is measured by the distance between the points: the closer the distance, the greater the similarity in the composition of the vocal repertoire between groups. We also investigated vocalization type contribution to group dissimilarities using the "simper" function in R, which performs pairwise comparisons of groups to find the average contribution of each vocalization type to the average overall Bray-Curtis dissimilarity, displaying the most important vocalization types for each pair of groups. These vocalization types contributed at least to 70% of the differences between the groups. We also visualized the vocal repertoire in pie charts by calculating the proportions of each vocalization type for each mouse and then averaging these proportions for each experimental group.

Ethical statement

After the recordings, all the mice were returned to their home cages and kept in our colony. This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and complies with the current laws of Austria. All the experiments were conducted at the Konrad Lorenz Institute of Ethology, Austria and the protocols have been approved and were in accordance with ethical standards and guidelines in the care and use of experimental animals of the Ethical and Animal Welfare Commission of the University of Veterinary Medicine, Vienna (Austria) (ETK-17/04/2015) in accordance with Good Scientific Practice guidelines and national legislation. We did not sacrifice any of the mice used for this study.

Results

Vocalization rate (ultrasonic and sonic)

USVs. The number of USVs emitted per individual was highly variable, ranging from 0 to 627 USVs with an overall mean±s.d. of 117±164 vocalizations per male during the 10 min recordings (median = 36 vocalizations/10 min). The USV count was right skewed: approximately half of the males emitted $\leq$ 50 USVs (n = 26) and the other half emitted between 51–627 USVs (n = 24) (Fig 2). The mean number of USVs increased from 50±25 among unprimed control males to 142±29 among primed males (negative binomial GLM; p = 0.013, Fig 3,

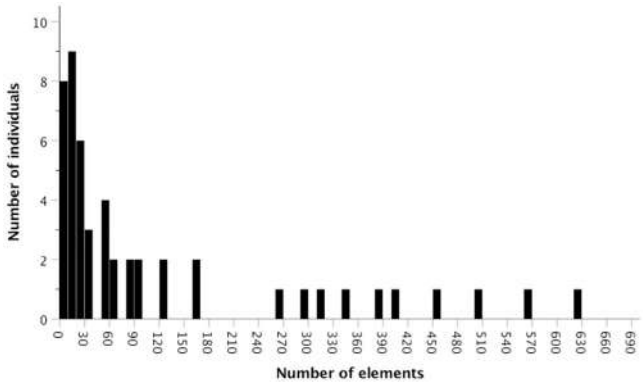

Fig 2. Histogram depicting variation in ultrasonic vocalization rate (number of USVs per 10 min) among individuals. Approximately half of the males emitted less than 50 USVs during the 10 min trials, though some mice were very vocal (n = 50).

https://doi.org/10.1371/journal.pone.0242959.g002

Table 1). The model comparing treatments had a lower AICc (554.25) than both the null model (AICc = 556.76) and the timepoints model (AICc = 560.49).

On average, it took the mice 120±136 s before emitting the first USV ("latency to vocalize"), with a minimum = 0.03 s to a maximum = 600 s (no vocalizations, n = 1) and group means±s.d. of 69.6±94 s for 0d, 64.3±117 s for 1d, 191.9±151 for 10d, 128.6±111 s for 20d and 147.3 ±173 s for 30d. Latency to vocalize was unaffected by priming, and the null model had the lowest AICc (Table 2). The "treatment" model was nearly undistinguishable from the null model (relative likelihood 0.63).

Sonic vocalizations. We counted number of low-frequency vocalizations (LFVs) and low-frequency harmonic vocalizations (squeaks) to investigate whether USVs correlate with sonic vocalizations and are a good representation of a male overall vocalizations (i.e., sonic and ultrasonic) as explained above. We found that LFVs showed high individual variability (overall mean±s.d.: 9.9±8 and range: 0–46; means per group: 0d = 7.3±5, 1d = 16±13, 10d = 7.6 ±6, 20d = 9.5±6 and 30d = 9.2±8). The number of LFVs and ultrasonic vocalizations positively correlated with each other (Spearman correlation: $\rho = 0.49$, n = 50, $p<0.0001$). Squeaks were also highly variable (overall mean±s.d.: 11.9±21.2 and range: 0–132; means per group: 0d = 17.6±40, 1d = 17.7±18, 10d = 2.9±3, 20d = 12.4±10 and 30d = 8.9±15) and positively correlated with USVs as well (Spearman correlation: $\rho = 0.469$, n = 50, p = 0.001). Taken together, USVs positively correlated with sonic vocalizations (Spearman correlation: $\rho = 0.502$, n = 50, p = 0.0001) and thus we also investigated priming effects on overall vocalization rates, as there were too few sonic vocalizations to be analysed separately and draw meaningful conclusions.

Sonic and ultrasonic vocalizations. We investigated priming effects on the overall vocalization rates, i.e., merging ultrasonic and sonic vocalizations. The mean number of all vocalizations increased from 75.1±28.4 to 163.7±29.8 (controls vs primed males; negative binomial GLM; p = 0.036, Fig 4). This treatment model (AICc = 576.57) was slightly better than the null model (intercept only; AICc = 577.95, relative likelihood 0.5), but much better than a model differentiating between timepoints (AICc = 582.14, Table 3).

Spectro-temporal features of USVs

We calculated the spectro-temporal features of all detected USVs omitting uc vocalizations due to their noisy and unstructured features, n = 5151 USVs. Frequency, slope and amplitude parameters could not be calculated for 198 USVs due to being too short or faint.

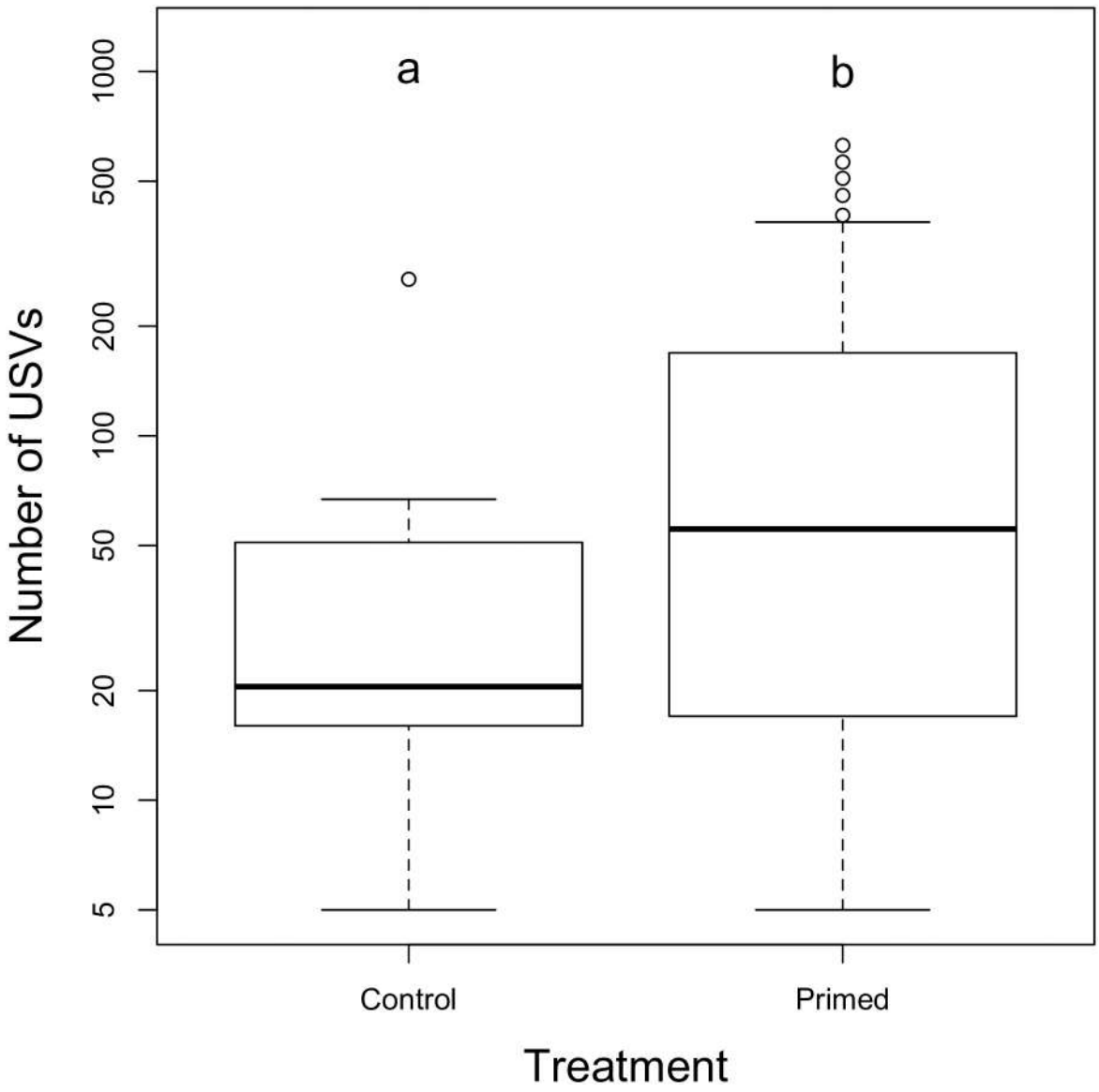

Fig 3. Number of USVs emitted with or without sexual priming. Boxplots of the number of USVs emitted by unprimed (control) and primed males. Boxes around the median (horizontal line) show the interquartile range (quartile 1 to 3) and whiskers extend to 1.5 times this range, or to the most extreme point, whichever is closer to the median. Extreme points are shown as circles. Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g003

The mean length of USVs more than doubled at day 1 after priming (14.93±0.71 ms vs. 32.5 ±0.60 ms; p = 0.006, Figs 5 and 6). GLMM analysis indicated that USV length was significantly longer after priming at day 1 (p = 0.006), but not on later trial dates. The model differentiating between timepoints (AICc = 43344.88) was better than both the null model (AICc = 43347.68) and the treatment model (AICc = 43348.74, Table 4). USVs' slope increased from -0.16±0.11 in unprimed to +0.16±0.01 in primed animals (GLMM; p = 0.022, Fig 7). This model comparing treatments had the lowest AICc (-27972.49). The model differentiating between all timepoints (AICc = -27969.47) and the null model (AICc = -27969.86) were inferior (Table 4). The mean frequency (kHz) of male vocalizations was unaffected by priming (GLMM; p>0.3); however, there was lower variability in the vocalizations of primed compared to unprimed males

Table 1. Tables 1 to 5 are regression tables of effects of priming on various aspects of vocalization in male mice.

Number of USVs (GLM negative binomial)					
Coefficients:					
	Estimate	**Std. Error**	**z value**	**P value**	**Relative Likelihood**
model timepoints AICc = 560.49					0.04
(Intercept)	3.92	0.37	10.6	<2e-16 ***	
groupa1	1.22	0.52	2.3	0.02 *	
groupa10	0.63	0.54	1.2	0.24	
groupa20	1.10	0.52	2.1	0.03 *	
groupa30	1.07	0.54	2.0	0.05 *	
model treatment AICc = 554.25					1.0
(Intercept)	3.92	0.37	10.5	<2e-16 ***	
treatPrimed	1.04	0.42	2.5	0.01 *	
Null model AICc = 556.76					0.28

In all models, intercepts represent the mean of the control group. Coefficient estimates are the differences of group means to the intercept (at 1, 10, 20, and 30 days after priming for timepoint models, and for pooled data after priming for treatment models). Tables also show standard errors, t-values, z-values and P-values for the deviation of these differences from zero. Null models are intercept only models. AICc values give Akaike's Information Criterion corrected for small sample size. The best AICc values are printed in bold face and relative likelihoods are the plausibilities of candidate models compared with the best model. The model type is given in parentheses. For mixed models (adjusting for repeated measurements) the random effects, e.g., standard deviations of intercepts of individuals, are also provided. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

https://doi.org/10.1371/journal.pone.0242959.t001

(Table 4). This decrease in the variability of frequency of primed males was a tendency at day 1 (p = 0.051), and significant on days 20 (p = 0.029) and 30 (p = 0.026, Fig 8). Neither the magnitude nor the variability of USV amplitude were affected by priming (Table 4).

Vocal repertoire of USVs (repertoire diversity and composition)

We found that number of USVs were positively correlated with repertoire diversity (Spearman correlation: $\rho = 0.91$, n = 50, $p<0.0001$). GLM analysis showed that the timepoint model for repertoire diversity had the lowest AICc (AICc = 289.68), but the treatment model was nearly

Table 2. Regression table of effects of priming on latency to call in male mice.

Latency to call (GLM gaussian)					
Coefficients:					
	Estimate	**Std. Error**	**t value**	**P value**	**Relative Likelihood**
model timepoints AICc = 580.29					0.15
(Intercept)	69.6	29.5	2.4	0.02 *	
groupa1	-5.3	41.7	-0.1	0.90	
groupa10	76.9	42.8	1.8	0.08	
groupa20	59.0	41.7	1.4	0.16	
groupa30	27.4	42.8	0.6	0.53	
model treatment AICc = 577.42					0.63
(Intercept)	70	30	2.3	0.02 *	
treatPrimed	39	34	1.2	0.25	
Null model AICc = 576.51					1.0

https://doi.org/10.1371/journal.pone.0242959.t002

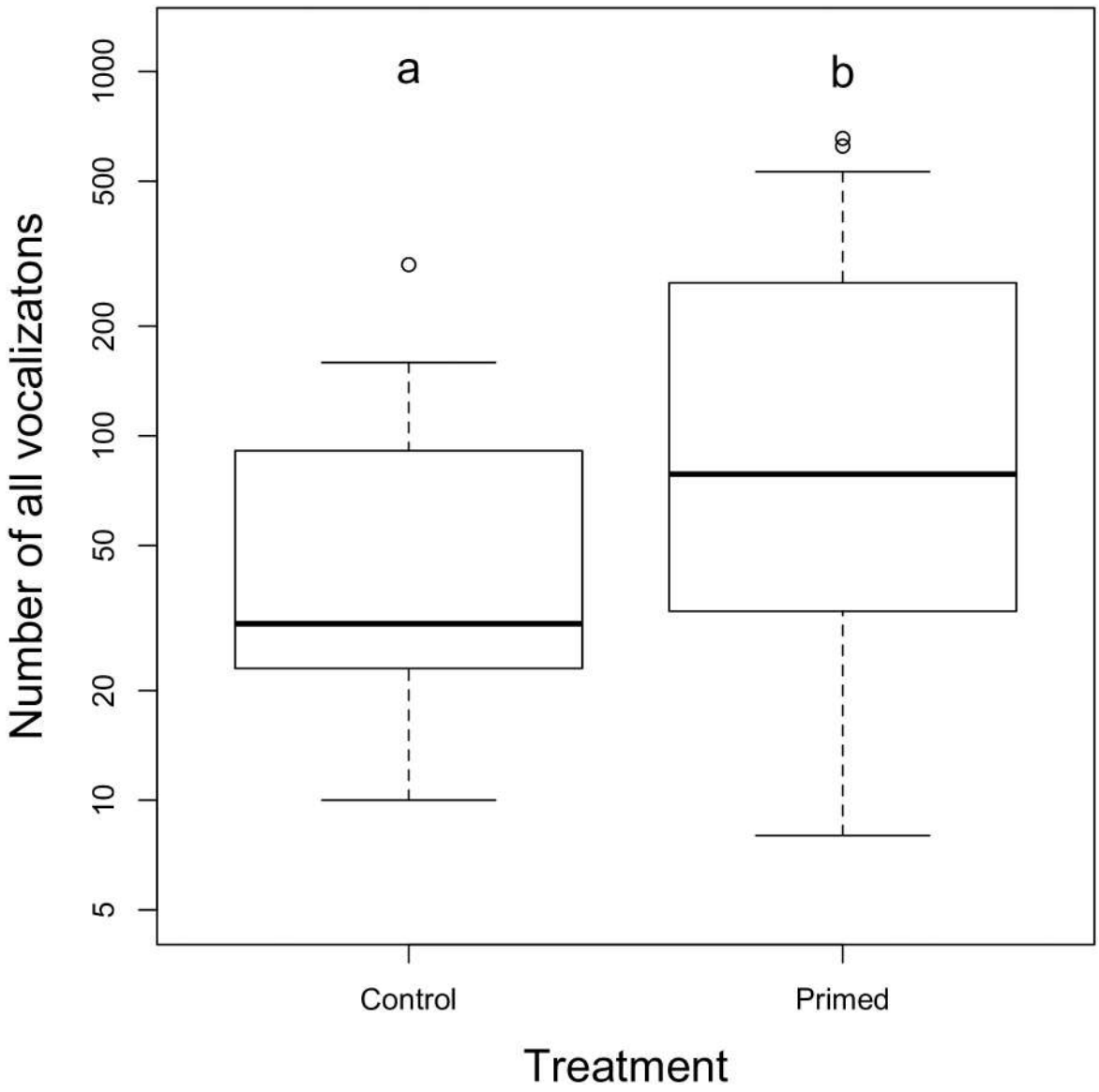

Fig 4. Number of all vocalizations (sonic and ultrasonic) emitted with or without sexual priming. Boxplots with medians of the number of the overall vocalizations emitted by unprimed (control) and primed males. Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g004

undistinguishable (AICc = 290.26, relative likelihood 0.74); the null model was clearly worse (AICc = 294.01, Table 5).

Compared with controls, repertoire diversity was significantly increased at days 1 ($p<0.002$) and 20 ($p<0.044$) after priming (Fig 9).

We examined priming effects in repertoire composition to assess vocalization type occurrence per group, using two statistical non-parametric multivariate approaches. Both analyses showed that groups differed and that the unprimed males had the most distinctive vocal repertoire (ANOSIM: R = 0.108, p = 0.01) and (PERMANOVA: F(4,44) = 1.98, p = 0.015). We visualized the data using non-metric multi-dimensional clustering (nMDS), and plotted the

Table 3. Regression table of effects of priming on number of all vocalizations in male mice.

Number of all vocalizations (GLM negative binomial)					
Coefficients:					
	Estimate	**Std. Error**	**z value**	**P value**	**Relative Likelihood**
model timepoints AICc = 582.14					0.06
(Intercept)	4.32	0.33	13.3	<2e-16 ***	
groupa1	1.00	0.46	2.2	0.03 *	
groupa10	0.34	0.47	0.7	0.47	
groupa20	0.84	0.46	1.8	0.07	
groupa30	0.80	0.47	1.7	0.09	
model treatment AICc = 576.57					1.0
(Intercept)	4.32	0.33	13.0	<2e-16 ***	
treatPrimed	0.78	0.37	2.1	0.04 *	
Null model AICc = 577.95					0.50

https://doi.org/10.1371/journal.pone.0242959.t003

occurrence of each vocalization type per experimental group (Fig 10). This graph shows the clustering of the vocalization types emitted by individuals (coloured symbols) to visualize the differences within versus between groups (individuals are connected to a centroid that minimizes the distances between individuals within each group). The different vocalization types (represented by letters) are positioned according to their highest clustering. The main difference appears to be due to unprimed males emitting more unclassified (uc) and fewer complex vocalizations than the other mice, whereas the 1d primed males have more ultra-high (uh), complex and less short (s) and ultra-short (us) vocalizations than the other mice (Fig 10). Further visualization of the proportions of vocalization types emitted by mice in the different groups are shown in pie charts (Fig 11), which indicate that the main differences were between the primed and unprimed males.

We also conducted the same analyses omitting all unclassified USVs, because, although confirmed by auditory inspection as vocalizations, they appeared unstructured and noisy compared to the other USVs. We found that the results on repertoire diversity are largely unchanged when omitting unclassified (uc) USVs. The nMDS graph displaying repertoire composition shows a separation of the 0d and 1d group, which is now mainly driven by short USVs (instead of the omitted uc) (see S1 Results in S1 File and S5-S7 Figs in S1 File).

Discussion

Our main aim was to experimentally test whether male house mice show increased rates of vocalizing following a direct interaction with an adult female (sexual priming), and our most important results include the following: First, we found that sexually primed males emitted significantly more USVs than unprimed controls, consistent with previous studies measuring 70 kHz vocalizations of laboratory mice [e.g. 39]. We also found that the rates of ultrasonic and sonic vocalizations were positively correlated with each other. Thus, USVs in our study provided a good estimate of the rates of sonic calls and the overall vocalization rates, and these relationships have not been previously compared to our knowledge. The effect of priming on overall vocalization rates were likely dominated by USVs, however, because sonic vocalizations were less common.

Second, we developed an improved version of automated USV detection (A-MUD, version 3.2) and we used the data from this study to evaluate its performance. We found that lowering the detection threshold reduced false negative error rates, but it also increased the risk of false

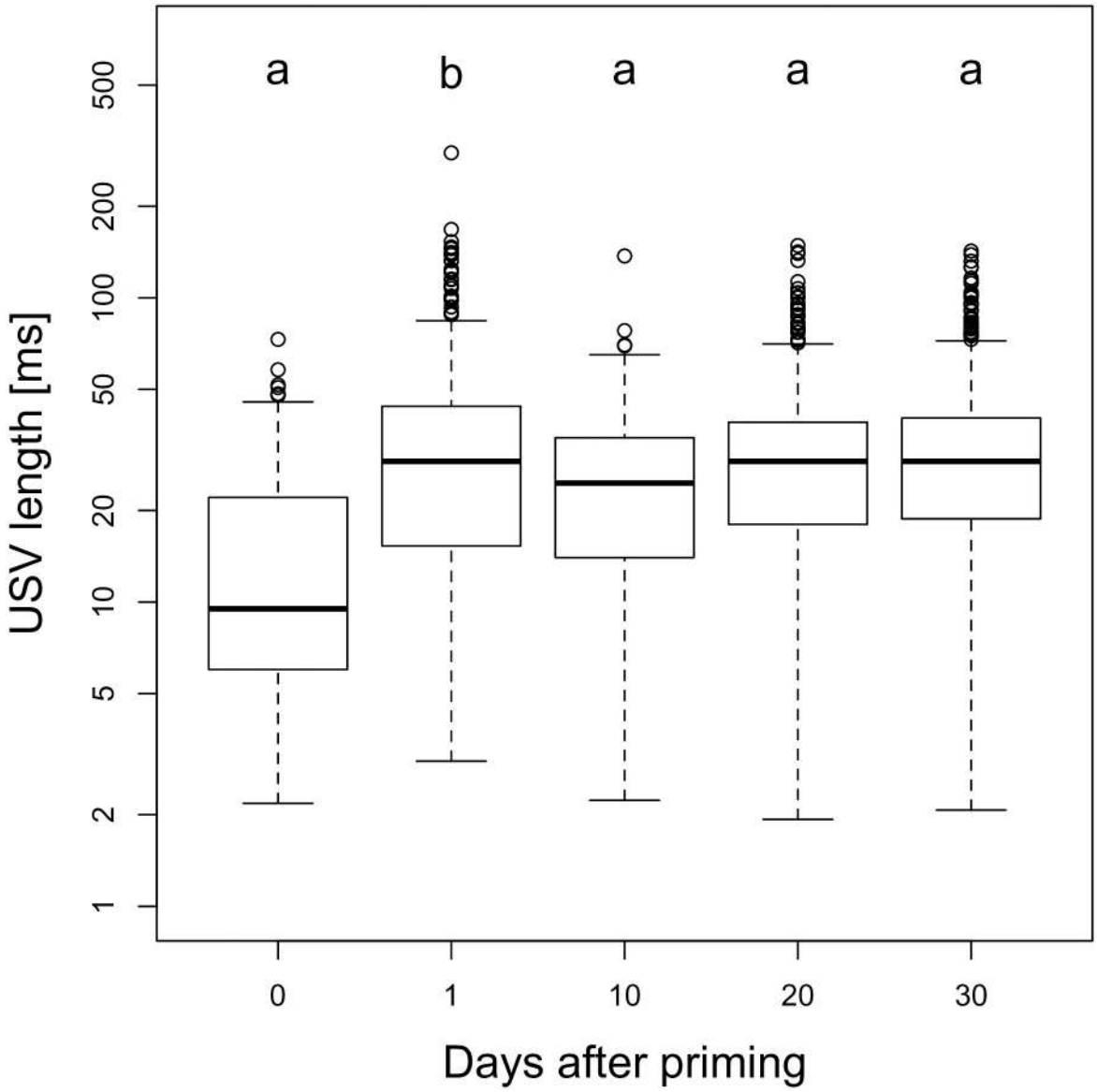

Fig 5. USV length with or without priming. Boxplots with medians of USV lengths emitted by unprimed (0) and primed males (≥1). Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g005

positives. This detection trade-off was acceptable because we also manually classified the vocalizations in this study, and thus we were able to correct the output as necessary. This new version of A-MUD (3.2) includes a *quality evaluation score* for each detected element (an estimate in the confidence of a true positive) and it also enables users to remove segments below a certain criterion. We evaluated A-MUD's performance using receiver operating characteristic (ROC) curve and found that the AUC value was excellent. This tool is free for scientific (non-profit) use and available here: https://www.kfs.oeaw.ac.at/doc/amud/AMUD1b.sts (Script); Readme: https://www.kfs.oeaw.ac.at/doc/amud/AMUD1b-Readme.odt.

Third, we quantified different types of vocalizations ('syllables'), and found that the repertoire diversity of male vocalizations significantly increased one day after priming, and

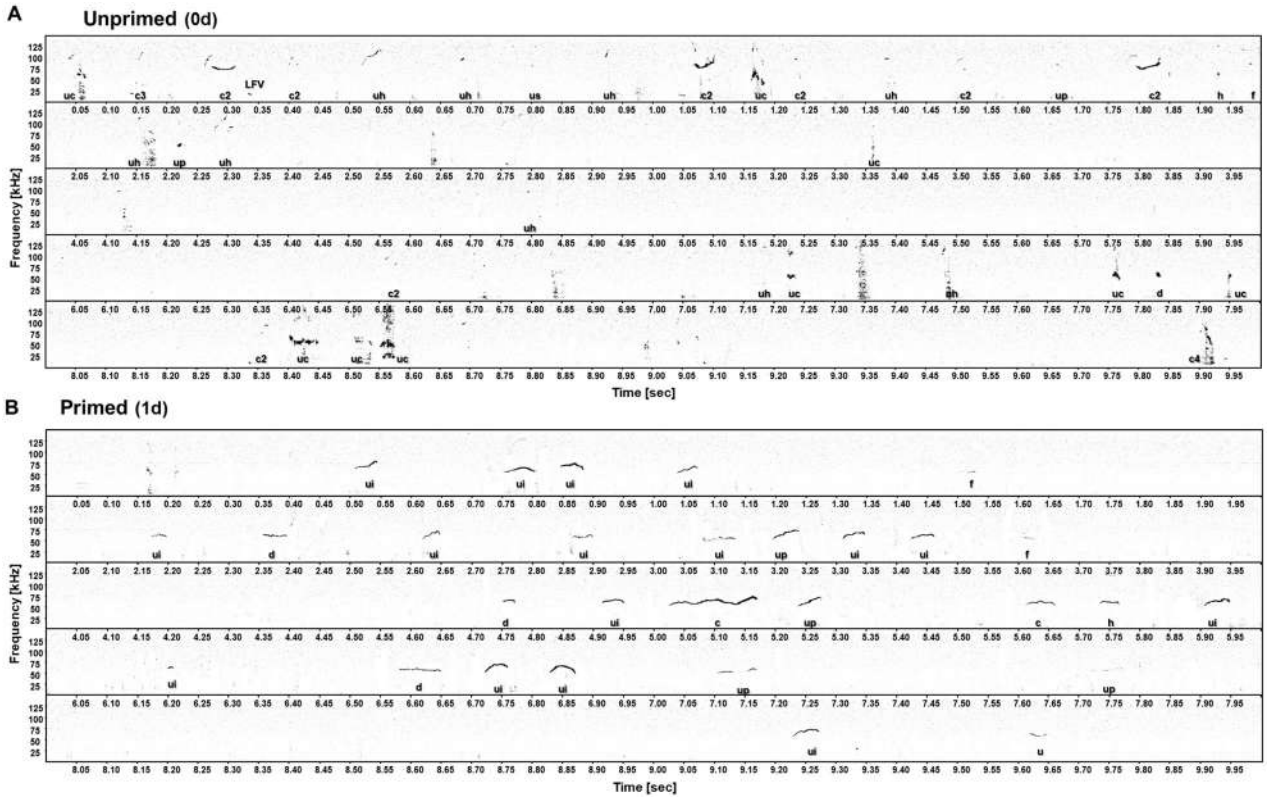

Fig 6. Spectrogram examples of an unprimed (0d) and a sexually primed (1d) male. The two spectrograms show a 10 s continuous sequence of the males that emitted most USVs in both groups, (A) the unprimed group and (B) in the group recorded 1d after priming. All lines of the spectrograms are continuous and each line shows 2 s (50 ms interval) of the 10 s sequence. Y-axes represent frequencies between 0–150 kHz with intervals of 25 kHz. Letters indicate examples of vocalization types, following the definitions and abbreviations in Fig 1. LFV = low-frequency vocalization.

https://doi.org/10.1371/journal.pone.0242959.g006

multivariate analyses indicated that the unprimed males had the most distinctive repertoire composition. Unprimed males emitted mostly unclassified calls and fewer complex types of vocalizations, whereas males tested one day after priming emitted more ultra-high and complex vocalizations and fewer short and ultra-short vocalizations. Vocal spectro-temporal features of USVs also differed after priming, and priming affected USV length, slope and frequency variability.

Fourth, primed males did not differ in the mean frequency (kHz) of their vocalizations from unprimed controls, but interestingly their calling frequencies showed significantly lower variability (Fig 8). As primed males showed more 'agreement' in their calling frequency than unprimed males, they might be 'targeting' female auditory perception or preferences. Future studies are needed to compare how male vocalization frequency matches female auditory sensitivity thresholds, and whether the frequency of male cUSVs influence female preferences for recorded playbacks.

Finally, the main differences in USVs were between the primed males versus the unprimed controls (treatment models), regardless of the time since priming. Since we observed changes in males' USVs after day 1, our results provide novel evidence for long-lasting effects from sexual priming. Timepoint models indicated that USV length increased 1 day after priming, that males' USV repertoire diversity increased 1 and 20 days after priming, and that the variability in the frequencies of vocalizations was lower 20 and 30 days after priming.

Table 4. Regression table of effects of priming on the spectro-temporal features of USVs in male mice.

Spectro-temporal features of USVs: Length (GLMM)					Relative Likelihood
model timepoints AICc = 43344.88					1.0
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	3.4e-06	0.0019		
Residual		1.6e-02	0.1269		
Number of obs: 5151, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	0.0121	0.0019	6.5	9e-11 ***	
groupa1	-0.0080	0.0029	-2.7	0.006 **	
groupa10	0.0016	0.0025	0.6	0.541	
groupa20	-0.0025	0.0026	-1.0	0.336	
groupa30	-0.0015	0.0030	-0.5	0.607	
model treatment AICc = 43348.74					0.15
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	4.8e-06	0.0022		
Residual		1.6e-02	0.1272		
Number of obs: 5151, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	0.0129	0.0021	6.1	1e-09 ***	
treatPrimed	-0.0023	0.0023	-1.0	0.3	
Null model AICc = 43347.68					0.25
Spectro-temporal features of USVs: slope (GLMM)					Relative Likelihood
model timepoints AICc = -27969.47					0.21
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.0020	0.045		
Residual		0.0039	0.062		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	16.20	0.14	119.0	<2e-16 ***	
groupa1	-0.29	0.16	-1.9	0.064	
groupa10	-0.30	0.17	-1.7	0.081	
groupa20	-0.42	0.16	-2.6	0.008 **	
groupa30	-0.47	0.16	-2.9	0.004 **	
model treatment AICc = -27972.49					1.0
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.0030	0.055		
Residual		0.0039	0.062		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	16.20	0.15	110.1	<2e-16 ***	

(*Continued*)

Table 4. (Continued)

treatPrimed	-0.36	0.16	-2.3	0.02 *	
Null model AICc = -27969.86					0.23
Spectro-temporal features of USVs: frequency (GLMM)					Relative Likelihood
model timepoints AICc = -8862.94					<0.01
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.308	0.55		
Residual		0.044	0.21		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	4.06	0.40	10.2	<2e-16 ***	
groupa1	-0.85	0.68	-1.3	0.2	
groupa10	-0.49	0.59	-0.8	0.4	
groupa20	-0.44	0.60	-0.7	0.5	
groupa30	-0.38	0.64	-0.6	0.6	
model treatment AICc = -8868.47					<0.01
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.308	0.56		
Residual		0.044	0.21		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	4.06	0.40	10.2	<2e-16 ***	
treatPrimed	-0.52	0.46	-1.1	0.3	
Null model AICc = -8889.24					1.0
Spectro-temporal features of USVs: variability of frequencies (GLMM)					Relative Likelihood
model timepoints AICc = 30480.26					0.23
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.00083	0.029		
Residual		0.60243	0.776		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value	P value	
(Intercept)	0.059	0.012	4.8	2e-06 ***	
groupa1	0.033	0.017	1.9	0.05	
groupa10	0.011	0.017	0.6	0.52	
groupa20	0.037	0.017	2.2	0.03 *	
groupa30	0.039	0.017	2.2	0.03 *	
model treatment AICc = 30477.29					1.0
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	0.00089	0.03		
Residual		0.60263	0.78		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					

(*Continued*)

Table 4. (Continued)

	Estimate	Std. Error	t value	P value	
(Intercept)	0.060	0.013	4.7	3e-06 ***	
treatPrimed	0.030	0.014	2.1	0.03 *	
Null model AICc = 30479.30					0.36
Spectro-temporal features of USVs: amplitude (GLMM)					Relative Likelihood
model timepoints AICc = 26662.96					1.0
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	2.3	1.5		
Residual		12.6	3.5		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value		
(Intercept)	15.316	0.664	23.0		
groupa1	0.878	0.838	1.0		
groupa10	-0.242	0.898	-0.3		
groupa20	0.048	0.876	0.1		
groupa30	-0.478	0.873	-0.5		
model treatment AICc = 26664.09					0.56
Random effects:					
Groups	Name	Variance	Std.Dev.		
Replicate	(Intercept)	2.3	1.5		
Residual		12.6	3.5		
Number of obs: 4948, groups: Replicate, 48					
Fixed effects:					
	Estimate	Std. Error	t value		
(Intercept)	15.32	0.67	22.9		
treatPrimed	0.10	0.73	0.1		
Null model AICc = 26663.3					0.84

https://doi.org/10.1371/journal.pone.0242959.t004

Our results could potentially be due to individual housing *reducing* the rates of USV emission of controls, and sexual priming restoring normal USV responses (though this alternative interpretation is not mutually exclusive to the hypothesis that sexual experience *increases* USV emission). Individual housing has been reported to influence the behaviour and physiology of laboratory mice in some, but not all studies [e.g. 67–69]. A recent study on laboratory mice reported that individually housed males *increased* the emission of USVs during male-male interactions compared to socially housed males [70]. Male mice were kept in either individual or social housing (4 mice per group) for five weeks, and USVs were recorded during direct interactions with other males (kept previously in individual or social housing). Interestingly, male USV emission was correlated with the male mounting behaviour of individually housed males, and the authors concluded that their findings were due to 'inappropriate' courtship and mating behaviour by individually housed male males towards same-sex conspecifics. If male mounting behaviour in this previous study was a consequence of sexual arousal, then our results are consistent. We have never observed male-male mounting in wild-caught or wild-derived mice, even under similar circumstances or in semi-natural conditions (wild male mice are more aggressive than most laboratory strains). Nevertheless, the effects of previous inter-

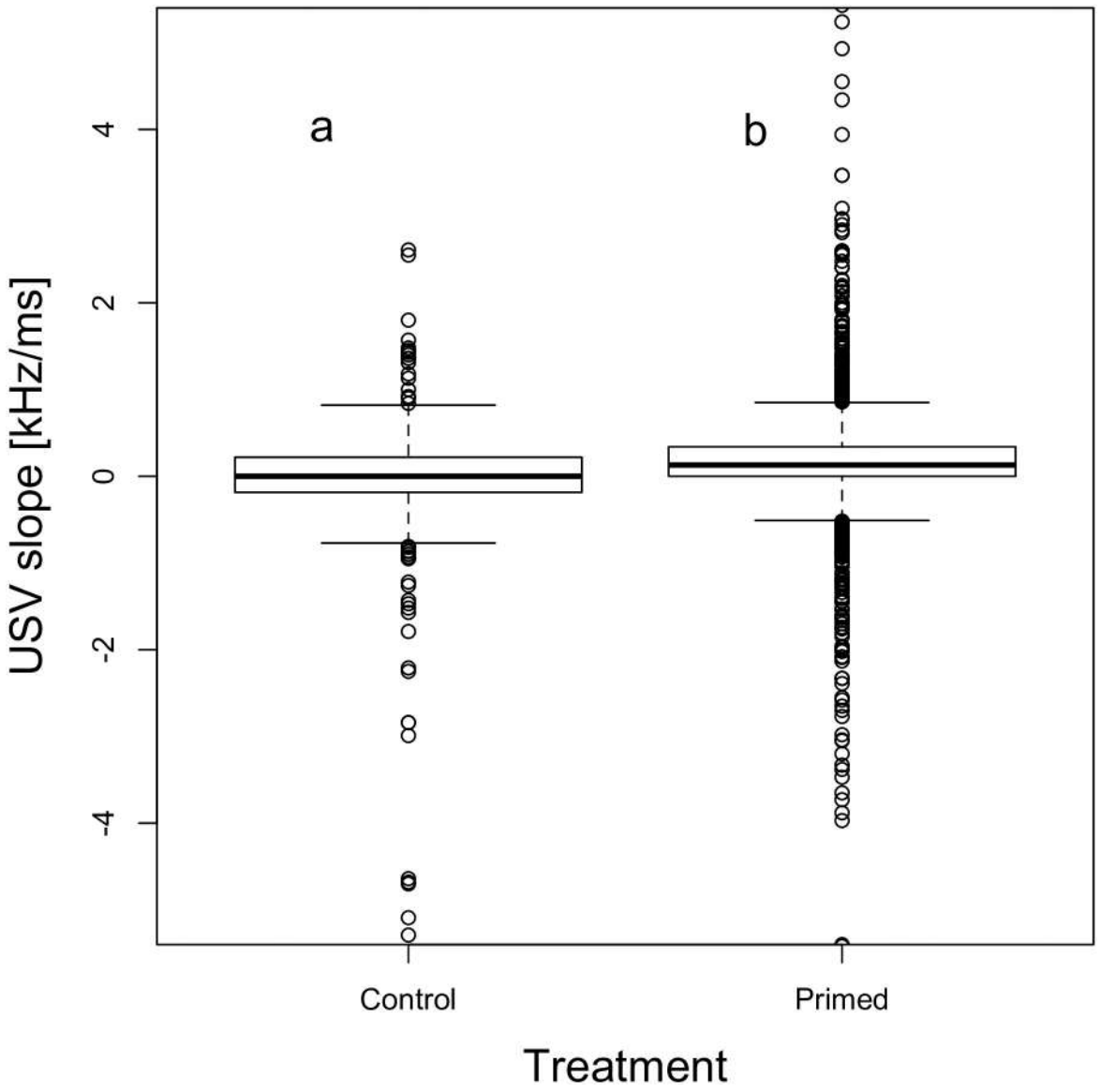

Fig 7. USV slope with or without priming. Boxplots with medians of USV slopes emitted by unprimed (control) and primed males. Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g007

and intra-sexual social experience on USV emission needs to be examined under more natural social conditions [32]. Also, we only examined the effects of previous experience of female interactions on male USVs, and future studies are needed to investigate same- versus opposite-sex priming on both sexes to determine whether such effects are sexual, social, or both (socio-sexual).

It is unclear how sexual experience induces changes in male vocalizations, but some potential neuro-endocrine mechanisms have been identified [see 39, 45, 71–73]. Sexual stimuli trigger a surge of androgens, which regulate male USV emission and other sexual behaviours [74–79]. Sexual experience induces long-term changes and selective elevations of androgen receptors in the medial preoptic area (mPOA) [80, 81], a key site for the integration of sensory

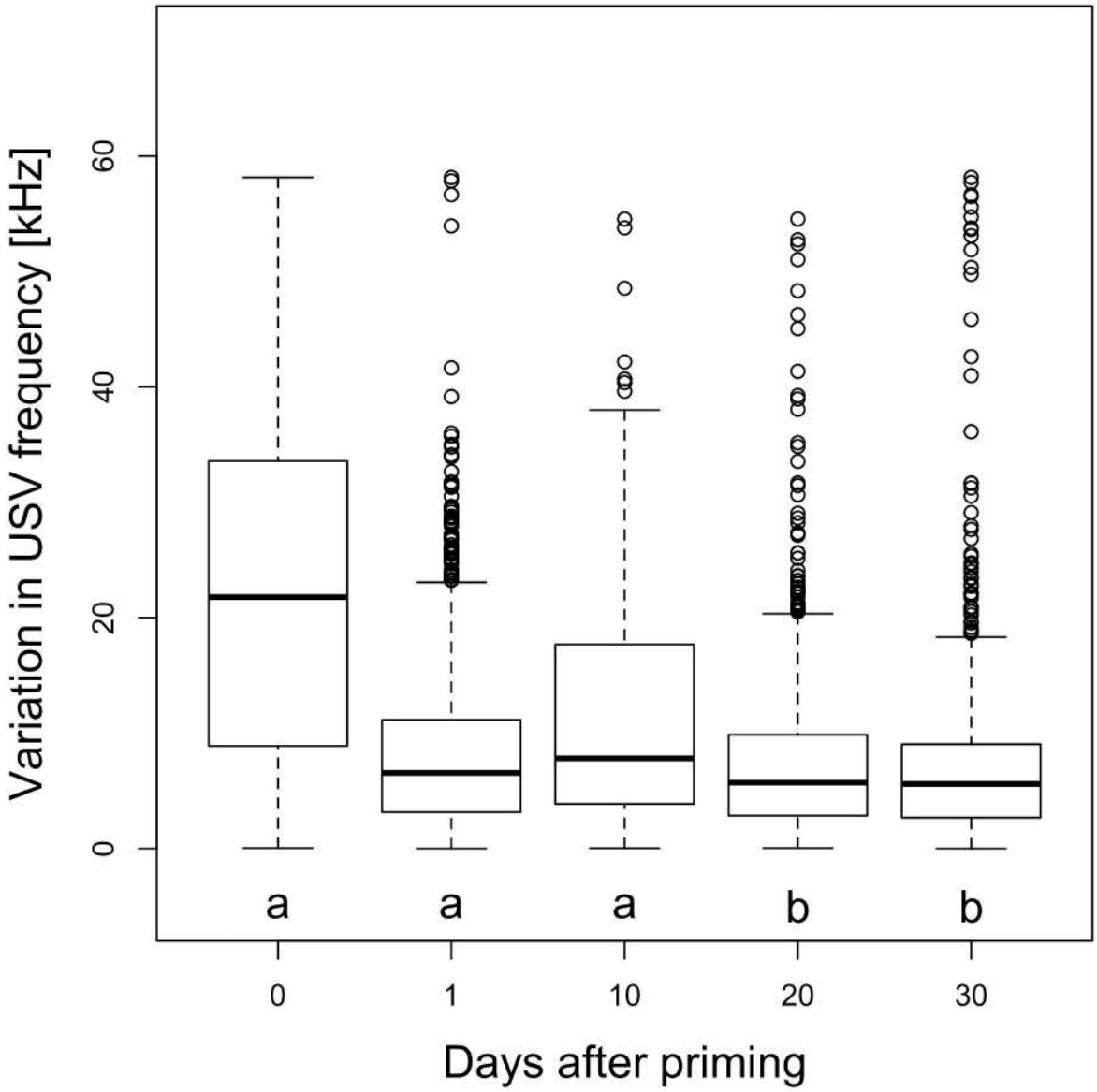

Fig 8. Variability in USV frequency between primed and unprimed males. Boxplots of absolute deviations of the USV frequency from the overall median. Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g008

inputs and control of motor behaviour, including courtship USVs [82]. Sexual priming may also influence the specialized neurons in the midbrain periaqueductal gray (PAG) that control USV emission [83]. Future studies are needed to better understand the neuro-endocrine mechanisms that control USV emission and how they are affected by sexual priming.

The functions of such experiential effects on male USV emission are also unclear. Effects from sexual priming are thought to motivate and prepare males for courtship and mating, as they also trigger increased scent-marking [84], sperm density [85], and copulatory behaviour [86]. Thus, our findings are consistent with the hypothesis that courtship USVs provide a reliable indicator of a male's sexual arousal [20, 34, 87] [reviewed in 45, 73]. They are also consistent with a study showing that sexual priming "emboldens" male mice and increases their

Table 5. Regression table of effects of priming on the vocal repertoire in male mice.

Vocal repertoire (GLM Poisson)					Relative Likelihood
Coefficients:					
	Estimate	**Std. Error**	**z value**	**P value**	
model timepoints AICc = 289.68					1.0
(Intercept)	1.758	0.131	13.4	<2e-16 ***	
groupa1	0.525	0.166	3.2	0.002 **	
groupa10	0.083	0.182	0.5	0.650	
groupa20	0.346	0.172	2.0	0.044 *	
groupa30	0.334	0.172	1.9	0.052	
model treatment AICc = 290.26					0.74
(Intercept)	1.76	0.13	13.4	<2e-16 ***	
treatPrimed	0.33	0.14	2.3	0.02 *	
Null model AICc = 294.01					0.11

https://doi.org/10.1371/journal.pone.0242959.t005

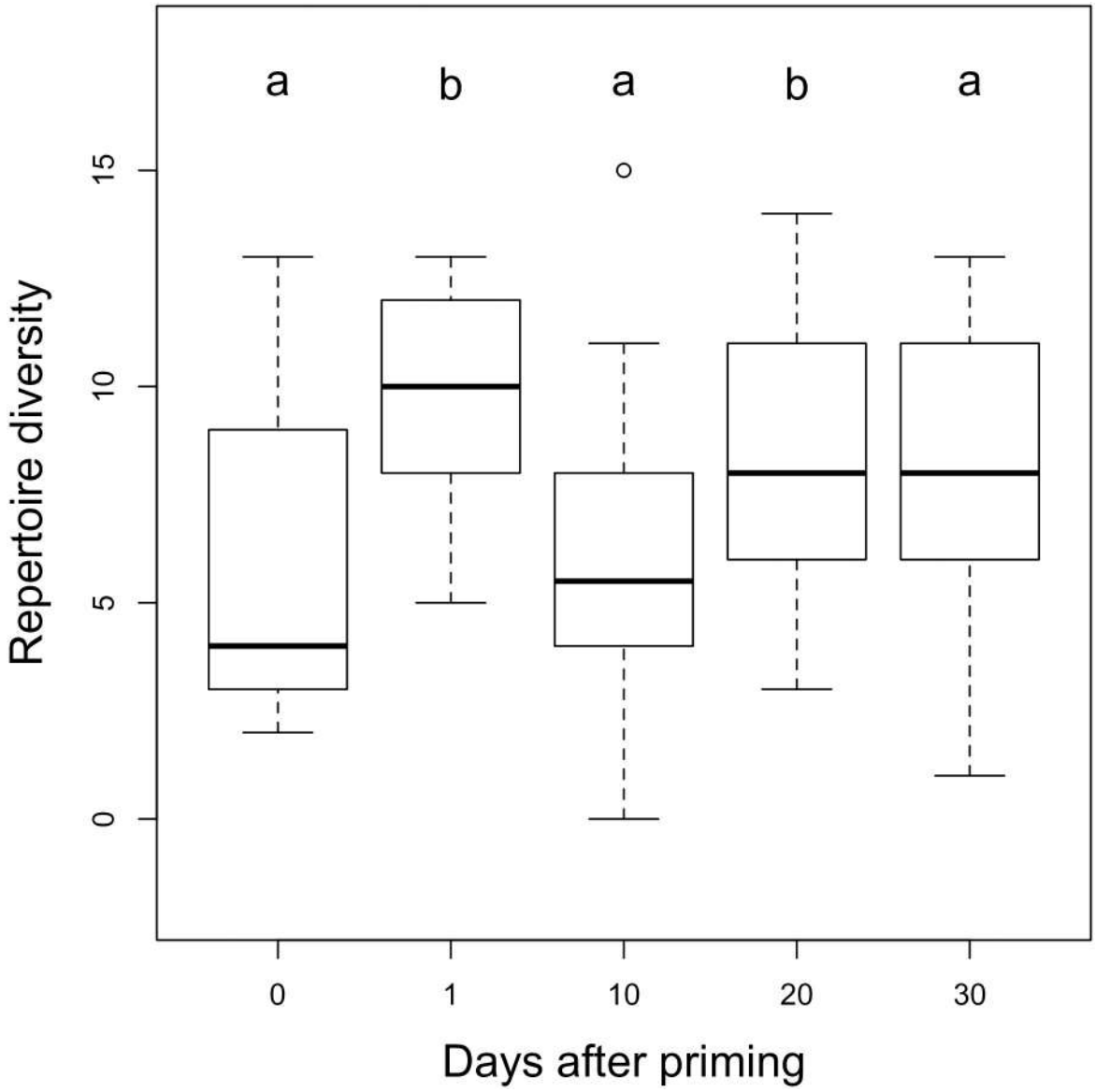

Fig 9. Repertoire diversity of USV with or without priming. Boxplots with medians of repertoire diversity of unprimed (0) and primed males (≥1). Different letters denote significant differences ($p<0.05$).

https://doi.org/10.1371/journal.pone.0242959.g009

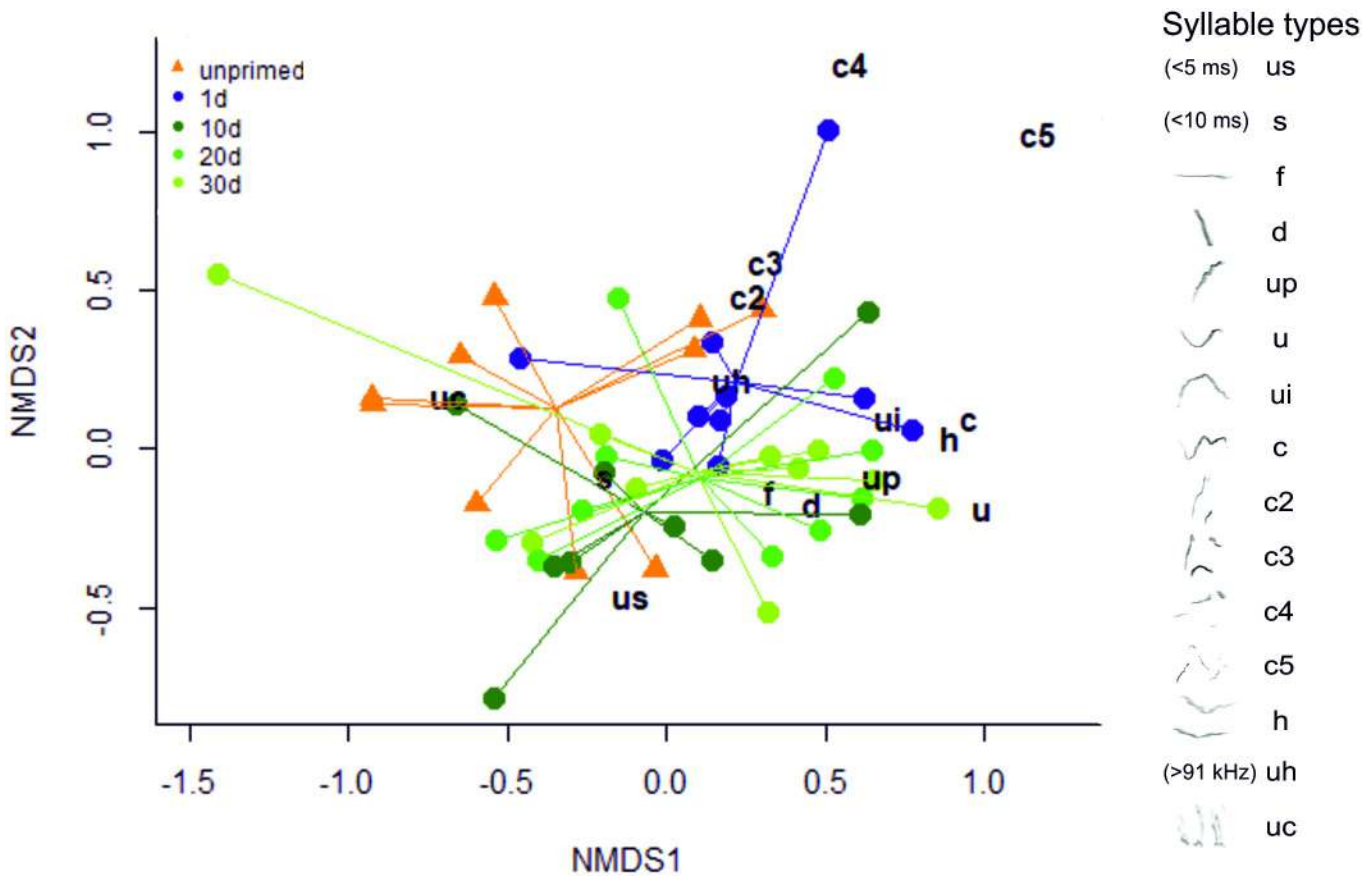

Fig 10. Non-metric multi-dimensional clustering of USV type according to priming groups (nMDS: Stress = 0.13). Mice are clustered according to the amount of each vocalization type emitted, and spectrograms of vocalization types are depicted on the legend on the right (see also Fig 1). Mice are color-coded by groups: unprimed males in orange triangles, 1d primed males in blue circles, 10d primed males in dark green circles, 20d primed males in green circles and 30d primed males in light green circles.

https://doi.org/10.1371/journal.pone.0242959.g010

boldness or risk-taking [88]. Male mice have been shown to alter the amount and types of USVs they emit after they detect a female or her scent and over the course of courtship and mating [15, 43]. Thus, the changes in male USVs induced by sexual stimuli may help to attract females and enhance their receptivity. One study found that females are attracted to playbacks of male vocalizations with more complex syllable types [33], and studies are now needed to investigate how females respond to other priming-induced changes in males USVs. The dynamic changes in male courtship USV emission after a sexual encounter might provide more reliable information about a male's identity (compatibility) or condition (quality).

In summary, our study is the first to experimentally test whether direct socio-sexual priming affects the USV emission of wild-derived house mice (*Mus musculus musculus)*, and the first to demonstrate that priming affects the repertoire diversity and composition, as well as the rate of vocalizations. We found that calls of primed males also showed altered USV spectro-temporal features, i.e., USV length, slope and variability in USV frequency. We found high individual variation in several vocalization parameters, as with previous studies of wild-derived mice (unlike many studies, we did not apply a screening procedure, such as omitting recordings of males that did not vocalize, or use threshold criteria for our analyses, as not to bias results). Given such variation, longitudinal measures are needed to further investigate priming effects on USV emission. Until then, our results suggest that USV studies should

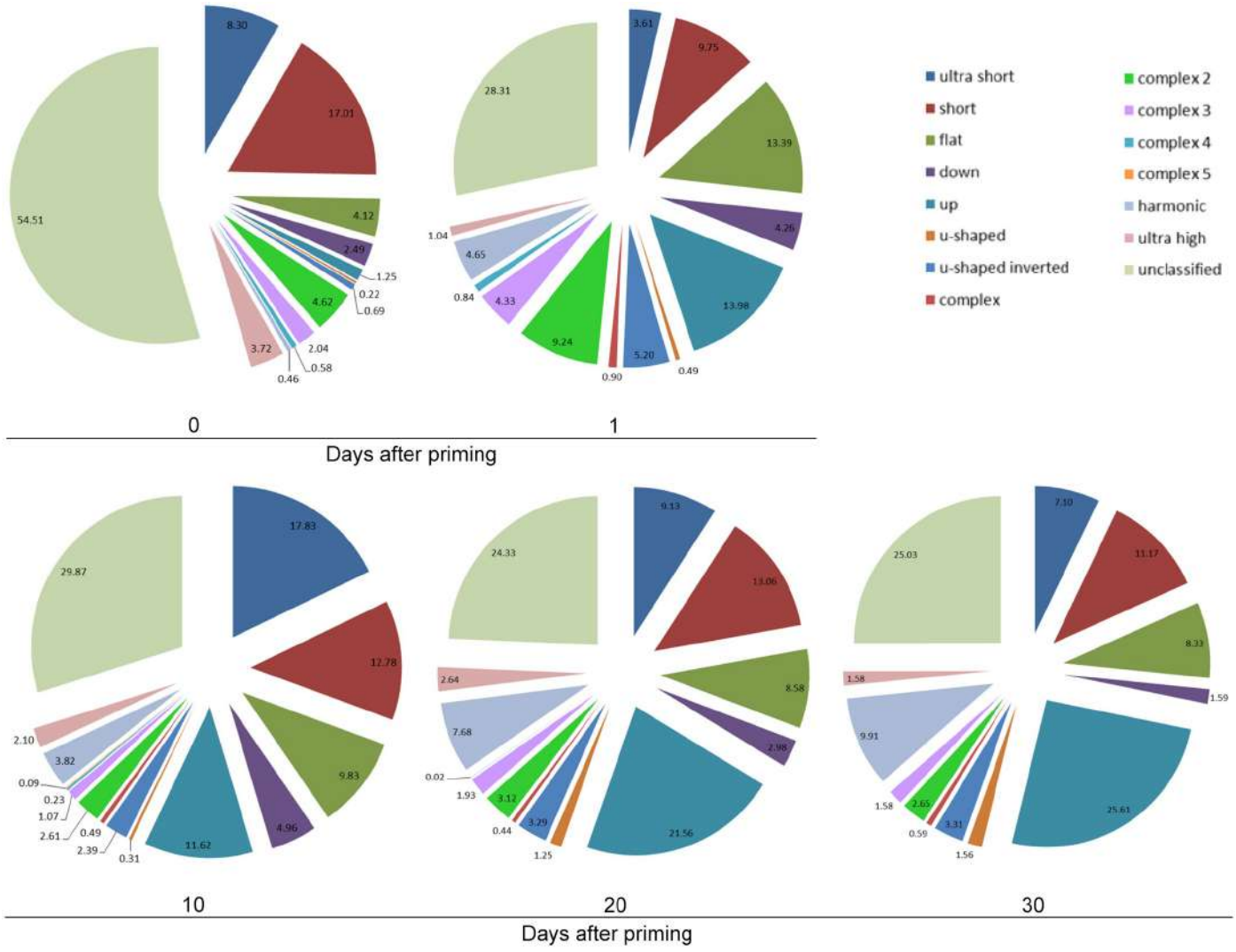

Fig 11. Proportions of the different types of vocalizations emitted by mice in the treatment and control groups. Pie charts show the mean proportions (%) of the occurrence of vocalization types emitted by each group, and the legend shows the 15 vocalization types (see also Fig 1).

https://doi.org/10.1371/journal.pone.0242959.g011

control differences in sexual priming (types of priming experience and duration after priming) as potential sources of variation.

Supporting information

S1 File.
(PDF)

S1 Data.
(XLSX)

S2 Data.
(XLSX)

Acknowledgments

We are grateful to T. Klaus and B. Wernisch for their help during the experimental and data processing work, H. Sasse and U. Madlsperger for animal care, and R. Hengsberger for help with the figures and formatting the manuscript. We thank F. Knauer for insightful discussions about statistical analyses, and M. Kalcounis-Rueppel and two anonymous reviewers for comments that improved the manuscript.

Author Contributions

Conceptualization: Sarah M. Zala, Doris Nicolakis.

Data curation: Thomas Ruf.

Formal analysis: Sarah M. Zala, Doris Nicolakis, Maria Adelaide Marconi, Thomas Ruf.

Funding acquisition: Sarah M. Zala, Dustin J. Penn.

Project administration: Sarah M. Zala, Dustin J. Penn.

Software: Anton Noll, Peter Balazs.

Writing – original draft: Sarah M. Zala, Dustin J. Penn.

Writing – review & editing: Sarah M. Zala, Doris Nicolakis, Maria Adelaide Marconi, Anton Noll, Peter Balazs, Dustin J. Penn.

References

1. Searcy WA, Andersson M. Sexual selection and the evolution of song. Annu Rev Ecol Syst. 1986; 17:507–33. http://www.jstor.org/stable/2097007.
2. Andersson M. Sexual selection. Princeton: Princeton University Press; 1994 1994. 599 p.
3. Gerhardt HC, Huber F. Acoustic Communication in Insects and Anurans. Common Problems and Diverse Solutions. Chicago: University of Chicago Press; 2002. 542 p.
4. Catchpole C, Slater PJB. Bird Song. Biological Themes and Variations. 2nd ed. Cambridge [England]; New York: Cambridge University Press; 2008. xi, 335 p. p.
5. Arch VS, Narins PM. 'Silent' signals: selective forces acting on ultrasonic communication systems in terrestrial vertebrates. Anim Behav. 2008; 76:1423–8. https://doi.org/10.1016/j.anbehav.2008.05.012
6. Feng AS, Narins PM, Xu CH, Lin WY, Yu ZL, Qiu Q, et al. Ultrasonic communication in frogs. Nature. 2006; 440(7082):333–6. https://doi.org/10.1038/nature04416
7. Portfors CV. Types and functions of ultrasonic vocalizations in laboratory rats and mice. J Am Assoc Lab Anim Sci. 2007; 46(1):28–34. https://www.ingentaconnect.com/contentone/aalas/jaalas/2007/00000046/00000001/art00005.
8. Musolf K, Penn DJ. Ultrasonic vocalizations in house mice: a cryptic mode of acoustic communication. In: Macholán M, Baird SJE, Munclinger P, Piálek J, editors. Evolution of the House Mouse. Cambridge Studies in Morphology and Molecules: New Paradigms in Evolutionary Biology. 3. Cambridge: Cambridge University Press; 2012. p. 253–77.
9. Holy TE, Guo Z. Ultrasonic Songs of Male Mice. PLoS Biol. 2005; 3(12):e386. https://doi.org/10.1371/journal.pbio.0030386
10. Heckman J, McGuinness B, Celikel T, Englitz B. Determinants of the mouse ultrasonic vocal structure and repertoire. Neurosci Biobehav Rev. 2016; 65:313–25. https://doi.org/10.1016/j.neubiorev.2016.03.029
11. Whitney G, Alpern M, Dizinno G, Horowitz G. Female odors evoke ultrasounds from male mice. Anim Learn Behav. 1974; 2(1):13–8. https://doi.org/10.3758/BF03199109
12. Nyby J, Wysocki CJ, Whitney G, Dizinno G. Pheromonal regulation of male mouse ultrasonic courtship (*Mus musculus*). Anim Behav. 1977; 25(2):333–41. https://doi.org/10.1016/0003-3472(77)90009-4
13. Wang H, Liang S, Burgdorf J, Wess J, Yeomans J. Ultrasonic Vocalizations Induced by Sex and Amphetamine in M2, M4, M5 Muscarinic and D2 Dopamine Receptor Knockout Mice. PLoS ONE. 2008; 3(4):e1893. https://doi.org/10.1371/journal.pone.0001893

14. Musolf K, Hoffmann F, Penn DJ. Ultrasonic courtship vocalizations in wild house mice, *Mus musculus*. Anim Behav. 2010; 79(3):757–64. https://doi.org/10.1016/j.anbehav.2009.12.034

15. Matsumoto YK, Okanoya K. Phase-Specific Vocalizations of Male Mice at the Initial Encounter during the Courtship Sequence. PLoS ONE. 2016; 11(2):e0147102. https://doi.org/10.1371/journal.pone.0147102

16. Hanson JL, Hurley LM. Female Presence and Estrous State Influence Mouse Ultrasonic Courtship Vocalizations. PLoS ONE. 2012; 7(7):e40782. https://doi.org/10.1371/journal.pone.0040782

17. Kanno K, Kikusui T. Effect of Sociosexual Experience and Aging on Number of Courtship Ultrasonic Vocalizations in Male Mice. Zoolog Sci. 2018; 35(3):208–14. https://doi.org/10.2108/zs170175

18. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (*Mus musculus musculus*). PLoS ONE. 2017; 12(12):e0188647. https://doi.org/10.1371/journal.pone.0188647

19. Warren MR, Spurrier MS, Roth ED, Neunuebel JP. Sex differences in vocal communication of freely interacting adult mice depend upon behavioral context. PLoS ONE. 2018; 13(9):e0204527. https://doi.org/10.1371/journal.pone.0204527

20. Nyby J. Ultrasonic Vocalizations during Sex Behavior of Male House Mice (*Mus musculus*): A Description. Behav Neural Biol. 1983; 39(1):128–34. https://doi.org/10.1016/S0163-1047(83)90722-7

21. Barthelemy M, Gourbal BEF, Gabrion C, Petit G. Influence of the female sexual cycle on BALB/c mouse calling behaviour during mating. Naturwissenschaften. 2004; 91(3):135–8. https://doi.org/10.1007/s00114-004-0501-4

22. Sirotin YB, Costa ME, Laplagne DA. Rodent ultrasonic vocalizations are bound to active sniffing behavior. Front Behav Neurosci. 2014; 8:399-. https://doi.org/10.3389/fnbeh.2014.00399

23. Sales GD, Pye JD. Ultrasonic Communication by Animals. London: Chapman and Hall; 1974. 281 p.

24. White NR, Prasad M, Barfield RJ, Nyby JG. 40- and 70-kHz Vocalizations of Mice (*Mus musculus*) during Copulation. Physiol Behav. 1998; 63(4):467–73. https://doi.org/10.1016/S0031-9384(97)00484-8

25. Hammerschmidt K, Radyushkin K, Ehrenreich H, Fischer J. Female mice respond to male ultrasonic 'songs' with approach behaviour. Biol Lett. 2009; 5(5):589–92. https://doi.org/10.1098/rsbl.2009.0317

26. Shepard KN, Liu RC. Experience restores innate female preference for male ultrasonic vocalizations. Genes Brain Behav. 2011; 10(1):28–34. https://doi.org/10.1111/j.1601-183X.2010.00580.x

27. Sugimoto H, Okabe S, Kato M, Koshida N, Shiroishi T, Mogi K, et al. A Role for Strain Differences in Waveforms of Ultrasonic Vocalizations during Male-Female Interaction. PLoS ONE. 2011; 6(7):e22093. https://doi.org/10.1371/journal.pone.0022093

28. Nicolakis D, Marconi MA, Zala SM, Penn DJ. Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success. Front Zool. 2020; 17:10. https://doi.org/10.1186/s12983-020-00353-1

29. Asaba A, Osakada T, Touhara K, Kato M, Mogi K, Kikusui T. Male mice ultrasonic vocalizations enhance female sexual approach and hypothalamic kisspeptin neuron activity. Horm Behav. 2017; 94:53–60. https://doi.org/10.1016/j.yhbeh.2017.06.006

30. Musolf K, Meindl S, Larsen AL, Kalcounis-Rueppell MC, Penn DJ. Ultrasonic Vocalizations of Male Mice Differ among Species and Females Show Assortative Preferences for Male Calls. PLoS ONE. 2015; 10(8):e0134123. https://doi.org/10.1371/journal.pone.0134123

31. Hammerschmidt K, Reisinger E, Westekemper K, Ehrenreich L, Strenzke N, Fischer J. Mice do not require auditory input for the normal development of their ultrasonic vocalizations. BMC Neurosci. 2012; 13(1):40. https://doi.org/10.1186/1471-2202-13-40

32. Hurley LM, Kalcounis-Rueppell MC. State and Context in Vocal Communication of Rodents. In: Dent ML, Fay RR, Popper AN, editors. Rodent Bioacoustics. 67. Cham: Springer International Publishing; 2018. p. 191–221.

33. Gaub S, Fisher SE, Ehret G. Ultrasonic vocalizations of adult male *Foxp2*-mutant mice: behavioral contexts of arousal and emotion. Genes Brain and Behavior. 2016; 15(2):243–59.

34. Dizinno G, Whitney G, Nyby J. Ultrasonic Vocalizations by Male Mice (*Mus musculus*) to Female Sex Pheromone: Experiential Determinants. Behav Biol. 1978; 22(1):104–13. https://doi.org/10.1016/S0091-6773(78)92094-1

35. Nyby J, Whitney G. Experience Affects Behavioral Responses to Sex Odors. In: Müller-Schwarze D, Silverstein RM, editors. Chemical Signals: Springer US; 1980. p. 173–92.

36. Roullet FI, Wöhr M, Crawley JN. Female urine-induced male mice ultrasonic vocalizations, but not scent-marking, is modulated by social experience. Behav Brain Res. 2011; 216(1):19–28. https://doi.org/10.1016/j.bbr.2010.06.004

37. Maggio JC, Whitney G. Ultrasonic Vocalizing by Adult Female Mice (*Mus musculus*). J Comp Psychol. 1985; 99(4):420–36. https://doi.org/10.1037/0735-7036.99.4.420

38. Sipos ML, Kerchner M, Nyby JG. An Ephemeral Sex Pheromone in the Urine of Female House Mice (*Mus domesticus*). Behav Neural Biol. 1992; 58(2):138–43. https://doi.org/10.1016/0163-1047(92)90375-E

39. Nyby J, Bigelow J, Kerchner M, Barbehenn F. Male Mouse (*Mus musculus*) Ultrasonic Vocalizations to Female Urine: Why is Heterosexual Experience Necessary? Behav Neural Biol. 1983; 38:32–46. https://doi.org/10.1016/S0163-1047(83)90354-0

40. Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Front Behav Neurosci. 2015; 9:76.

41. Arriaga G, Zhou EP, Jarvis ED. Of Mice, Birds, and Men: The Mouse Ultrasonic Song System Has Some Features Similar to Humans and Song-Learning Birds. PLoS ONE. 2012; 7(10):e46610. https://doi.org/10.1371/journal.pone.0046610

42. Burke K, Screven LA, Dent ML. CBA/CaJ mouse ultrasonic vocalizations depend on prior social experience. PLoS ONE. 2018; 13(6):e0197774. https://doi.org/10.1371/journal.pone.0197774

43. Marconi MA, Nicolakis D, Abbasi R, Penn DJ, Zala SM. Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures. Anim Behav. 2020; 169:169–97. https://doi.org/10.1016/j.anbehav.2020.09.006

44. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Automatic mouse ultrasound detector (A-MUD): A new tool for processing rodent vocalizations. PLoS ONE. 2017; 12(7):e0181200. https://doi.org/10.1371/journal.pone.0181200

45. Fernández-Vargas M. Vocal Signals of Sexual Motivation in Male and Female Rodents. Curr Sex Health Rep. 2018; 10(4):315–28. https://doi.org/10.1007/s11930-018-0179-9

46. Zala SM, Bilak A, Perkins M, Potts WK, Penn DJ. Female house mice initially shun infected males, but do not avoid mating with them. Behav Ecol Sociobiol. 2015; 69:715–22. https://doi.org/10.1007/s00265-015-1884-2

47. Moosmüller S, Brandstätter J. Phonotactic information in the temporal organization of Standard Austrian German and the Viennese dialect. Lang Sci. 2014; 46(A):84–95. https://doi.org/10.1016/j.langsci.2014.06.016

48. Moosmüller S. Evaluation of the speech behaviour of reference speakers. Int J Speech, Lang Law. 2012; 18(2):179–205. https://doi.org/10.1558/ijsll.v18i2.179

49. Gygi B, Deutsch W, Noll A, Hirner A, Becker T, editors. NOIDESc: Incorporating Feature Descriptors into a Novel Railway Noise Evaluation Scheme. 19th International Congress on Acoustics; 2007 2–7 Sept. 2007; Madrid.

50. Kasess C, Noll A, Majdak P, Waubke H. Effect of train type on annoyance and acoustic features of the rolling noise. J Acoust Soc Am. 2013; 134(2):1071–81. https://doi.org/10.1121/1.4812771

51. Balazs P, Laback B, Eckel G, Deutsch WA. Introducing Time-Frequency Sparsity by Removing Perceptually Irrelevant Components Using a Simple Model of Simultaneous Masking. IEEE/ACM Trans Audio, Speech, Language Process. 2010; 18(1):34–49.

52. Grimsley JMS, Sheth S, Vallabh N, Grimsley CA, Bhattal J, Latsko M, et al. Contextual Modulation of Vocal Behavior in Mouse: Newly Identified 12 kHz "Mid-Frequency" Vocalization Emitted during Restraint. Front Behav Neurosci. 2016; 10:38. https://doi.org/10.3389/fnbeh.2016.00038

53. Neilans EG, Holfoth DP, Radziwon KE, Portfors CV, Dent ML. Discrimination of Ultrasonic Vocalizations by CBA/CaJ Mice (*Mus musculus*) Is Related to Spectrotemporal Dissimilarity of Vocalizations. PLoS ONE. 2014; 9(1):e85405. https://doi.org/10.1371/journal.pone.0085405

54. Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism. PLoS ONE. 2008; 3(8):e3067. https://doi.org/10.1371/journal.pone.0003067

55. Hoffmann F, Musolf K, Penn DJ. Ultrasonic courtship vocalizations in wild house mice: spectrographic analyses. J Ethol. 2012; 30(1):173–80. https://doi.org/10.1007/s10164-011-0312-y

56. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes Brain Behav. 2011; 10(1):44–56. https://doi.org/10.1111/j.1601-183X.2010.00623.x

57. R Core Team. R: A language and environment for statistical computing. 2019. Vienna, Austria. R Foundation for Statistical Computing. http://www.R-project.org/

58. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. 495 p.

59. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. J Stat Softw. 2015; 64(4):1–34. https://doi.org/10.18637/jss.v064.i04

60. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015; 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

61. Barton K. MuMIn: Multi-model inference. R package version 1.43.6 2019. http://CRAN.R-project.org/package=MuMIn.

62. Burnham KP, Anderson DR. Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach. 2nd ed. New York: Springer; 2002. XXVI, 488 p.

63. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Efffects Models and Extensions in Ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors. New York: Springer; 2009.

64. R Core Team. R: A language and environment for statistical computing. 2018. Vienna, Austria. R Foundation for Statistical Computing. http://www.R-project.org/

65. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2018. R package version 2.5–2. http://CRAN.R-project.org/package=vegan

66. Clarke KR, Warwick RM. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation. 2nd ed. Plymouth, UK: Primer-E Ltd.; 2001. 176 p.

67. Arndt SS, Laarakker MC, van Lith HA, van der Staay FJ, Gieling E, Salomons AR, et al. Individual housing of mice—Impact on behaviour and stress responses. Physiol Behav. 2009; 97(3):385–93. https://doi.org/10.1016/j.physbeh.2009.03.008

68. Miczek KA, Maxson SC, Fish EW, Faccidomo S. Aggressive behavioral phenotypes in mice. Behav Brain Res. 2001; 125(1–2):167–81. https://doi.org/10.1016/S0166-4328(01)00298-4

69. Võikar V, Polus A, Vasar E, Rauvala H. Long-term individual housing in C57BL/6J and DBA/2 mice: assessment of behavioral consequences. Genes Brain Behav. 2005; 4(4):240–52.

70. Keesom SM, Finton CJ, Sell GL, Hurley LM. Early-Life Social Isolation Influences Mouse Ultrasonic Vocalizations during Male-Male Social Encounters. PLoS ONE. 2017; 12(1):e0169705. https://doi.org/10.1371/journal.pone.0169705

71. Sipos ML, Nyby JG. Concurrent androgenic stimulation of the ventral tegmental area and medial preoptic area: Synergistic effects on male-typical reproductive behaviors in house mice. Brain Res. 1996; 729 (1):29–44. https://doi.org/10.1016/0006-8993(96)00148-5

72. Sipos ML, Nyby JG. Intracranial Androgenic Activation of Male-Typical Behaviours in House Mice: Concurrent Stimulation of the Medial Preoptic Area and Medial Nucleus of the Amygdala. J Neuroendocrinol. 1998; 10(8):577–86.

73. Pfaus JG, Kippin TE, Centeno S. Conditioning and Sexual Behavior: A Review. Horm Behav. 2001; 40 (2):291–321. https://doi.org/10.1006/hbeh.2001.1686

74. Coquelin A, Bronson FH. Release of Luteinizing Hormone in Male Mice During Exposure to Females: Habituation of the Response. Science. 1979; 206(4422):1099–101. https://doi.org/10.1126/science.573924

75. Coquelin A, Bronson FH. Secretion of Luteinizing Hormone in Male Mice: Factors that Influence Release during Sexual Encounters. Endocrinology. 1980; 106(4):1224–9.

76. Batty J. Acute changes in plasma testosterone levels and their relation to measures of sexual behaviour in the male mouse (Mus musculus). Anim Behav. 1978; 26(2):349–57. https://doi.org/10.1016/0003-3472(78)90053-2

77. Macrides F, Bartke A, Dalterio S. Strange Females Increase Plasma Testosterone Levels in Male Mice. Science. 1975; 189(4208):1104–6. https://doi.org/10.1126/science.1162363

78. Nyby JG. Reflexive Testosterone Release: A Model System for Studying the Nongenomic Effects of Testosterone Upon Male Behavior. Front Neuroendocrin. 2008; 29(2):199–210. https://doi.org/10.1016/j.yfrne.2007.09.001

79. James PJ, Nyby JG, Saviolakis GA. Sexually stimulated testosterone release in male mice (*Mus musculus*): Roles of genotype and sexual arousal. Horm Behav. 2006; 50(3):424–31. https://doi.org/10.1016/j.yhbeh.2006.05.004

80. Murphy AZ, Shupnik MA, Hoffman GE. Androgen and Estrogen (α) Receptor Distribution in the Periaqueductal Gray of the Male Rat. Horm Behav. 1999; 36(2):98–108. https://doi.org/10.1006/hbeh.1999.1528

81. Swaney WT, Dubose BN, Curley JP, Champagne FA. Sexual experience affects reproductive behavior and preoptic androgen receptors in male mice. Horm Behav. 2012; 61(4):472–8. https://doi.org/10.1016/j.yhbeh.2012.01.001

82. Gao S-C, Wei Y-C, Wang S-R, Xu X-H. Medial Preoptic Area Modulates Courtship Ultrasonic Vocalization in Adult Male Mice. Neurosci Bull. 2019; 35:697–708. https://doi.org/10.1007/s12264-019-00365-w

83. Tschida K, Michael V, Takatoh J, Han B-X, Zhao S, Sakurai K, et al. A Specialized Neural Circuit Gates Social Vocalizations in the Mouse. Neuron. 2019; 103(3):459–72.e4. https://doi.org/10.1016/j.neuron.2019.05.025

84. Zala SM, Potts WK, Penn DJ. Scent-marking displays provide honest signals of health and infection. Behav Ecol. 2004; 15(2):338–44.

85. Koyama S, Kamimura S. Influence of social dominance and female odor on the sperm activity of male mice. Physiol Behav. 2000; 71(3–4):415–22. https://doi.org/10.1016/S0031-9384(00)00361-9

86. James PJ, Nyby JG. Testosterone rapidly affects the expression of copulatory behavior in house mice (*Mus musculus*). Physiol Behav. 2002; 75(3):287–94. https://doi.org/10.1016/S0031-9384(01)00666-7

87. Whitney G, Coble JR, Stockton MD, Tilson EF. Ultrasonic emissions: do they facilitate courtship of mice. J Comp Physiol Psychol. 1973; 84(3):445–52. https://doi.org/10.1037/h0034899

88. Kavaliers M, Choleris E, Colwell DD. Brief Exposure to Female Odors "Emboldens"Male Mice by Reducing Predator-Induced Behavioral and Hormonal Responses. Horm Behav. 2001; 40(4):497–509. https://doi.org/10.1006/hbeh.2001.1714

CHAPTER 3: Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success

Nicolakis, Doris; Marconi, Maria Adelaide;
Zala, Sarah, M.; Penn, Dustin J.

Frontiers in Zoology: Impact factor: 3.3; **published**

Nicolakis et al. Frontiers in Zoology (2020) 17:10
https://doi.org/10.1186/s12983-020-00353-1

Frontiers in Zoology

RESEARCH Open Access

Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success

Doris Nicolakis*, Maria Adelaide Marconi, Sarah M. Zala*† and Dustin J. Penn†

Abstract

Background: Courtship vocalizations are used by males of many species to attract and influence the behavior of potential mating partners. Our aim here was to investigate the modulation and reproductive consequences of courtship ultrasonic vocalizations (USVs) in wild-derived house mice (*Mus musculus musculus*). The courtship USVs of male mice are surprisingly complex and are composed of a variety of different syllable types. Our specific aims were to test whether (1) the emission of courtship USVs depends upon the kinship of a potential mating partner, and (2) whether USV emission during courtship affects the pairs' subsequent reproductive success.

Results: We experimentally presented males with an unfamiliar female that was either genetically related or unrelated, and we recorded USV emission, first while the sexes were separated by a perforated partition and then during direct interactions, after removing the partition. USVs were detected by the Automatic Mouse Ultrasound Detector (A-MUD) and manually classified into 15 syllable types. The mice were kept together to test whether and how courtship vocalizations predict their subsequent reproductive success. We found that the mice significantly increased their amount of vocalizations (vocal performance) and number of syllable types (vocal repertoire) after the partition was removed and they began interacting directly. We show that unrelated pairs emitted longer and more complex USVs compared to related pairs during direct interactions. Unrelated pairs also had a greater reproductive success compared to related pairs, and in addition we found a negative correlation between the mean length and amount of vocalizations with the latency to their first litter.

Conclusion: Our study provides evidence that house mice modulate the emission of courtship USVs depending upon the kinship of potential mating partners, and that courtship USVs correlate with reproductive success.

Keywords: *Mus musculus musculus*, House mice, Wild derived, USV, Ultrasonic vocalizations, Reproductive success, Genetic relatedness, Adaptive functions

* Correspondence: Doris.Nicolakis@vetmeduni.ac.at; Sarah.Zala@vetmeduni.ac.at
†Sarah M. Zala and Dustin J. Penn contributed equally to this work.
Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine, Vienna, Savoyenstraße 1a, 1160 Vienna, Austria

Background

Courtship vocalizations are produced in many species, usually by males, as a mechanism to attract and influence the behavior of potential mates [1, 2]. In some birds, exposure to male courtship song can induce ovarian development in females [3] and copulation solicitation behaviors [4]. Courtship vocalizations can reveal a surprising amount of information about a male to potential mates, including their fertility, genetic quality and species or individual identity [reviewed in 1]. Many studies on sexual selection in mammals, however, have focused on vocalizations emitted during competitive male-male interactions, and though there is increasing evidence for female-choice [5], these studies have mainly investigated chemical signals (olfactory communication). Surprisingly little is known about whether and how mammalian vocalizations evolve through female choice. Here, we investigated the functions of the courtship ultrasonic vocalizations (USVs) of wild-derived male house mice (*Mus musculus musculus).*

The USVs of male house mice are complex and have features similar to birdsong [6]. Males emit USVs mainly during courtship and mating, however, their functions are still unclear (reviewed in [7–10]). The vast majority of studies on mouse USVs have been conducted on inbred laboratory strains (*Mus laboratorius)*, and USVs are often used as a tool to investigate neurodevelopmental and speech disorders [11, 12]. Previous studies suggest that USVs provide a reliable signal of male sexual arousal or motivation (reviewed in [13]). The complexity of male USVs is increased during the course of courtship and particularly just before copulation [14] and mice of both sexes emit vocalizations at a higher rate and higher frequencies during opposite- compared to same-sex interactions [15]. Female mice show approach behavior towards playbacks of male vocalizations [16–18], and they show preferences for USVs with more frequency jumps [19]. Several hypotheses have been proposed to explain why females are attracted to male USVs (reviewed in [7]), and so far playback studies provide evidence for two potential functions: (1) species recognition, as *Mus musculus* females are more attracted to playbacks of male USV of their own species compared to those of *Mus spicilegus* [20]; and (2) kin recognition, as females are more attracted to the USVs of non-kin compared to those of their siblings [17]. In this study we aimed to test whether males modulate their USV emission depending upon the genetic compatibility of a potential mating partner (kin recognition), and we tested whether courtship vocalizations predict a mating pairs' subsequent reproductive success.

It has long been suggested that the courtship USVs of male mice influence mating and reproductive success, and yet only two studies have tested this hypothesis to our knowledge. First, Asaba et al. [18] recorded vocalizations during interactions of males with a female after being housed with a different female for 4 months. They found a correlation between the number of deliveries during this time and the number of USVs males emitted when they were later recorded with the other female. However, it is unclear whether male USV emission influenced male mating success, or vice versa. Second, Kanno and Kikusui [21] recorded males with a novel virgin female both before and after they were housed with another female for 2 weeks. They compared males that emitted USVs with males that did not vocalize before or after co-housing, and they found that vocalizing males sired more offspring than non-vocalizing males during the co-housing phase. This study recorded males before and after housing with a female, but only compared vocalizing versus non-vocalizing males. Thus, it is still unknown whether any other variation in male USVs predicts reproductive success. Also, these studies were both conducted on laboratory mice (C57BL/6 J), which are very different from wild mice (laboratory strains are selected for rapid reproduction, and differ in their vocalizations, courtship and mating behavior), and therefore, we aimed to investigate the adaptive functions of courtship USVs in wild-derived house mice.

In the present study we experimentally manipulated the genetic compatibility (relatedness) of breeding pairs by presenting males with an unfamiliar female, which was either genetically related or not (no-choice mate preference). We recorded the USVs emitted both, before and after removing a perforated divider, and we then tested whether the mice modulate the emission of courtship USVs depending on their genetic relatedness. This experiment allowed us to test whether males alter their USV emission during the early phases of courtship and also to test whether males are able to recognize and show preferences for unrelated over genetically related potential mating partners. We expected that if the mice show kin recognition and inbreeding avoidance, then they will emit more USVs and a more complex repertoire when paired with unrelated females. Finally, we tested whether male courtship USV emissions influence the pairs' subsequent reproductive success.

Results

Phases of courtship

We first investigated whether and how the males modulated their USV emission when they were presented with a female, first while separated by a clear, perforated divider (*introduction phase*) and then during direct contact, after the divider was removed (*interaction phase*). Overall, the mice emitted 5x more USVs and produced more types of syllables during than before direct interactions (vocal performance: Wilcoxon test: $n = 26$, $Z = -3.264$, $p = 0.001$, Fig. 1a; vocal repertoire: Wilcoxon test: $n = 26$,

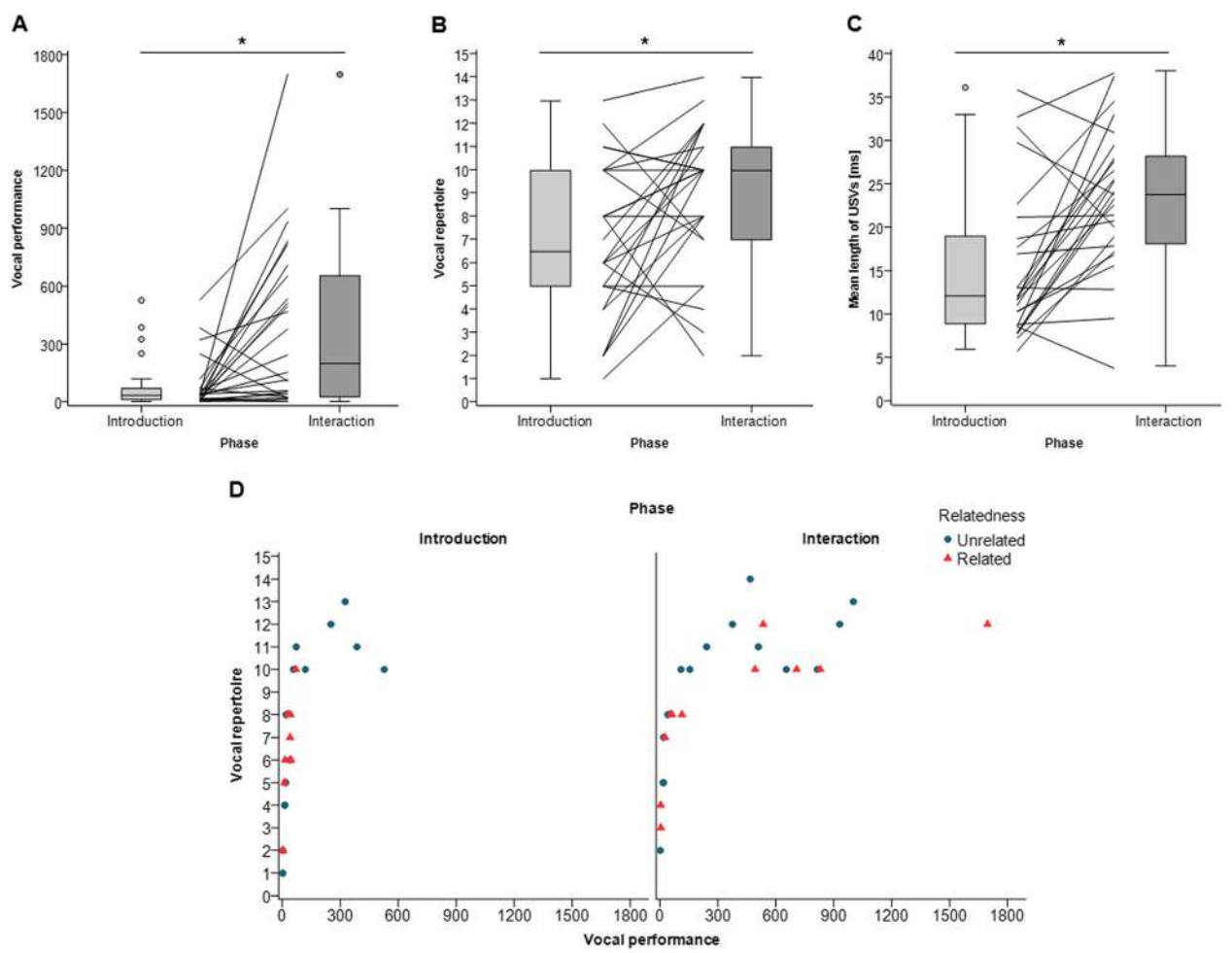

Fig. 1 USV emission during the introduction phase compared to the interaction phase. Boxplots show median (center line), interquartile range (box), 95% variation (whiskers) and outliers (circles) of (**a**) the vocal performance (total number of USVs emitted during 10 min), (**b**) the vocal repertoire (number of syllable types emitted during 10 min) and (**c**) the mean length of USVs (ms). **d** Relationship between the total number of USVs (vocal performance) and number of syllable types (vocal repertoire) emitted by unrelated (blue circles) and related (red triangles) pairs during the introduction (left) and interaction phase (right). * $p < 0.05$

Z = – 3.912, $p < 0.001$, Fig. 1b; Additional file 1: Table S1). In both phases there was a positive correlation between vocal performance and vocal repertoire, so that the mice that emitted more USVs also emitted more syllable types (Spearman correlation: introduction: $n = 26$, $r_s = 0.926$, $p < 0.001$, interaction: $n = 26$, $r_s = 0.852$, $p < 0.001$) (Fig. 1d). The vocal repertoire first increased with the number of vocalizations but then plateaued after circa 10 syllable types. Hence, the relationship between vocal performance and repertoire follows a logarithmic curve. During direct interactions the mice also emitted longer syllables compared to the introduction phase (Wilcoxon test: $n = 26$, Z = – 3.467, $p = 0.001$, Additional file 1: Table S1) (Fig. 1c). Therefore, we examined these two phases separately for our subsequent analyses.

Female sexual receptivity

We examined whether USV emission was influenced by female estrous state. During the introduction phase, vocal performance did not differ when males were exposed to females of any of the four estrous states (Kruskal-Wallis test: $n = 26$, $\chi^2 = 3.169$, $p = 0.366$). Visual inspection suggests that during the interaction phase more USVs were emitted when females were in proestrus than other stages; however, there was also no significant difference among the four estrous states (Kruskal-Wallis test: $n = 26$, $\chi^2 = 5.469$, $p = 0.141$). Next, we combined females in proestrus and estrus into "receptive females" and females in metestrus and diestrus into "unreceptive females." Males had a higher median vocal performance during direct interactions with a receptive female compared to unreceptive females, however, this difference was not statistically significant (Mann-Whitney

U test: $n = 26$, introduction: $Z = -1.313$, $p = 0.189$, interaction: $n = 26$, $Z = -1.698$, $p = 0.090$, Additional file 1: Table S2) (Fig. 2a). Female receptivity had no significant effect on the mean length of USVs during either phase (Mann-Whitney U test: introduction: $n = 26$, $Z = -1.234$, $p = 0.217$, interaction: $n = 26$, $Z = -0.309$, $p = 0.758$, Additional file 1: Table S2) (Fig. 2b). Mice produced a larger vocal repertoire when presented with an unreceptive female (vs. a receptive female) during the introduction phase (Mann-Whitney U test: $n = 26$, $Z = -2.434$, $p = 0.015$), but not during the direct interactions (Mann-Whitney U test: $n = 26$, $Z = -1.643$, $p = 0.100$) (Fig. 2c, Additional file 1: Table S2). Female receptivity also influenced the grand mean frequency of USVs emitted during introduction (Mann-Whitney U test: $n = 25$, $Z = -2.502$, $p = 0.012$) but not during direct interactions (Mann-Whitney U test: $n = 26$, $Z = -0.463$, $p = 0.643$) (Fig. 2d, Additional file 1: Table S2) such that USVs emitted in the presence of receptive females had a lower grand mean frequency (50.67 ± 13.42 kHz) compared to unreceptive females (62.42 ± 7.73 kHz).

Genetic relatedness

We next tested whether USVs were modulated by presentation of a genetically related or unrelated partner. During the introduction phase, males tended to have a higher vocal performance when presented with an unrelated female compared to a related female (Welch's t-test: $n = 26$, $t = 1.963$, $p = 0.066$), though not during direct interactions (t-test: $n = 26$, $t = -0.038$, $p = 0.9$) (Fig. 3a, Additional file 1: Table S3). This trend was mainly due to males emitting more simple syllables when presented with unrelated

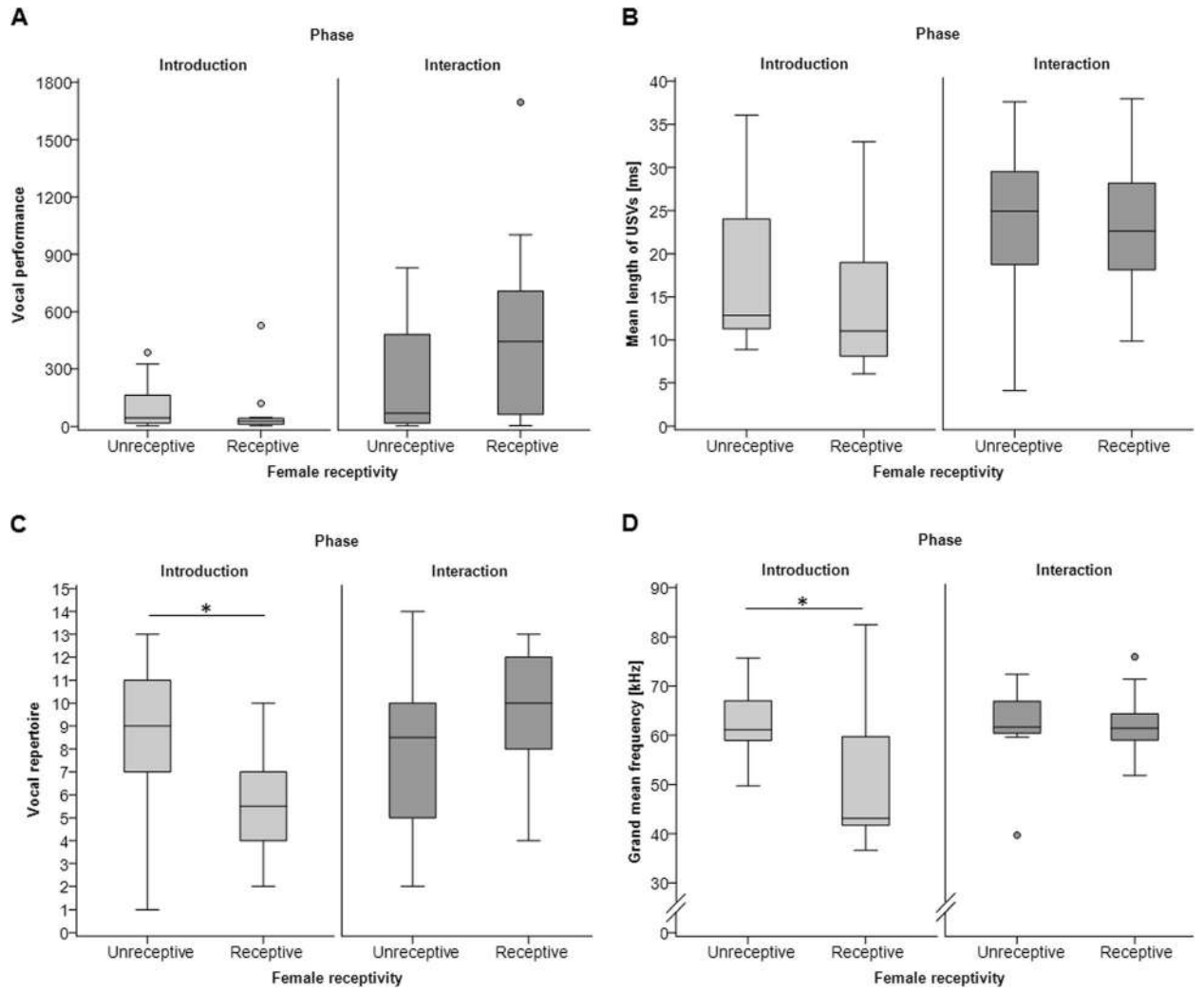

Fig. 2 USV emission during the introduction versus interaction phase when females were unreceptive or receptive. Boxplots show median (center line), interquartile range (box), 95% variation (whiskers) and outliers (circles) of (**a**) the vocal performance (total number of USVs emitted during 10 min), (**b**) the mean length of USVs (ms) (**c**) the vocal repertoire (number of syllable types emitted during 10 min) and (**d**) the grand mean frequency of USVs (kHz). * $p < 0.05$

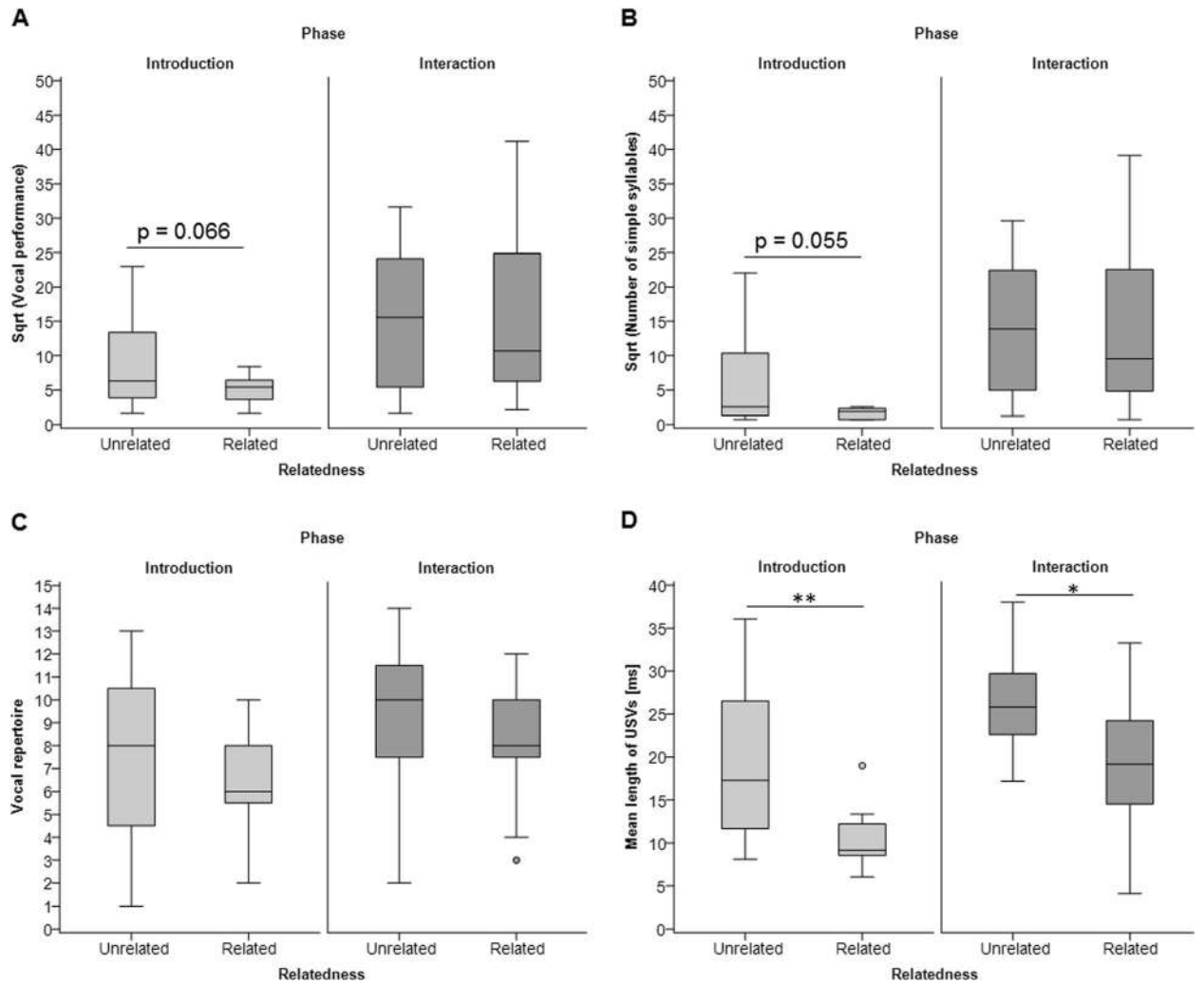

Fig. 3 USV emission of unrelated and related pairs during the introduction versus interaction phase. Boxplots show median (center line), interquartile range (box), 95% variation (whiskers) and outliers (circles) of the (**a**) vocal performance (square-root-transformed total number of USVs emitted during 10 min) (**b**) number of simple syllables (square-root-transformed total number of simple syllables emitted during 10 min), (**c**) vocal repertoire (number of syllable types emitted during 10 min) and (**d**) mean length of USVs (ms). * $p < 0.05$, ** $p < 0.01$

compared to related females in the introduction phase (introduction: Mann-Whitney U test: $n = 26$, $Z = -1.917$, $p = 0.055$**,** interaction: t-test: $n = 26$, $t = -0.005$, $p = 0.996$) (Fig. 3b, see Additional file 1: Table S3). The vocal repertoire did not differ between unrelated and related pairs in any phase (introduction: Welch's t-test: $n = 26$, $t = 1.035$, $p = 0.311$, interaction: t-test: $n = 26$, $t = 0.773$, $p = 0.447$) (Fig. 3c, Additional file 1: Table S3), however, unrelated mice always emitted longer USVs than related mice in both phases (introduction: Welch's t-test: $n = 26$, $t = 3.161$, $p = 0.005$, interaction: t-test: $n = 26$, $t = 2.449$, $p = 0.020$) (Fig. 3d, Additional file 1: Table S3). These results were not influenced by female estrous state as there was no interaction between female receptivity and relatedness to the male (GZLM, interaction of receptivity*relatedness: vocal performance: introduction: $n = 26$, Wald-$\chi^2 = 0.133$, $p = 0.715$, interaction: $n = 26$, Wald-$\chi^2 = 0.756$, $p = 0.388$; vocal repertoire: introduction: $n = 26$, Wald-$\chi^2 = 0.006$, $p = 0.937$, interaction: $n = 26$, Wald-$\chi^2 = 0.446$, $p = 0.504$; mean USV length: introduction: $n = 26$, Wald-$\chi^2 = 0.290$, $p = 0.590$, interaction: $n = 26$, Wald-$\chi^2 = 0.017$, $p = 0.896$; Additional file 1: Table S4).

We further investigated whether other features of USVs were influenced by a pair's genetic relatedness running a discriminant function analysis (DFA) with the following parameters: mean USV length (ms), grand mean USV frequency (kHz), vocal repertoire, total number of short syllables (square-root-transformed), total number of simple syllables (square-root-transformed), and total number of complex syllables (square-root-transformed). These USV parameters tended to discriminate pairs of different relatedness during direct interactions (DFA:

$n = 26$, Wilks' Lambda = 0.558, canonical correlation = 0.665, $p = 0.057$), but not during introduction (DFA: $n = 25$, Wilks' Lambda = 0.655, canonical correlation = 0.588, $p = 0.206$) (Fig. 4). Using cross-validation, the DFA was able to correctly classify 73% unrelated and 50% related pairs into the respective group in the introduction phase (overall: 64%, not cross-validated: 84%), and 66% unrelated and 63% related pairs in the interaction phase (overall 65.4%, not cross-validated: 80.8%). For each phase USV features could be combined into one discriminant function, which was plotted against the latency to the first litter (LFL) (Fig. 4) as a measure of reproductive success (see below). The parameters with the greatest discriminatory ability between related and unrelated pairs were number of short syllables, grand mean frequency and mean USV length in the introduction phase and number of simple syllables, mean USV length and number of short syllables in the interaction phase. Thus, in the introduction phase males emitted a larger number of simple syllables with a longer duration and higher frequency to unrelated females, whereas they emitted a larger number of short syllables at lower frequencies to related females (Fig. 4a). During direct interactions, unrelated mice emitted USVs with a longer duration and used a larger number of complex syllables, while related mice emitted a larger number of short and simple syllables (Fig. 4b).

We then compared the different syllable types emitted by related versus unrelated pairs using multivariate analyses. The results showed that the number of syllables used per syllable type tended to differ between related and unrelated mice during the introduction phase (PERMANOVA: $n = 26$, $F = 1.942$, $p = 0.062$), but not during direct interactions (PERMANOVA: $n = 26$, $F = 0.797$, $p = 0.481$). Variances were larger in unrelated than related pairs during the introduction phase (permutation based analysis of multivariate group dispersions: $n = 26$, $F = 4.314$, $p = 0.041$). Since PERMANOVA assumes similar multivariate dispersions, these results should be treated with caution and interpreted only for exploratory purposes. In detail, 80% of the difference between the unrelated and related pairs during the introduction phase was explained by five syllable types ("up", "uc", "s", "c2" and "us") (Fig. 5a). Three syllable types ("up", "c2" and "s") showed a greater abundance when males were presented with an unrelated partner and two syllable types ("uc" and "us") were emitted more often by related pairs (Fig. 5a, c). Visual inspection of the pie charts suggests that related mice emitted more ultrashort, short and unclassified syllables ("us", "s" and "uc") in both phases, whereas unrelated mice emitted more "up", "u", "ui", "c2", "c3" and "c4" syllables in both phases (Fig. 5c). These results are also consistent with the previous DFA showing that related mice emitted larger number of

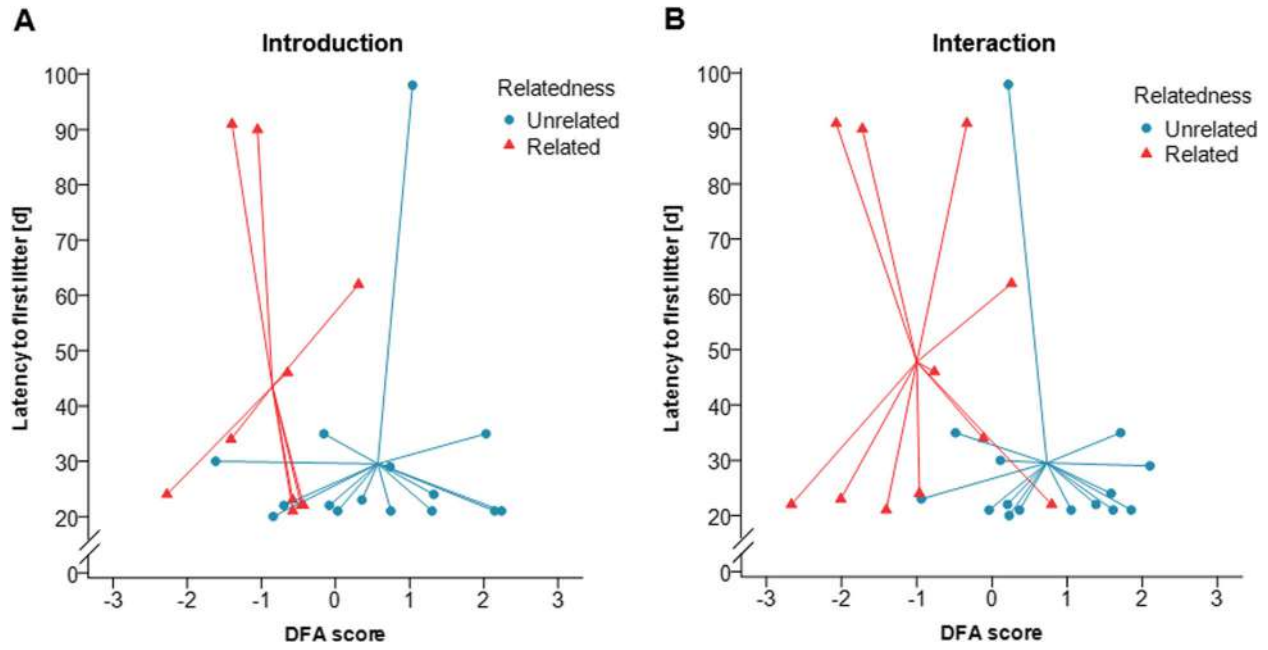

Fig. 4 Discriminant function scores of USV emission and latency to reproduce. Each symbol represents one genetically unrelated (blue circles) or related (red triangles) breeding pair and lines connect individual pairs to the corresponding group centroid. **a** Introduction phase: larger DFA scores represent a higher number of simple syllables, longer durations and higher frequencies; smaller scores indicate a higher number of short syllables and lower frequencies. **b** Direct interaction phase: larger DFA scores represent longer USVs and a higher number of complex syllables; smaller scores indicate a higher number of short syllables.

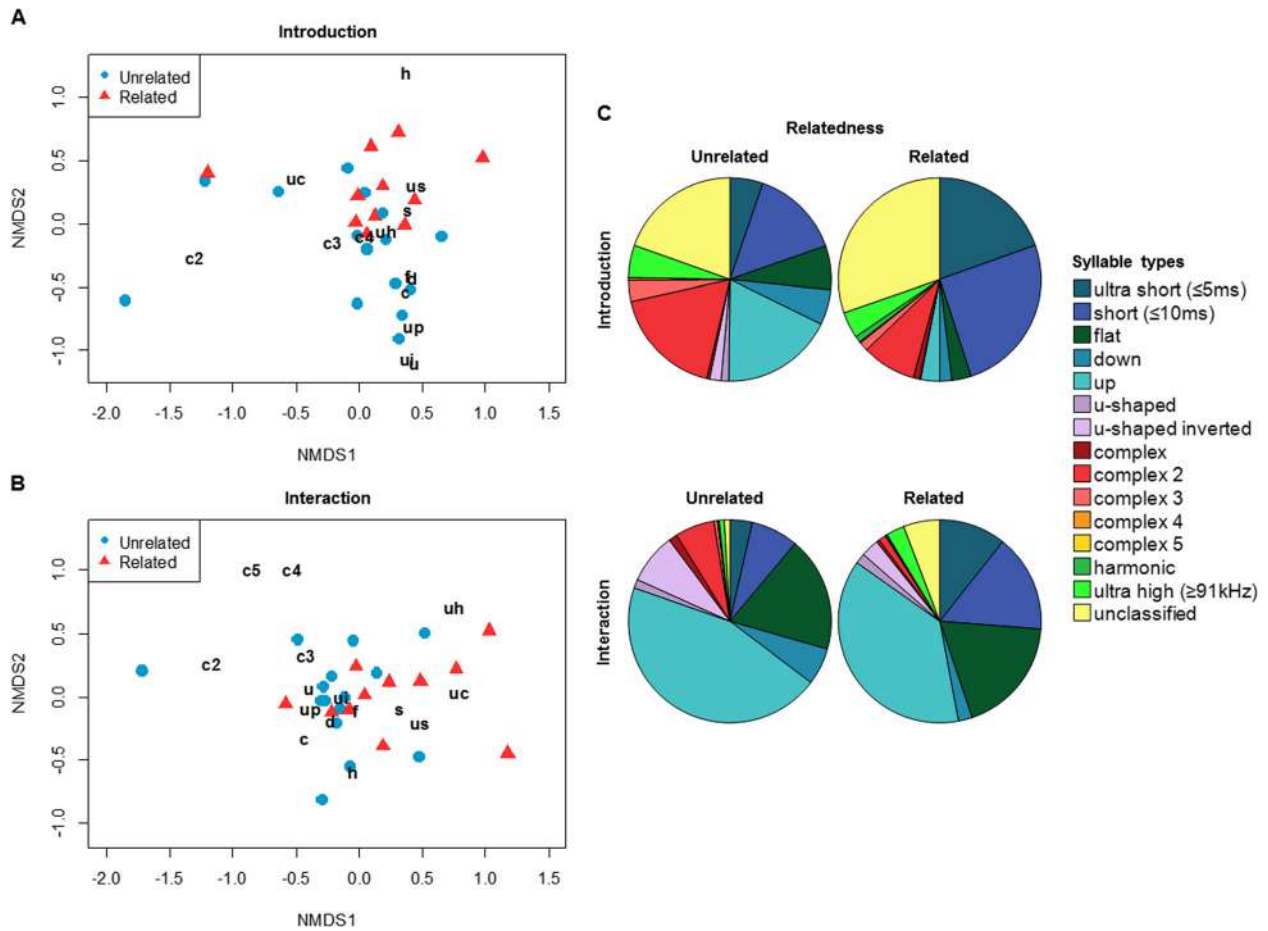

Fig. 5 Syllable type usage presented in non-metric multidimensional scaling (nMDS) plots and pie charts. NMDS plots of syllable types emitted during (**a**) introduction and (**b**) interaction phase comparing unrelated pairs (blue dots) versus related pairs (red triangles). Letters in black indicate the syllable types; each symbol represents one breeding pair. Distances between the symbols represent similarities of breeding pairs in the syllable type usage. Short distances of symbols to letters indicate syllable types that were most representative for each breeding pair. **c** Pie charts representing proportions of each syllable type used by unrelated and related pairs during introduction and interaction phase. Both visualizations show similar results of syllable type usage

ultrashort and short syllables, whereas unrelated mice emitted a larger number of simple syllables during the introduction phase and a larger number of complex syllables during interactions (Fig. 4). Visualization of syllable type usage in a 2-dimensional space using non-metric multidimensional scaling (nMDS) plots provides a good representation of the data during the introduction phase (stress = 0.109) (Fig. 5a) and an intermediate representation for the interaction phase (stress = 0.150) (Fig. 5b). Visual comparison of pairs consisting of siblings and cousins show a similar distribution in syllable type usage, however, we did not conduct a statistical comparison due to the low sample sizes within groups (see Additional file 1: Figure S1).

Reproductive success

We tested whether genetic relatedness influenced the reproductive success of the pairs, and found that unrelated pairs sired significantly more offspring than related pairs during the entire breeding period (t-test: $n = 26$, $t = 2.215$, $p = 0.036$) (Fig. 6a) When comparing the number of offspring born within 70d (i.e. the same time period for all breeding pairs), unrelated pairs still sired more offspring (13 ± 7) than related pairs (8 ± 7), however, this difference was not significant (t-test: $n = 26$, $t = 1.783$, $p = 0.087$) (Fig. 6b). Unrelated pairs gave birth to more litters (Mann-Whitney U test: $n = 26$, $Z = -2.381$, $p = 0.017$) (Fig. 6c, Table 1), while the litter size did not significantly differ between unrelated and related pairs (t-test: $n = 26$, $T = 1.344$, $p = 0.191$; Table 1). Furthermore, unrelated pairs tended to have a shorter latency to the first litter (Mann-Whitney U test: $n = 26$, $Z = -1.832$, $p = 0.067$) (Fig. 6d, Table 1) compared to related pairs, however this effect of relatedness depended on female receptivity. We found an interaction between the female's receptivity and her relatedness to the male on the latency to the first litter

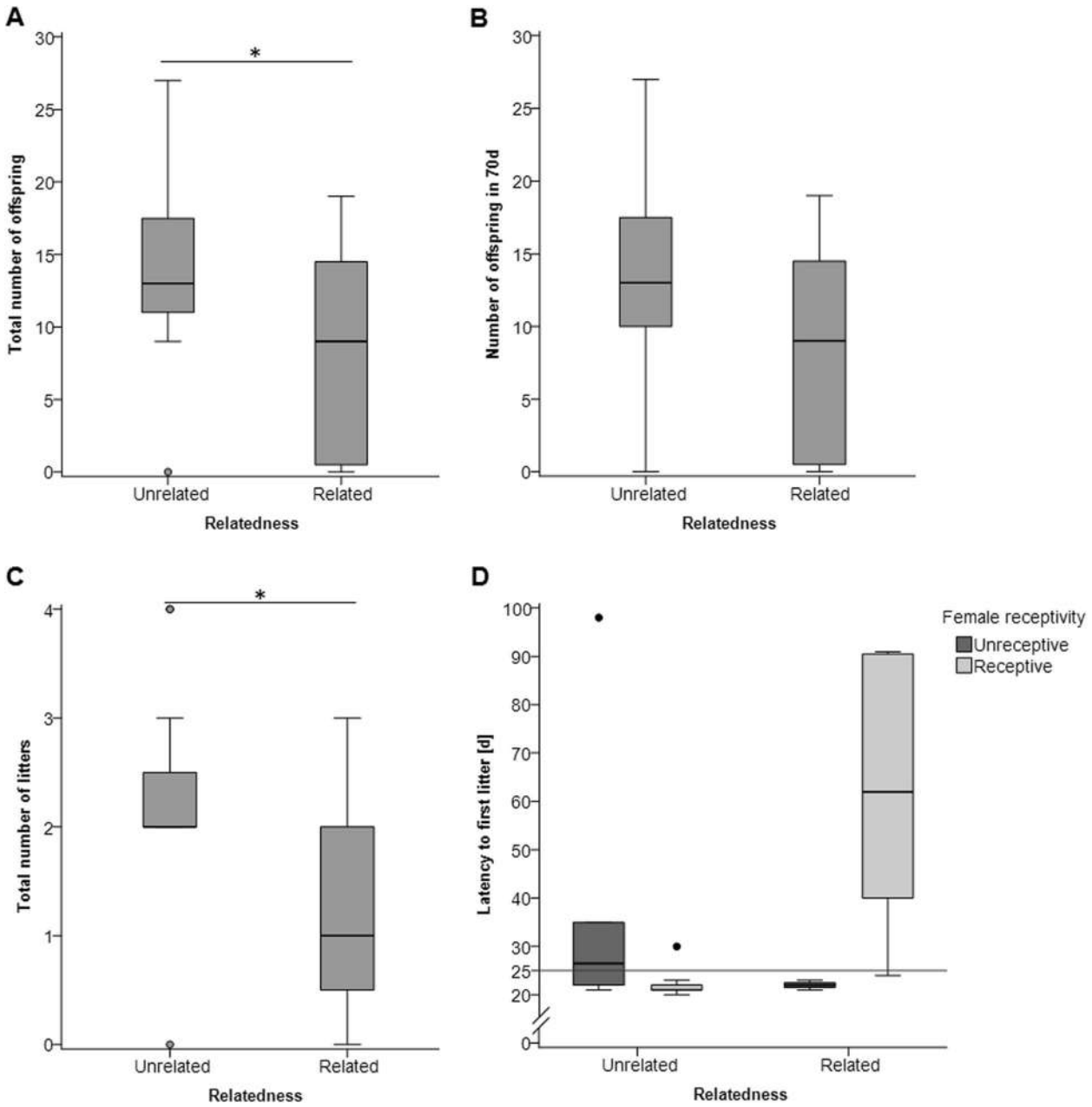

Fig. 6 Reproductive success of unrelated and related breeding pairs. Boxplots show median (center line), interquartile range (box), 95% variation (whiskers) and outliers (circles) of (**a**) the total number of offspring during the entire breeding period, (**b**) the number of offspring born within 70d, and (**c**) the total number of litters during the entire breeding period. **d** Latency (days) to the first litter (LFL) comparing males paired with unrelated or related females, which were either sexually unreceptive (dark grey, $n = 8$ unrelated, 4 related) or receptive (light grey, $n = 7$ unrelated, 7 related) on the recording day. There was a significant interaction between the female receptivity and relatedness to the male. The line at 25d represents the division into pairs with a low LFL (<25d) vs. high LFL (>25d) (see later). * $p < 0.05$

(GZLM: $n = 26$, effect of relatedness: Wald-$\chi^2 = 5.135$, $p = 0.023$, effect of receptivity: Wald-$\chi^2 = 5.530$, $p = 0.019$, interaction relatedness*receptivity: Wald-$\chi^2 = 24.391$, $p < 0.001$). Among pairs with females that were initially receptive, unrelated pairs had a significantly shorter LFL than related pairs. When females were initially unreceptive, there was no difference in LFL between related and unrelated pairs (Fig. 6d).

USV emission and reproductive success

We tested whether USV emission correlated with the pair's reproductive success. We found that several USV

Table 1 Statistical comparison of reproductive success (RS) between unrelated ($n = 15$) and related ($n = 11$) breeding pairs

		RS within entire breeding period		RS within 70 days	
Mann-Whitney U test					
Variable	**Comparison**	**Z**	***p*-value**	**Z**	***p*-value**
Latency to first litter	Unrelated vs. Related	−1.832	0.067	−1.913	0.056
Number of litters		−2.381	**0.017**	−2.096	**0.036**
t-test					
Variable	**Comparison**	**t**	***p*-value**	**t**	***p*-value**
Number of offspring	Unrelated vs. Related	2.215	**0.036**	1.783	0.087
N. offspring in first litter		1.356	0.188	1.356	0.188
N. offspring/litter		1.344	0.191	1.264	0.218
GZLM					
Variable	**Effect**	**Wald-Chi-Square**	***p*-value**	**Wald-Chi-Square**	***p*-value**
Log (Latency to first litter)	Relatedness	5.135	**0.023**	6.201	**0.013**
	Receptivity	5.530	**0.019**	6.744	**0.009**
	Relatedness*Receptivity	24.391	**< 0.001**	29.298	**< 0.001**

Results are shown for data including offspring delivered during the entire breeding period (85 ± 15d) and when using only offspring delivered within 70d. The second dataset represents equal breeding opportunities for all pairs. Results showing $p < 0.05$ are reported in bold

parameters correlated with the latency to the first litter. Surprisingly, we found that the results for unrelated (UR) and related (R) pairs depended upon the experimental phase. During the introduction phase, the grand mean frequency of USVs and the vocal repertoire emitted by males in related pairs was negatively correlated with the LFL (Spearman correlation: UR: $n = 15$, $r_s = 0.270$, $p = 0.331$, R: $n = 10$, $r_s = -0.632$, $p = 0.0498$, Fig. 7a and Spearman correlation: UR: $n = 15$, $r_s = 0.363$, $p = 0.184$, R: $n = 11$, $r_s = -0.632$, $p = 0.037$, Fig. 7b, respectively; see Additional file 1: Table S5). Thus, related mice emitting USVs at a higher grand mean frequency and with a larger vocal repertoire in the introduction phase had a shorter latency to the first litter. Unrelated pairs' USV emission during the introduction phase did not correlate with LFL. During direct interactions, however, we found that the mean length of USVs negatively correlated with LFL but only in unrelated pairs (Spearman correlation: UR: $n = 15$, $r_s = -0.523$, $p = 0.046$, R: $n = 11$, $r_s = 0.123$, $p = 0.718$) (Fig. 7c, Additional file 1: Table S5). Furthermore, unrelated pairs that had a higher vocal performance, tended to have a shorter LFL (Spearman correlation: UR: $n = 15$, $r_s = -0.502$, $p = 0.056$, R: $n = 11$, $r_s = 0.306$, $p = 0.360$, Additional file 1: Table S5). When analyzing short, simple and complex syllable types separately, we found a significant negative correlation between the number of simple syllables and LFL (Spearman correlation: UR: $n = 15$, $r_s = -0.526$, $p = 0.040$, R: $n = 11$, $r_s = 0.346$, $p = 0.298$) (Fig. 7d, Additional file 1: Table S5), and a trend in the correlation between the number of complex syllable types and LFL (Spearman correlation: UR: $n = 15$, $r_s = -0.472$, $p = 0.076$, R: $n = 11$, $r_s = 0.388$, $p = 0.238$, Additional file 1: Table S5). Thus, unrelated mice emitting longer USVs and with a higher number of simple syllables during direct interactions had a shorter latency to the first litter. When using the DFA scores, which combine the USV parameters for each phase, there was no correlation between the DFA score and the latency to the first litter (Spearman correlation: introduction: $n = 25$, $r_s = -0.249$, $p = 0.230$, interaction: $n = 26$, $r_s = -0.282$, $p = 0.163$) (Fig. 4).

USV emission and reproductive success were not affected by male age or age differences; however, we found a negative correlation between female age and the reproductive success in unrelated but not in related pairs. Unrelated pairs with older females had a higher latency to the first litter (Spearman correlation: UR: $n = 15$, $r_s = 0.826$, $p < 0.001$, R: $n = 11$, $r_s = 0.151$, $p = 0.658$, Additional file 1: Table S6) and produced less offspring within 70 d (Spearman correlation: UR: $n = 15$, $r_s = -0.586$, $p = 0.022$, R: $n = 11$, $r_s = -0.396$, $p = 0.228$, Additional file 1: Table S6). Furthermore, the age of females in unrelated pairs was correlated with the vocal performance (Spearman correlation: $n = 15$, $r_s = -0.573$, $p = 0.026$, Additional file 1: Table S7), vocal repertoire (Spearman correlation: $n = 15$, $r_s = -0.531$, $p = 0.042$, Additional file 1: Table S7) and grand mean frequency (Spearman correlation: $n = 15$, $r_s = -0.526$, $p = 0.044$, Additional file 1: Table S7) of USVs emitted during direct interactions. However, we did not find any correlation of the females' age and USV emission in related pairs or during the introduction phase (see Additional file 1: Table S7).

Next, to test whether syllable type usage was associated with LFL, we divided the breeding-pairs into pairs that gave birth within 25d (short LFL) and after 25d

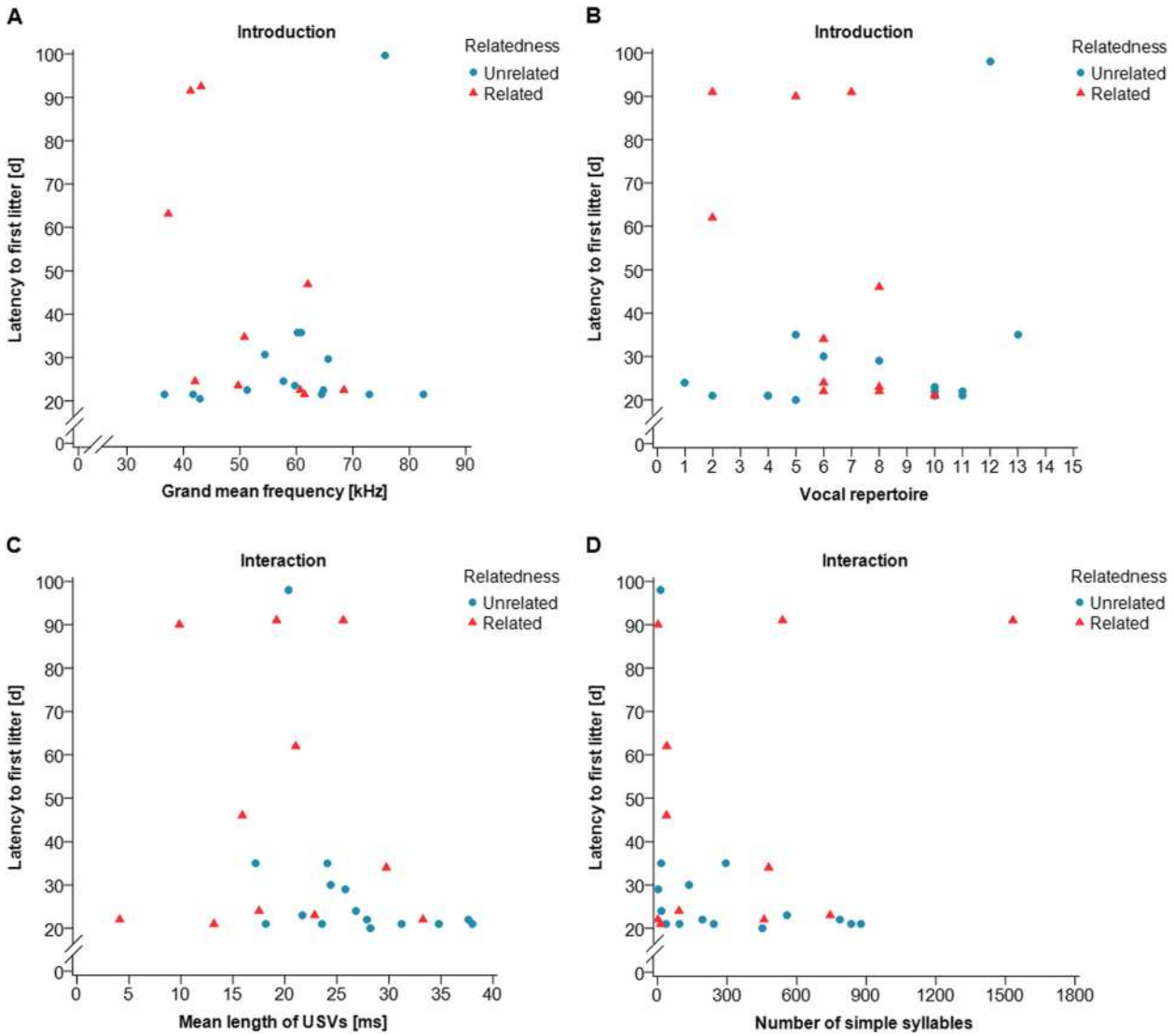

Fig. 7 Correlation between USV emission and the latency to the first litter (LFL). Each symbol represents one breeding pair consisting of either unrelated (blue circles) or related (red triangles) individuals. During the introduction phase (**a**) the grand mean frequency of USVs (kHz) and (**b**) the vocal repertoire (number of syllable types emitted during 10 min) were correlated with LFL in related pairs. During the interaction phase (**c**) the mean length of USVs (ms) and (**d**) the number of simple syllable types (total number of simple syllables emitted during 10 min) were correlated with LFL in unrelated pairs

(long LFL). We chose a cut-off at 25d for two reasons. First, visual inspection of the data showed a skewed distribution of LFL. Fifteen pairs had their first litter within 24d (20-24d), while 11 pairs had a latency of ≥29d (Fig. 8, see Additional file 1: Figure S2). Second, since the expected gestation period of mice is 21d and one estrous cycle lasts for approximately 4 days, mice with a latency to the first litter of <25d were expected to mate within the first estrous cycle. However, we found that syllable type usage in both phases did not differ between pairs with a short or long latency to the first litter (PERMANOVA: introduction: $n = 26$, $F = 0.203$, $p = 0.997$, interaction: $n = 26$, $F = 0.835$, $p = 0.481$) (Fig. 9).

Discussion

This is the first study to record wild-derived mice during direct sexual interactions to our knowledge, and we tested whether the USVs of mice emitted during courtship depend upon the genetic compatibility of a potential mating partner, and whether USV emission is correlated with the pair's subsequent reproductive success. Our main findings include the following results: (1) once males were allowed to directly interact with an

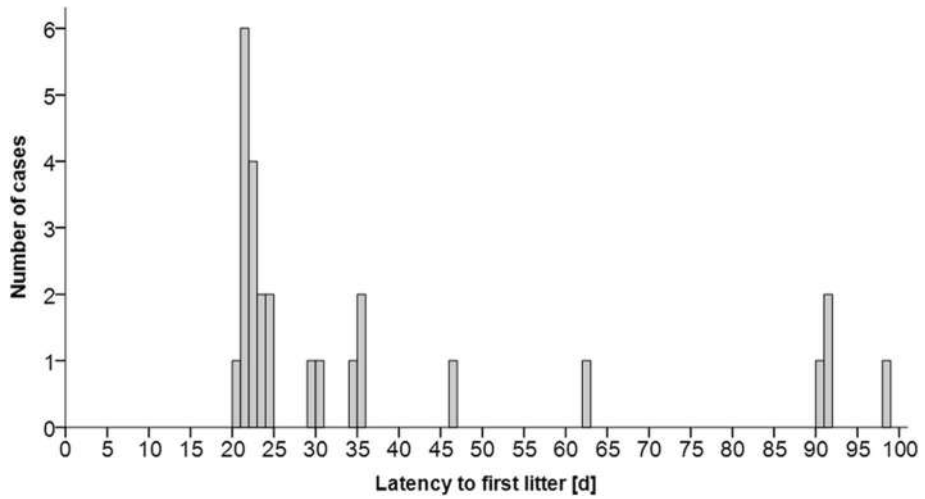

Fig. 8 Histogram showing the frequency of the latency to the first litter. 15 pairs had a latency of <25d and 11 pairs had a latency of > 28 days. The visible cut-off at 25 days was used to divide mice into pairs with a short vs. long latency to the first litter

unfamiliar female, we detected a significant increase in the number of USVs emitted, which shows that mice modulate their vocal performance during the early phases of courtship. (2) We detected longer and more complex USVs when males were experimentally paired with genetically unrelated compared to related females. This result provides further evidence that house mice show genetic kin discrimination, and the first evidence that male courtship USVs depend upon the relatedness of a potential mating partner. (3) We found that unrelated pairs of mice had higher reproductive success compared to incestuous pairings, which is consistent with inbreeding avoidance, though prenatal offspring mortality due to inbreeding depression cannot be ruled out. (4) We found that mean number and length of vocalizations of unrelated pairs were negatively correlated with the latency of the pairs' first litter. This is the first study to our knowledge to show that USV emission depends upon genetic compatibility of mating partners, and the first to find a relationship between USV emission and *subsequent* reproductive success. Future studies are needed to test whether USV emission influences mating, whether mating influences USV emission, or both. Below we address our main findings in more detail.

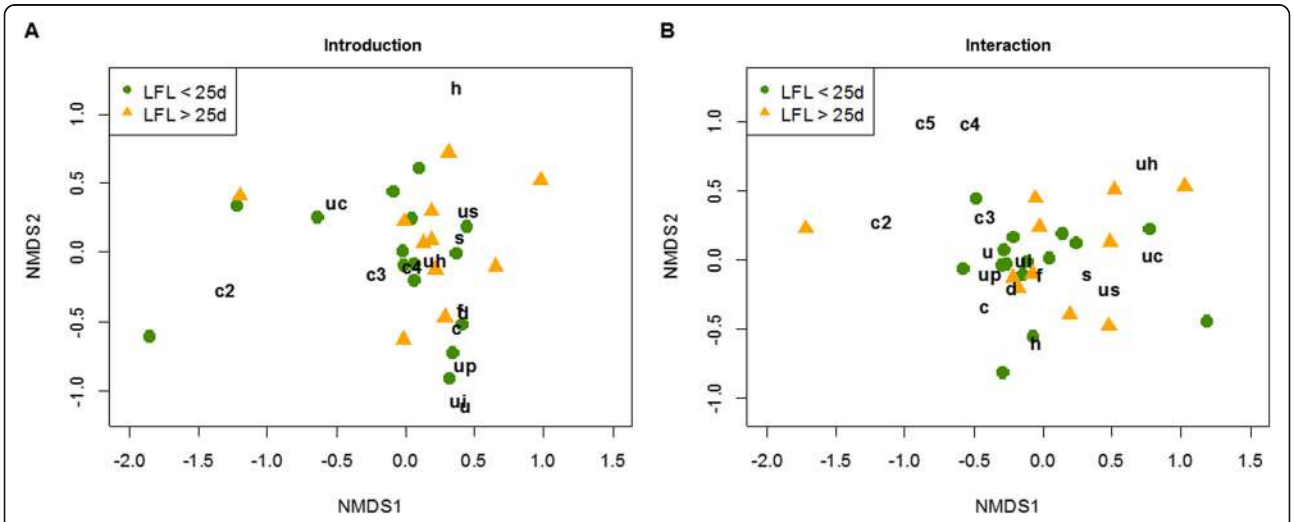

Fig. 9 Syllable type usage depending on the latency to the first litter (LFL). NMDS plots showing syllable type usage during (**a**) introduction and (**b**) interaction phase comparing pairs with a short latency to the first litter (LFL < 25d, green circles) and pairs with a long latency to the first litter (LFL > 25d, orange triangles). Letters in black indicate the syllable types; each symbol represents one breeding pair. Distances between the symbols represent similarities of breeding pairs in the number of emitted syllable types. Short distances of symbols to letters indicate syllable types which were most representative for each breeding pair

Dynamics of courtship USVs

Wild-derived house mice, unlike laboratory mice, rarely vocalize in laboratory conditions until they are presented with a stimulus mouse or its scent (Marconi et al. unpublished ms). Most males, but not all, begin vocalizing after presenting a mouse on the opposite side of a perforated partition, and especially if it is the opposite sex [15]. We expected males in this study to vocalize once they detected females through the partition and then to increase the amount and the types of USVs that they emit once they could contact and directly interact with the female. As expected, we found a large (5x) increase in the number of USVs emitted during the *interaction phase*, and the USVs were 1.5x longer compared to the preceding *introduction phase* (Fig. 1). The mice also produced a more diverse vocal repertoire, as they emitted more types of syllables during the interaction than the introduction phase. The high rates of USV emission during the interaction phase in our study were ca. 2x greater than in previous studies of wild-derived mice recorded while presenting males with only a female olfactory stimulus (female urine) [17, 20, 22, 23] or with a stimulus female separated with a divider [15, 24]. Our results indicate that once males are presented with a stimulus female, they modulate the rate of USV emission from low to higher rates of calling upon detecting and then directly interacting with and pursuing a potential mating partner. A previous study on laboratory mice similarly found that males emitted few vocalizations while the sexes were separated, and then produced high levels of USV emission during direct interactions [25]. The amount and types of USVs that mice emit during opposite-sex interactions are associated with mounting behavior [26] and the types of USVs emitted change over time during courtship and become longer and more complex (i.e., multiple frequency jumps and harmonic elements) at the end of courtship, and just before copulation [14]. Taken together, these results indicate that when male mice encounter an unfamiliar adult female, they begin vocalizing, and then continue to modulate the amount and types of USVs that they emit over time as they initiate courtship and attempt mating. Modulating USV emission over time during courtship could potentially influence male mating success, and how a male modulates his vocalizations might be more important than the total number of calls that he produces (e.g., producing too many or the wrong types of USVs too soon or too late might repel females). We observed chasing, nose-to-nose sniffing, and anogenital sniffing, but we did not observe any mating or mounting attempts in our study. Future studies are necessary with longer observation times (especially for wild mice) to document how USV emission changes over time from first encounter to copulation – and to determine how and why males modulate USV emission. One study found that females are more attracted to playbacks of complex USVs (containing more frequency jumps) than simple ones (without frequency jumps) [19]. Playback studies are now needed to manipulate the amount, types and order of USVs that females perceive during opposite-sex interactions over the stages of courtship, and to examine female responses to differences in the rate and other features of male USVs. Determining the function of dynamic modulation of male courtship USVs will be a challenge, especially since courtship vocalizations appear to be an interactive exchange between the sexes (duetting) [27–29].

The increased USV emission we found during direct interactions might be explained, at least in part, by vocalizations emitted from females during direct interactions. It was previously concluded that female mice do not vocalize during courtship, as the rate of USV (70 kHz) emission did not differ when males interacted with a surgically muted versus an intact, control female [30]. For this reason, studies on USV emission recorded during sexual interactions often assume that only males vocalize [14, 26]. More recently, however, female mice have been shown to vocalize during direct opposite-sex interactions [27–29]. However, females in these studies contributed only up to 18% of the total USVs emitted, and this is insufficient to explain the 5x increase during direct interactions that we detected in our study. Wild-derived female house mice have been shown to emit USVs when presented with a stimulus mouse separated by a divider [15, 24]. Here, we did not detect any instances of overlapping USVs, which can be an indicator that females rarely vocalized, unless they alternate their calls to avoid overlapping. While we can attribute most calls to the males during the introduction phase due to the covered compartment of the female, we cannot assume that the USVs recorded during direct interactions were emitted solely by males in our study. Nevertheless, it is unlikely that female vocalizations explain the large increase of vocalizations that we found during the direct interaction phase, and more importantly, this uncertainty does not change the interpretation of our main findings. Thus, future studies are needed to record the USV emission of both sexes during direct opposite-sex interactions in wild house mice.

Female sexual receptivity

We expected males to modulate their USV emission depending on whether they are presented with a sexually receptive versus a non-receptive female. During the introduction phase, we found that males emitted USVs at significantly lower frequencies (43 vs. 61 kHz), and surprisingly they had a significantly *lower* vocal repertoire (6 vs. 9 syllable types) when presented with a

sexually receptive female (proestrus or estrus) compared to an unreceptive (metestrus or diestrus) female (Fig. 2). However, during direct interactions we found no significant effect of female sexual receptivity on USV emission. A previous study that recorded the USVs of opposite-sex pairs of laboratory mice found that their USVs were also lower in frequency when females were in proestrus compared to diestrus [26]. The recordings were made during direct interactions, and it was assumed that this difference was due to males changing the frequency of their vocalizations, but female vocalizations were not controlled. In our study the female compartment was covered, so that the differences in USV emission we found during the introduction phase were only due to male vocalizations. Unlike this previous study, we found no evidence that female receptivity influenced the length of USVs emitted during direct interactions, and surprisingly, we found that female estrus had a negative effect on the vocal repertoire. Thus, the effects of female estrous status on male USV emission may depend on whether the mice are directly interacting and on the stage of courtship.

Genetic relatedness and USV emission

We experimentally paired males with either a genetically related or an unrelated female, and we expected the mice to modulate their USV emission depending on the kinship of their potential mating partner. The males were unfamiliar with the stimulus females, as their sisters were from different litters. We found that mice emitted more vocalizations (introduction phase) and longer USVs (interaction phase) when presented with unrelated compared to related females. While we found that more simple syllables were emitted for an unrelated compared to a related female during the introduction phase, our multivariate (discriminant function) analysis indicated that during the interaction phase unrelated pairs emitted a higher number of complex USVs, whereas related pairs emitted a higher number of short USVs. Thus, our result shows that male house mice discriminated between genetically related versus unrelated stimulus females, and that they emitted more, longer and a higher number of complex USVs for unrelated compared to related females, indicating that males can assess the relatedness of potential mating partners, even if unfamiliar (genetic kin recognition) (reviewed in [31]). The recognition mechanisms involved here are still unclear, but mice can discriminate kinship through odor cues [32]. If USV emission provides an index of a male's sexual arousal, as often proposed [13], then our findings suggest that male mice show mating preferences for non-kin over kin potentially to avoid inbreeding. Females have been shown to recognize siblings by their USVs, as they are more attracted to playbacks of USVs from non-siblings than siblings [17], however, we cannot rule out the possibility that females might be more attracted to unrelated males due to odor or their enhanced sexual arousal. Taken together, our findings provide further evidence for genetic kin recognition in house mice, and though USVs might mediate inbreeding avoidance [17], more playback experiments are needed to study female preferences and the recognition mechanisms. We did not compare the different degrees of relatedness (siblings vs. first cousins) due to inadequate sample sizes, and future studies are needed to investigate this question.

Genetic relatedness and reproductive success

As expected, unrelated pairs had higher reproductive success compared to mice that we experimentally assigned to mate with close kin. Unrelated pairs produced more litters than related pairs, though their litter sizes did not differ. This result may have been due to inbreeding avoidance, though prenatal offspring mortality due to inbreeding is a non-mutually exclusive explanation for such differential reproductive success. This result did not differ when using the total number of offspring sired during the entire breeding period or only offspring born within 70d. We also found a lower latency to first litter (LFL) among unrelated pairs, though this result was also influenced by the receptivity of the female at pairing. The difference in LFL between unrelated versus related pairs was only observed when the female was initially sexually receptive. In wild-derived inbred strains of mice, it has been shown that females derived from *Mus musculus musculus* (PWD/PhJ) show a strong assortative choice when they are in estrous but not in diestrous when they could choose between *M. m. musculus* (PWK/PhJ) or *M. m. domesticus* (C57BL/6 J) males [33]. Thus, our results could be explained by females that were sexually receptive at pairing showing a greater attraction toward unrelated than related males, whereas the subsequent timing of reproduction must be explained by other factors. Our results show that unrelated pairs reproduced with a shorter latency (especially when the female was receptive) and at a faster rate (more litters) compared to related pairs.

USV emission and reproductive success

We expected that the USV emission would predict a pair's subsequent reproductive success (i.e., reduced latency to the first litter (LFL), increased offspring number, or both). The results supported our prediction, but surprisingly, we found different results depending upon the genetic relatedness of the pairs and the recording phase. We found significant results only during the introduction phase for the related mice and only during the interaction phase for unrelated mice. Among the

genetically related pairs, we found a significant correlation between male USV emission and LFL: the males that emitted USVs at higher frequencies and with a larger vocal repertoire had a shorter latency to the first litter. This result was found during the introduction phase (which is why we can attribute the effect to the male vocalizations), but not the interaction phase. Our previous study found that mice emit USVs at higher frequencies during opposite- compared to same-sex interactions [15]. The potential functions of USVs emitted at different frequencies is not known, however, one possible explanation might be that USV frequency is related to sexual contact. Thus, if higher frequencies indicate sexual arousal, then this could explain the association with faster reproduction in the present study, however studies are needed to test for this effect. Similarly, the negative correlation between the vocal repertoire and the LFL might indicate that emitting a larger number of different syllable types can signal a higher sexual arousal of males, or might be perceived as more attractive by the female partner. Among unrelated pairs, USV emission was also correlated with LFL, but only during the interaction phase. We found a shorter latency to the first litter when unrelated males emitted longer USVs and a higher number of simple USVs during the direct interactions. We also found a negative correlation between complex USVs and LFL, though this trend was not significant. This latter result is consistent with a study in laboratory mice that found that the number of long USVs with multiple frequency jumps and harmonic USVs increase over time during courtship and mounting [14]. Additionally, this previous study found that the distribution of the duration of syllables emitted during the early phase of an interaction was different in pairs that only showed sniffing behavior compared to pairs that also showed mounting behavior. Mice exhibiting both sniffing and mounting seemed to emit longer USVs than mice that did not show mounting behavior [14]. Thus, emission of long USVs and more complex USVs might be an indicator of a higher male's sexual arousal, and might facilitate mating. Complex and long USVs might be used by males to signal their sexual motivation [14], and simultaneously provide information about their genetic relatedness, which then might increase female receptivity [18]. Furthermore, we found that USV emission depended upon female age (among unrelated pairs), suggesting that USV emission might signal male sexual motivation. As fertility decreases with age [34], males are expected to prefer younger than older females, and here we found that unrelated pairs reproduced faster and with a larger number of offspring when the female was younger.

Finally, since our results are based on correlational evidence we cannot conclude any causal links and further studies are needed to experimentally test the effect of USV emission on mating behavior and reproductive success. Our results could be due to male preferences, female preferences, or both. Male USV emission appears to signal sexual arousal [13], and males might be more attracted to unrelated than related females. Females might be more attracted to these calls and mate faster with males that are more sexually aroused. Alternatively, females might discriminate individual males or kin versus nonkin using male USV emission [35], and might mate faster with unrelated than related males [17]. Additionally, dynamic interactions between males and females can influence the partner's behavior, and therefore female and male preferences might not be independent from each other.

Conclusions

In summary, our study provides evidence for dynamic modulation of courtship USVs, genetic kin recognition, and that the courtship USVs of male mice predicts their subsequent reproductive success. Our results can be useful for future breeding regimes, as USV emission could be used to screen breeding pairs during their first contact to anticipate their subsequent latency to reproduce and reproductive success. Since wild mice often show a long latency to reproduce or do not reproduce at all, this could save time and resources in the laboratory, especially when working with wild mice. Future studies are needed to manipulate the USV emission to experimentally test the effect of USV emission on mating and reproductive success. Furthermore, it is possible that USVs might even have a larger effect on male reproductive success in more natural conditions. Mice move around during courtship over a much larger area than small cages, and if male USVs help coordinate mating by keeping females nearby [36], then studies are also needed in larger areas.

Methods

Subjects and housing

We used wild-derived (F3) house mice (*Mus musculus musculus*). Wild mice were trapped at the Konrad Lorenz Institute of Ethology, Vienna, Austria (48°12′38″ N, 16°16′54″E) in 2012 and maintained as breeding stock [for more details see 15]. We used wild-derived mice to control for age and rearing conditions. Mice were weaned at 21d and kept in mixed-sex groups with ≤4 siblings per cage until the age of 5 weeks (35d). After this time, adult males were housed individually to prevent fighting and females were housed in sister-pairs whenever possible. Mice were housed in standard Type IIL cages (36.5 × 20 × 14 cm cages, Tecniplast, Germany), with food (rodent diet 1324, Altromin, Germany) and water provided *ad libitum*. Cages were covered with

stainless-steel covers (1 cm mesh width) and provided with bedding (ABEDD, Austria) and nesting material (Nestlet, Ehret, Austria). A nest box (Tecniplast, Germany) and a cardboard paper roll were provided for environmental enrichment. Home cages were kept at standard conditions (mean ± SD room temperature: 22 ± 2 °C) under a 12:12 h light-red light cycle (red lights on at 15:00). We used 26 males and 26 females, which were 249 ± 36 d old (mean ± SD) and sexually naïve at the beginning of the experiment.

Breeding pairs and their reproductive success (RS)

Using our colony pedigree, we assigned individual males and females to two types of experimental breeding pairs: (1) 15 unrelated pairs (UR) and (2) 11 related pairs (R) with an average coefficient of relatedness (CoR) of 0.29 ± 0.2 (mean ± SD). This group included 5 pairs of siblings from different litters (CoR = 0.5) and 6 pairs of cousins that shared either two grandparents (1st degree cousins, CoR = 0.125, $n = 3$), four grandparents (double 1st degree cousins, CoR = 0.25, $n = 1$) or two great-grandparents (2nd degree cousins CoR = 0.03125, $n = 2$). Differences in sample sizes and degree of relatedness were due to constrains on the number of individuals in our colony. The age difference between males and females of the breeding pairs was 30 ± 28d (mean ± SD; median = 21d), and was not significantly different between unrelated (32 ± 30d, median = 21d) and related (28 ± 27d, median = 21d) pairs (Mann-Whitney U test: $n = 26$, $Z = -0.286$, $p = 0.775$). Furthermore there was no difference of male or female age between unrelated and related pairs (Mann-Whitney U test: male age: $n = 26$, $Z = -1.272$, $p = 0.203$; female age: $n = 26$, $Z = -1.326$, $p = 0.185$). Breeding pairs were housed in the males' home cage after conducting the USV recordings and under the same housing conditions described above. After 21d, pairs were checked daily for litters, and each pair's reproductive success was documented using birth dates, litter sizes, number of litters and days that mice were kept in pairs. For further analyses of the reproductive success we used the following parameters: latency (in days) to the first litter (LFL), number of offspring in the first litter, total number of litters, total number of offspring and number of offspring/litter. The LFL of mice that failed to reproduce ($n = 4$ pairs) was defined as the number of days mice were kept in breeding pairs plus 21d (gestation time). Since mice were also bred for maintenance of our general colony, the time mice were kept in breeding pairs differed between pairs (mean ± SD: 9 ± 2 wks, range: 7–13 wks). Nevertheless, the number of days that mice were kept as breeding pairs was not significantly different between unrelated and related pairs (Mann-Whitney U test: $n = 26$, $Z = -0.523$ $p = 0.610$). The minimum time pairs were kept together was 48d. To compare and standardize the breeding opportunities for all pairs, we further analyzed the parameters using only offspring born until day 70 (48d in breeding pairs +22d gestation period). Accordingly, we adjusted the number of offspring, number of litters and number of offspring/litter for pairs that reproduced also after 70d ($n = 3$ unrelated pairs). Since all pairs that reproduced delivered their first litter within 70d, the LFL only needed to be reduced to 70d for pairs that did not reproduce ($n = 4$). The analyses using the restricted dataset, did not change our main results (see Table 1), and therefore, we only present results of the full dataset, unless stated otherwise. After the breeding was terminated, all parental mice (males and females) were housed individually under standard housing conditions.

Female estrous state

Estrous state was checked using vaginal smears 3 to 5 h prior to USV recordings and staged according to the presence or absence of vaginal cell types (light microscope with 200x magnification using a 20x objective and 10x ocular): diestrus (mainly leukocytes), proestrus (mainly nucleated epithelial cells), estrus (mainly cornified cells) and metestrus (equal combination of all three cell types) [37]. For further analyses, we pooled females in proestrus and estrus as "sexually receptive" (indicated by the absence of leukocytes) and females in metestrus and diestrus as "sexually unreceptive" (indicated by the presence of leukocytes) [33]. We used this classification since we were interested in assessing sexual receptivity rather than a particular estrous state.

Recording apparatus and procedure

USV recordings were conducted during the mice active period under red light, i.e. after the onset of the dark phase (15:00–18:00 h) in a separate, closed room. Mice were recorded in a plexiglass cage (modified from a Type III cage, Tecniplast, Germany; floor measurements: 36.5 × 21 × 15 cm, top measurements: 42.5 × 27 × 15 cm) equally divided into two compartments by a perforated plexiglass divider (described in [15]). A clean recording cage was used for each breeding pair. Both compartments were provided with equal amount of soiled bedding from the male's home cage. Before each recording, a male mouse was gently transferred into one of the two compartments, which was covered with a standard cage cover (1 cm mesh width). A female was then transferred into the other compartment of the cage, which allowed both olfactory and visual cues through the perforated divider, but restricted physical contact. Recordings were conducted in two consecutive phases, lasting 10 min each. In phase 1 (*introduction phase*), we aimed to record only male vocalizations, while exposed to the female on the other side of the divider, i.e. with visual, and

chemical communication. To ensure that we only recorded male USVs, the female's compartment was additionally covered with a 0.5 cm plexiglass cover, which prevented recording USVs from the female compartment (see [15]). An ultrasound microphone (USG Electret Ultrasound Microphone, Avisoft Bioacoustics / Knowles FG) was placed in a fixed position 10 cm above the center of the male compartment. For phase 2 (*interaction phase*), we removed the divider at the end of phase 1 to allow direct, physical interactions. We also exchanged the plexiglass cover with a stainless-steel cage cover, and placed the microphone 10 cm above the middle of the entire cage to ensure that USVs would be recorded from all positions in the cage. The microphone was connected to an A/D-converter (UltraSoundGate 416Hb, Avisoft Bioacoustics). Recordings were conducted on a computer (Lenovo T540p, Windows 7) using the RECORDER USGH software (Avisoft-RECORDER Version 4.2) with a sampling rate of 300 kHz and 16 bit format. During USV recordings, we videotaped the mice using an IP-camera (D-Link DCS-3710) and open source software (iSpy - Video Surveillance Software), which allowed us to observe the behavior of the mice from another room. We did not observe any mating or mating attempts (i.e. mounting, intromission, copulation) during the 10 min recordings of direct interactions. Wild-derived mice typically show a long latency to mate (compared to laboratory strains, which are selected for fast reproduction), and we never observed any mating events during such brief interactions. After the end of phase 2, both mice were gently removed from the recording cage using plastic cylinders and the male bedding was returned to the male's home cage. Both mice were placed together into the male's home cage to allow breeding.

Processing and analyzing vocalizations

Sound files were processed semi-automatically in STx (S_TOOLS-ST^{x} Version 4.3.8 (9374), Acoustics Research Institute, Vienna, Austria). USVs were automatically detected using the Automatic Mouse Ultrasound Detector (A-MUD, version 3.1 [38]) and we set the threshold for element duration at 5 ms (rather than 10 ms) to increase the sensitivity in detecting ultrashort and faint elements. This threshold reduces false negatives, but increases the risk of false positive detections. We visually inspected all sound files and removed false positive and retained false negative segments. We also adjusted the length (start and end time) of the detected segments when necessary. This semi-automatic method ensured that we would include all USVs and exclude false positive segments from our analysis. The USVs were manually classified into one of 15 categories (adapted from [20, 23, 26, 39, 40]) depending upon their frequency, length and frequency modulation (Table 2). Ambiguous syllables or other sounds were verified by listening to the sounds (slowed down 15- to 20-fold). Additionally, syllables types were grouped into 3 different classes ("short syllables", "simple syllables" and "complex syllables", see Table 2), to reduce the number of variables in some analysis. Spectrograms for visual inspection were created using the transcription function in STx, which enabled us to scroll through the spectrogram in 2 s steps. Spectrograms were generated with a range of 50 dB (floor at -80 dB to obtain a comparable representation for all recordings), a frame of 4 ms and an overlap of 75%. The spectrograms were displayed in a Hanning window showing frequencies between 0 and 150 kHz. After classification, we ran the function *compute/update segment info* in A-MUD to compute spectrographic parameters of each detected element, including time and frequency parameters (start time, duration, mean frequency, minimum frequency and maximum frequency of each element). For one related pair, during the introduction phase only 2 USVs were emitted, which had low amplitude and the program was unable to detect frequency parameters. Thus, this pair was not included in statistical analyses when using frequency parameters of the introduction phase. All parameters were exported into an Excel-file (Microsoft) using the export-function of A-MUD, and processed for further analysis.

Statistical analyses

To quantify the total USV emission rate, we used the total number of USVs (*vocal performance*) recorded in each 10 min phase. To quantify the usage of different syllable types, we used the total number of USVs of each syllable type emitted per 10 min phase. The amount of different syllable types used by each pair is defined as their *vocal repertoire* (0–15 different syllable types). For quantifying additional spectrographic parameters (time and frequency parameters), we calculated means (e.g., mean length) or grand means (e.g., grand mean frequency) for each pair, separately for each parameter and for each 10 min recording. The mean length was calculated from the length of each USV averaged over all USVs in each recording. To calculate the grand mean frequency we used the mean frequency of each USV (i.e. the average frequency of the frequency track (contour) measured by AMUD) and calculated the average over all USVs in each recording [15]. We examined data distributions and homogeneity of variances using the Kolmogorov–Smirnov test and Levene's test, respectively, and we used non-parametric statistical tests if the assumptions for parametric statistics were not met. We tested for normal distribution separately for each phase and the different groups depending on the question (introduction vs interaction, receptive vs unreceptive, related vs. unrelated). If possible, we transformed the data to reach normal distribution. USV count data were square-root transformed, after adding 0.5 to the data (sqrt(x + 0.5)),

Table 2 Classification of the 15 different syllables types and grouping into 3 different syllable classes used in this study. Ambiguous syllables or other sounds were verified by acoustical inspection

Syllable shape	Syllable label	Syllable type	Syllable class	Definition	References
(< 5 ms)	**us**	ultrashort	Short syllables	Syllables < 91 kHz that are < 5 ms regardless of the shape	[39]
(< 10 ms)	**s**	short		Syllables < 91 kHz that are < 10 ms regardless of the shape	[26]
	f	flat		Syllables < 91 kHz with < 5 kHz frequency modulation	[26]
	d	down		Syllables < 91 kHz that decreases in frequency for > 5 kHz	[26]
	up	up	Simple syllables	Syllables < 91 kHz that increase in frequency for > 5 kHz	[26]
	u	u-shaped		Syllables < 91 kHz that first decrease, and then increase in frequency for > 5 kHz each	[26]
	ui	u-shaped inverted		Syllables < 91 kHz that first increase, then decrease in frequency for > 5 kHz each	[26]
	c	complex		Syllables < 91 kHz that contain ≥2 directional changes in frequency and > 5 kHz modulation of frequency	[26]
	c2	complex 2		Syllables < 91 kHz consisting of 2 elements separated by 1 frequency-jump without time separation	[20]
	c3	complex 3	Complex syllables	Syllables < 91 kHz consisting of 3 elements separated by 2 frequency-jumps without time separation	[20]
	c4	complex 4		Syllables < 91 kHz consisting of 4 elements separated by 3 frequency-jumps without time separation	Added category for our classification
	c5	complex 5		Syllables < 91 kHz consisting of ≥5 elements separated by ≥4 frequency jumps without time separation	Added category for our classification
	h	harmonic		Syllables < 91 kHz that have an harmonic element	[26]
(> 91 kHz)	**uh**	ultra high		All syllables > 91 kHz regardless of the shape	[23]
	uc	unclassified		Syllables that do not fit any other of the 14 categories due to background noise or that lack clearly defined spectrographic features (shape)	[40, 41]

and LFL was log-transformed (log(x)). Sqrt transformation of the USV count data resulted in normal distribution when comparing related vs unrelated pairs. Log-transformation of LFL was used to test for interactions between relatedness and receptivity. Detailed variable definitions and raw data which were used for statistical analyses are provided as additional file (Additional file 2). We used two-tailed tests, and results were considered statistically significant at $\alpha \leq 0.05$ and presented as mean ± SD, unless stated otherwise. Statistical tests were conducted using SPSS (IBM SPSS Statistics 24) and RStudio (R-Version 3.5.1 [42], using the functions "vegdist", "anosim", "adonis2" and "metaMDS" included in the package "vegan" [43]).

Whenever we tested for interactions between the female's sexual receptivity and relatedness to the male, we conducted a generalized linear model (GZLM) including the pairs' relatedness, the females' receptivity and the interaction of both as fixed factors. Additionally, we performed different multivariate methods to investigate whether different USV parameters and the syllable type usage depend upon the relatedness of the pairs. We conducted a discriminant function analysis (DFA) to test whether mice could be classified into unrelated or related pairs and which USV features had the main effect in discriminating between the two groups. We conducted DFA separately for the introduction and interaction phase, and included the following features: mean USV length, grand mean USV frequency, vocal repertoire, number of short syllables, number of simple syllables and number of complex syllables. These parameters were included in the DFA in order to combine spectrotemporal features with parameters of syllable diversity and complexity. Since we only compared two groups (unrelated vs. related pairs), the DFA resulted in only one discriminant function axis. For visual representation of the results, we plotted the DFA score against the latency to first litter as a variable describing the reproductive success, as this was one of our main study

questions. To describe differences in syllable type usage between unrelated and related pairs, we conducted PERMANOVA (permutational multivariate analysis of variance) on the number of USVs emitted within each of the 15 syllable types. PERMANOVA is a non-parametric alternative to other multivariate statistics (such as MANOVA), which works on permutations of a dissimilarity measure [44, 45]. We conducted the analysis with 999 permutations by running the function "adonis2" in the R-Package "Vegan" [43], using the rank based Bray–Curtis dissimilarity indices. Non-metric multidimensional scaling (nMDS) plots were created for visual representation of the results. The stress value of the plots describes whether the 2-dimensional nMDS plot sufficiently summarizes the relationship of the multidimensional data [46]. Stress values < 0.05, < 0.1 and < 0.2 will give an excellent, good or intermediate representation of the data, respectively. Results with stress values of 0.2–0.3 should be interpreted carefully, while stress values > 0.3 indicate arbitrary representation of the data in the 2-dimensional space [46]. To examine the relationship between USV emission and reproductive success, we conducted Spearman rank correlations separately for each phase and for unrelated and related pairs.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12983-020-00353-1.

Additional file 1. Supplementary material, containing additional figures (Figures S1 and S2) and summary tables of statistical results (Tables S1 – S7).

Additional file 2. Original dataset containing the raw data and variable definitions used in this article.

Acknowledgments
This research was supported by a grant (FWF P 28141-B25) of the Austrian Science Fund (http://www.fwf.ac.at). We thank U. Madlsperger and H. Sasse for animal care; T. Klaus and B. Wernisch for their help during data collection and processing; three anonymous reviewers for valuable comments that improved the manuscript; and R. Hengsberger for her help with formatting figures and the manuscript.

Authors' contributions
DN, MAM, SMZ and DJP designed the experiment. DN and MAM conducted the experiments and analyzed the data. DN, DJP and SMZ wrote the paper, and all authors reviewed and approved the final manuscript.

Funding
This research was supported by a grant (FWF P 28141-B25) of the Austrian Science Fund (http://www.fwf.ac.at). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files (Additional file 2).

Ethics approval
This study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. Experiments were conducted at the Konrad Lorenz Institute of Ethology, Austria and the protocols have been approved and were in accordance with ethical standards and guidelines in the care and use of experimental animals of the Ethical and Animal Welfare Commission of the University of Veterinary Medicine, Vienna (Austria) (ETK-17/04/2015).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 11 October 2019 Accepted: 4 March 2020
Published online: 03 April 2020

References
1. Searcy WA, Andersson M. Sexual selection and the evolution of song. Annu Rev Ecol Syst. 1986;17:507–33.
2. Kroodsma DE, Byers BE. The function(s) of bird song. Am Zool. 1991;31:318–28.
3. Brockway BF. Stimulation of ovarian development and egg laying by male courtship vocalization in budgerigars (*Melopsittacus undulatus*). Anim Behav. 1965;13:575–8.
4. Dunning JL, Pant S, Bass A, Coburn Z, Prather JF. Mate choice in adult female Bengalese finches: females express consistent preferences for individual males and prefer female-directed song performances. PLoS One. 2014;9:e89438.
5. Clutton-Brock T, McAuliffe K. Female mate choice in mammals. Q Rev Biol. 2009;84:3–27.
6. Holy TE, Guo Z. Ultrasonic songs of male mice. PLoS Biol. 2005;3:e386.
7. Musolf K, Penn DJ. Ultrasonic vocalizations in house mice: a cryptic mode of acoustic communication. In Evolution of the House Mouse. Volume 3. Edited by Macholán M, Baird SJE, Munclinger P, Piálek J. Cambridge: Cambridge University Press; 2012: 253-77: Cambridge Studies in Morphology and Molecules: New Paradigms in Evolutionary Biology.
8. Portfors CV. Types and functions of ultrasonic vocalizations in laboratory rats and mice. J Am Assoc Lab Anim Sci. 2007;46:28–34.
9. Portfors CV, Perkel DJ. The role of ultrasonic vocalizations in mouse communication. Curr Opin Neurobiol. 2014;28:115–20.
10. Asaba A, Hattori T, Mogi K, Kikusui T. Sexual attractiveness of male chemicals and vocalizations in mice. Front Neurosci. 2014;8:231.
11. Scattoni ML, Crawley J, Ricceri L. Ultrasonic vocalizations: a tool for behavioural phenotyping of mouse models of neurodevelopmental disorders. Neurosci Biobehav Rev. 2009;33:508–15.
12. Chabout J, Sarkar A, Patel SR, Radden T, Dunson DB, Fisher SE, Jarvis ED. A Foxp2 mutation implicated in human speech deficits alters sequencing of ultrasonic vocalizations in adult male mice. Front Behav Neurosci. 2016;10:197.
13. Fernández-Vargas M. Vocal signals of sexual motivation in male and female rodents. Curr Sex Health Rep. 2018;10:315–28.
14. Matsumoto YK, Okanoya K. Phase-specific vocalizations of male mice at the initial encounter during the courtship sequence. PLoS One. 2016;11:e0147102.
15. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (*Mus musculus musculus*). PLoS One. 2017;12:e0188647.
16. Hammerschmidt K, Radyushkin K, Ehrenreich H, Fischer J. Female mice respond to male ultrasonic 'songs' with approach behaviour. Biol Lett. 2009;5:589–92.
17. Musolf K, Hoffmann F, Penn DJ. Ultrasonic courtship vocalizations in wild house mice, *Mus musculus*. Anim Behav. 2010;79:757–64.
18. Asaba A, Osakada T, Touhara K, Kato M, Mogi K, Kikusui T. Male mice ultrasonic vocalizations enhance female sexual approach and hypothalamic kisspeptin neuron activity. Horm Behav. 2017;94:53–60.
19. Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Front Behav Neurosci. 2015;9:76.
20. Musolf K, Meindl S, Larsen AL, Kalcounis-Rueppell MC, Penn DJ. Ultrasonic vocalizations of male mice differ among species and females show Assortative preferences for male calls. PLoS One. 2015;10:e0134123.

21. Kanno K, Kikusui T. Effect of Sociosexual experience and aging on number of courtship ultrasonic vocalizations in male mice. Zool Sci. 2018;35:208–14.
22. Hoffmann F, Musolf K, Penn DJ. Freezing urine reduces its efficacy for eliciting ultrasonic vocalizations from male mice. Physiol Behav. 2009;96:602–5.
23. Hoffmann F, Musolf K, Penn DJ. Ultrasonic courtship vocalizations in wild house mice: spectrographic analyses. J Ethol. 2012;30:173–80.
24. von Merten S, Hoier S, Pfeifle C, Tautz D. A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (*Mus musculus domesticus*). PLoS One. 2014;9:e97244.
25. Wang H, Liang S, Burgdorf J, Wess J, Yeomans J. Ultrasonic vocalizations induced by sex and amphetamine in M2, M4, M5 muscarinic and D2 dopamine receptor knockout mice. PLoS One. 2008;3:e1893.
26. Hanson JL, Hurley LM. Female presence and estrous state influence mouse ultrasonic courtship vocalizations. PLoS One. 2012;7:e40782.
27. Neunuebel JP, Taylor AL, Arthur BJ, Egnor SER. Female mice ultrasonically interact with males during courtship displays. eLife 2015;4:e06203.
28. Warren MR, Spurrier MS, Roth ED, Neunuebel JP. Sex differences in vocal communication of freely interacting adult mice depend upon behavioral context. PLoS One. 2018;13:e0204527.
29. Heckman JJ, Proville R, Heckman GJ, Azarfar A, Celikel T, Englitz B. High-precision spatial localization of mouse vocalizations during social interaction. Sci Rep. 2017;7:3017.
30. White NR, Prasad M, Barfield RJ, Nyby JG. 40- and 70-kHz vocalizations of mice (*Mus musculus*) during copulation. Physiol Behav. 1998;63:467–73.
31. Penn DJ, Frommen JG. Kin recognition: an overview of conceptual issues, mechanisms and evolutionary theory. In: Kappeler P, editor. Animal Behavior: Evolution and Mechanisms. Springer: Berlin Heidelberg; 2010. p. 55–85.
32. Winn BE, Vestal BM. Kin recognition and choice of males by wild female house mice *(Mus musculus)*. J Comp Psychol. 1986;100:72–5.
33. Zinck L, Lima SQ. Mate Choice in *Mus musculus* Is Relative and Dependent on the Estrous State. PLoS One. 2013;8:e66064-e.
34. Suzuki H, Togashi M, Moriguchi Y, Adachi J. Relationship between age-related decline in fertility and *In Vitro* fertilization rate in IVCS mice. J Reprod Develop. 1994;40:107–16.
35. Hoffmann F, Musolf K, Penn DJ. Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. Physiol Behav. 2012;105:766–71.
36. Pomerantz SM, Nunez AA, Bean NJ. Female behavior is affected by male ultrasonic vocalizations in house mice. Physiol Behav. 1983;31:91–6.
37. Byers SL, Wiles MV, Dunn SL, Taft RA. Mouse estrous cycle identification tool and images. PLoS One. 2012;7:e35538.
38. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Automatic mouse ultrasound detector (A-MUD): a new tool for processing rodent vocalizations. PLoS One. 2017;12:e0181200.
39. Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism. PLoS One. 2008;3:e3067.
40. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes Brain Behav. 2011;10:44–56.
41. Grimsley JMS, Sheth S, Vallabh N, Grimsley CA, Bhattal J, Latsko M, Jasnow A, Wenstrup JJ. Contextual modulation of vocal behavior in mouse: newly identified 12 kHz "mid-frequency" vocalization emitted during restraint. Front Behav Neurosci. 2016;10:38.
42. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. [http://www.R-project.org/].
43. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, et al: vegan: Community Ecology Package; R package version 2.5–2; 2018. [http://CRAN.R-project.org/package=vegan].
44. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26:32–46.
45. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA). In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL, editors. Wiley StatsRef: Statistics Reference Online; 2017. p. 1–15. 1–15.
46. Clarke KR, Warwick RM. Change in marine communities: an approach to statistical analysis and interpretation. 2nd ed. Plymouth: Primer-E Ltd.; 2001.

Publisher's Note

CHAPTER 4: Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures

Marconi, Maria Adelaide; Nicolakis, Doris; Abbasi, Reyhaneh; Zala, Sarah, M.; Penn, Dustin J.

Animal Behaviour, impact factor: 3.04; **published**

Animal Behaviour 169 (2020) 169–197

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures

Maria Adelaide Marconi [a, *], Doris Nicolakis [a], Reyhaneh Abbasi [a, b], Dustin J. Penn [a, *, 1], Sarah M. Zala [a, 1]

[a] Konrad Lorenz Institute of Ethology, University of Veterinary Medicine, Vienna, Austria
[b] Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

ARTICLE INFO

Article history:
Received 17 April 2020
Initial acceptance 3 June 2020
Final acceptance 11 August 2020
Available online 14 October 2020
MS. number: 20-00265R

Keywords:
animal signals
courtship behaviour
individual consistency
individual recognition
individual signatures
individual variation
Mus musculus musculus
ultrasonic vocalizations
wild house mice

Acoustic communication mediates many types of social interactions; however, few studies have investigated whether courtship vocalizations contain distinctive individual signatures necessary for individual recognition. Male house mice, *Mus musculus*, produce spectrally complex ultrasonic vocalizations (USVs) during courtship and mating, which appear to attract females and increase male reproductive success. Our goals were to (1) describe quantitative and qualitative changes in the vocalizations of wild-derived male house mice, *M. m. musculus*, induced by a female odour stimulus; (2) measure individual variation and consistency in male USV emission over time; (3) test whether the variation in USVs is greater between than within individuals; and (4) identify individual signatures in spectrotemporal features using univariate statistics, multivariate models and machine learning methods. We recorded males once per week for 3 weeks, used an automated method (A-MUD) to detect USVs, and manually classified them into distinct syllables. We found that most males did not vocalize until they encountered female scent, and then most males dramatically increased the number and the types of different vocalizations (repertoire size and composition). Male USVs showed high interindividual variation and most showed intraindividual consistency, and we found greater inter- than intraindividual variation for both USV count and repertoire size. Male USVs contained individual signatures in most spectrotemporal features, regardless of the method of analysis. Males sometimes produced few, if any, vocalizations when presented with female scent, but consistent nonvocalizers were rare. Our study provides the first evidence that individual signatures in USVs of male house mice are stable over time and across recording trials, although studies are still needed to investigate repeatability across social contexts and to test whether USVs mediate individual recognition.

Individual recognition mediates a wide variety of social interactions (Tibbetts & Dale, 2007; Wiley, 2013; Yorzinski, 2017), but it requires sensory signals that contain distinctive 'individual signatures' (i.e. phenotypic cues with high interindividual and low intraindividual variation, Penn et al., 2007; Favaro, Gamba, Alfieri, Pessani, & McElligott, 2015; Jornod & Roche, 2015). Many species have been found to produce vocalizations that contain individual signatures (reviewed in Shapiro, 2010; Kershenbaum et al., 2016, cited in Stowell, Petrusková, Šálek, & Linhart, 2019). Yet, surprisingly few studies have tested whether male courtship displays, or other secondary sexual traits, contain individual signatures. Sexual signals are often assumed to function either to influence or to provide information to conspecifics ('persuasive' versus 'informative' signals, Grafen, 1990). It has also been argued that signals function to influence rather than provide information to receivers (Rendall, Owren, & Ryan, 2009). Yet, influence and information are not alternative explanations for signalling, and male courtship signals probably influence female receivers because they convey information (e.g. species, sex, kinship and individual identity). In this study, our main aim was to test whether the courtship ultrasonic vocalizations (USVs) of wild-derived male house mice, *Mus musculus musculus*, contain individual signatures.

* Correspondence: M. A. Marconi and D. J. Penn, Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine Vienna, Savoyenstraße 1a, A-1160, Vienna, Austria.
E-mail addresses: Maria.Marconi@vetmeduni.ac.at (M. A. Marconi), Dustin.Penn@vetmeduni.ac.at (D. J. Penn).
[1] These authors made equal contributions.

https://doi.org/10.1016/j.anbehav.2020.09.006

The courtship USVs of male house mice are surprisingly complex and have song-like features, that is, mice produce 10–15 different syllable types in phrases of specific sequences repeated over time (Holy & Guo, 2005; Scattoni, Gandhy, Ricceri, & Crawley, 2008; Grimsley, Monaghan, & Wenstrup, 2011; Mahrt, Perkel, Tong, Rubel, & Portfors, 2013; von Merten, Hoier, Pfeifle, & Tautz, 2014; Matsumoto & Okanoya, 2016; Zala et al., 2019). Males mainly emit USVs when exposed to adult females or their scent (Musolf, Hoffmann, & Penn, 2010). The spectrographic features of USVs have been described in laboratory strains (reviewed in Heckman, McGuinness, Celikel, & Englitz, 2016; Ehret, 2018) and wild-derived mice (Musolf et al., 2010; Hoffmann, Musolf, & Penn, 2012b; von Merten et al., 2014; Musolf, Meindl, Larsen, Kalcounis-Rueppell, & Penn, 2015; Zala, Reitschmidt, Noll, Balazs, & Penn, 2017a). Although both sexes emit USVs during courtship, males emit most (83%) of the vocalizations (Neunuebel, Taylor, Arthur, & Egnor, 2015; Heckman et al., 2017; Warren, Spurrier, Roth, & Neunuebel, 2018), and the USVs of interacting male–female pairs become more complex over time during courtship and mating (Wang, Liang, Burgdorf, Wess, & Yeomans, 2008; Matsumoto & Okanoya, 2016). Females are more attracted to vocalizing than devocalized males (Asaba et al., 2017), recorded playbacks of male USVs versus controls (Pomerantz, Nunez, & Bean, 1983; Hammerschmidt, Radyushkin, Ehrenreich, & Fischer, 2009; Musolf et al., 2010; Nomoto et al., 2018) and especially males with more complex USVs (Chabout, Sarkar, Dunson, & Jarvis, 2015). Males that vocalize obtain higher reproductive success than males that do not vocalize (before and after housing with a female, Kanno & Kikusui, 2018), and males with high reproductive success later vocalize more than other males when presented with an unfamiliar female (Asaba et al., 2017). Males that emit more and longer USVs and larger vocal repertoires when presented with a novel female have higher reproductive success when mated with those same females compared to other males (Nicolakis, Marconi, Zala, & Penn, 2020). Male courtship USVs contain information about a male's species and kinship, and this information influences female preferences. (1) The spectral features of male USVs differ between *Mus* species and females prefer playbacks of USVs from conspecific versus heterospecific males (species recognition, Musolf et al., 2015). (2) USVs also contain distinctive kinship signatures (Hoffmann, Musolf, & Penn, 2012a), and females prefer playbacks of males that are unrelated versus kin (kin recognition, Musolf et al., 2010). Thus, USVs potentially mediate species and kin recognition, and our question here is whether they also provide information about individual identity.

Individual variation is rarely acknowledged in most USV studies, and it is often treated as undesirable noise when comparing groups of mice. Some studies provide evidence for interindividual variation in USVs (e.g. Holy & Guo, 2005; Musolf et al., 2015; Castellucci, McGinley, & McCormick, 2016; Zala et al., 2017a). Interindividual variation is necessary, but not sufficient to show individual signatures. Signals used to mediate individual recognition must contain higher interindividual variation than intraindividual variation. Moreover, repeated measures of individual vocalizations must show high repeatability over time (consistency) and more similarity to each other than to those of other individuals, so that individuals are not easily confused with each other.

Four studies to our knowledge have tested whether mouse vocalizations contain individual signatures, but the results are mixed and inconclusive. (1) A study with laboratory mice (F1 of C57BL/6 x DBA2/J) found evidence for individually distinctive USVs (Holy & Guo, 2005); however, the analysis was conducted only on a subset of seven of 45 mice that emitted the most USVs and had the most diverse syllable repertoire. Thus, a larger and unbiased sample size is needed. (2) A study with CBA/J mice found greater intra- than interindividual variability in most parameters (Mahrt et al., 2013), contrary to predictions of individual signatures. However, males were recorded during direct interactions with females, and the recordings probably included female vocalizations. Each male was recorded with five different females, which may have further inflated intraindividual variation. (3) A study with C57BL/6J mice found that individual vocalizations were consistent and individually distinctive in call rate and duration, but not in other spectrotemporal features (e.g. pitch jumps, median pitch, peak power; Rieger & Dougherty, 2016). However, males in this study were also recorded in opposite-sex pairs (males were recorded only twice and with a different stimulus female each time). Evidence for individual consistency was still found, despite changing stimulus females over trials, but nevertheless, definitive tests for individual signatures require recording individual mice. Moreover, all these studies were conducted with inbred, laboratory mice, *Mus laboratorius*, which are expected to have artificially low interindividual variation compared to wild populations. Therefore, conclusive tests of individual signatures also require analyses of the USVs of outbred, wild or wild-derived mice. (4) The first and only study on individual signatures in the USVs of wild-derived male house mice, *M. m. musculus* (Hoffmann et al., 2012a) found strong evidence for individual signatures, as well as greater similarity between kin than nonkin. Individuals in this study were recorded during a single (90 min) session, indicating that USVs could potentially mediate individual recognition during short-term interactions. However, to determine the stability of individual signatures, studies need to record individuals repeatedly over time (e.g. across recording sessions on different days) and across different social contexts and environmental conditions.

The main aims of our study were to (1) make detailed qualitative and quantitative measures of the vocalizations of wild-derived male house mice, which include syllable repertoire as well as the number of USVs, before and after the mice were presented with female odour, (2) quantify individual variation and individual consistency in USV emission over time and (3) test whether USVs show greater inter- than intraindividual variation (also called repeatability or consistency) over time (i.e. individual signatures). This approach has been used for measuring individual microbial and volatile odour profiles of humans (Penn et al., 2007; reviewed in Parma et al., 2017), as well as individual vocal signatures (Petrusková et al., 2016). (4) We also aimed to determine whether nonvocalizing is a consistent individual trait, since studies on male wild-derived house mice found that some males surprisingly did not emit USVs in response to a sexual olfactory stimulus (Musolf et al., 2010; Hoffmann et al., 2012a; Musolf et al., 2015). We repeatedly recorded wild-derived mice (F3 from wild-caught *M. m. musculus)* before and during the presentation of a female urinary odour stimulus (three trials over 3 consecutive weeks), and described USV count, repertoire size and composition. We quantified the interindividual variation and intraindividual consistency, and used the potential of individual coding (PIC) to identify parameters as candidates for individual discrimination. Additionally, we paired these methods with multivariate statistics and machine learning methods to test the potential of selected acoustic features for individual signatures.

METHODS

Subjects and Housing

We recorded vocalizations from 22 sexually experienced adult male mice, which were F3 generation of wild-caught house mice living at the Konrad Lorenz Institute of Ethology (48°12′38′N, 16°16′54′E) in Vienna, Austria. Individuals were kept in mixed-sex

family groups (standard Type IIL cages, 36.5 × 20 cm and 14 cm high, stainless steel cover, 1 cm mesh width, Tecniplast, Hohenpeißenberg, Germany) until weaning (21 days). At weaning, mice were housed in mixed-sex groups (maximum of four animals per cage) until 5 weeks of age when females were housed in sister pairs and males were singly housed to avoid fighting. Each cage contained wood shavings (ABEDD, Vienna, Austria), nesting material (Nestlet, Ehret, Austria), a nestbox (Tecniplast, Buguggiate, Italy) and a cardboard paper roll for environmental enrichment. Mice were provided with food (rodent diet 1324, Altromin, Lage, Germany) and water ad libitum. Colony rooms were kept at standard conditions (room temperature: mean ± SD = 22 ± 2 °C, in a 12:12 h light:dark cycle with red light on at 1500). Test males ($N = 22$, mean age ± SD = 369 ± 40 days at first recording day) were recorded in a previous experiment with a female and then housed in a cage with her for breeding (breeding regime: mean ± SD = 9 ± 2 weeks). At the end of the breeding regime, individual males were separated and singly housed for a mean ± SD of 9 ± 2 weeks until recording.

Female Urinary Odour Collection (Sexual Stimulus)

We used female urine to elicit male USV emission, and we collected urine from 28 wild-derived female house mice (mean age ± SD = 356 ± 28 days at first sampling day) over 10 days. Two mice were resampled once to achieve sufficient volume. Urine collection and pooling were performed 2 weeks before the recordings. Females were handled using a small, clean plastic cylinder, where they urinated, and urine was then pipetted and stored at −80 °C. We pooled urine samples, and systematically assigned females to males to control for several factors that could potentially generate biases due to variation in female odour. (1) Urine from each female was combined over the 10 days to reduce intra-individual variation in scent due to oestrus or other reasons, i.e. we sampled individuals over 5 consecutive days for 2 weeks to get females in oestrus on 1–2 sampling days (Byers, Wiles, Dunn, & Taft, 2012). (2) We combined 20 µl of urine from three females to reduce interindividual variation (i.e. 60 µl urine stimulus). (3) We controlled for potential variation in urinary scent due to females' sexual experience combining the scent of one sexually experienced female (housed with a male in our breeding colony) with a female that had lived in seminatural conditions with males and one socially experienced virgin (by a repeated 2 h exposure through a perforated separation wall to males and females). Thus, each male was weekly presented with a urine pool, which included two sexually experienced females and one virgin. (4) Donor females were not closely related (not siblings) or familiar to the male subjects. (5) Urine from each female was used a maximum of three times in different urine pools in the same recording week and nine times over 3 weeks, and urine samples from each female were dispersed across different males, so that there was more variation in individual female donors within than between males. The same male was never exposed to the same females again to minimize potential effects of female identity. Thus, our design for collecting and assigning female stimuli could potentially increase intra- versus interindividual variation in odour, but not vice versa, and therefore our results should be conservative.

Experimental Assay and Recording Procedure

We recorded male vocalizations first without and then during presentation of a female urine stimulus and we compared individual USV count and repertoire size during these two phases, and over 3 recording weeks. We conducted our experiments under red light during the most active period of the day (1500–1830) for the mice. Each male was removed from his home cage and transferred in a plastic cylinder where he was weighed and then placed in a clean cage (36.5 × 20 cm and 14 cm high) with fresh bedding and a stainless-steel cover (1 cm mesh width). The cage was put in a separate room from the observers under red light and the male was recorded for 5 min without any stimulus (prestimulation phase). After 5 min, we offered a pool of thawed female urine on a cotton swab clipped on the cage cover and recorded the individual for 10 min (stimulation phase). Each male was recorded once per week over 3 consecutive weeks. Every week, each male was recorded at approximately the same time of day, to avoid any time effects between days. Mice were video recorded with an IP camera (D-Link DCS-3710) and audio recorded using an ultrasound microphone (USG Electret Ultrasound Microphone, Avisoft Bioacoustics/Knowles FG, Brandenburg, Germany) positioned 10 cm over the centre of the cage. The audio recording set-up included the microphone, an A/D converter (UltraSoundGate 416Hb, Avisoft Bioacoustics) and a laptop (Lenovo T540p, Windows 7) with RECORDER USGH software (Avisoft-RECORDER Version 4.2). Recording settings included a 300 kHz sampling rate and 16-bit format. We video recorded the behaviour of our subjects using an open source software (iSpy–Video Surveillance Software, www.ispyconnect.com).

USV Detection and Analyses

To analyse spectrograms, we used the software S_TOOLS-STx (Version 4.3.8, 9374, Acoustics Research Institute, Vienna, Austria) and we ran the automatic mouse ultrasound detector (A-MUD 3.0) for USV detection and parameter extraction (Zala, Reitschmidt, Noll, Balazs, & Penn, 2017b). We enabled the detection function with a minimum length of 5 ms for each USV and, when necessary, we manually adjusted the length of the USVs (see Nicolakis et al., 2020). The transcription function produced the spectrogram with a time window overlap of 75% (4 ms frame, Hanning window between 0 and 150 kHz) and a range of 50 dB (floor set at −80 dB for all recordings). All the sound files were visually and acoustically checked and USVs were classified according to 15 categories modified from previous literature (Scattoni et al., 2008; Scattoni et al., 2011; Hanson & Hurley, 2012; Hoffmann et al., 2012b; Musolf et al., 2015) according to their length, frequency and frequency modulation (as described in Zala et al., 2019, see Fig. 1). The spectrogram visualization was in 2 s steps, and 25% of the spectrogram view overlapped with the previous one. After the visual check, A-MUD 3.1 was run to extract spectrotemporal features for each vocalization (starting time, USV length and minimum, mean and maximum frequency). Computed parameters and manually classified USVs for each sound file were saved into an Excel file (Excel 2016, Microsoft Office Professional Plus) for statistical analysis. We classified 24 376 USV syllables from 66 recordings of 22 individual males (ca. 17 h of recording in total).

For analyses, each 15 min recording was divided into a 5 min prestimulation phase (phase 0) and a 10 min odour stimulation period, which was divided in half (phases 1 and 2) so that we could compare these three 5 min blocks. To compare the 5 min prestimulation phase (phase 0) with the stimulation period, we used only the first 5 min (phase 1). The two halves of the stimulation period (phases 1 and 2) were highly correlated and showed similar results for all the tests performed and did not indicate increased or short-term habituation (see Appendix Table A1). We defined three main vocal parameters: USV count (total number of USVs emitted during the recording time), repertoire size (number of different syllable types uttered during each recording, 1–15 USV categories, see Fig. 1) and repertoire composition (number of USVs emitted for each syllable type). The USV mean frequency was calculated as the grand mean of the mean frequency over all USVs of each recording

Definition	Spectrograhic shape	USV category	USV label
USV categories with frequency > 91 kHz			
Frequency > 91 kHz (all the shapes and duration < or > 10 ms)	Any shape	Ultra high	uh
USV categories with frequency < 91 kHz and duration (USV length) < 10 ms			
Frequency < 91 kHz Duration < 5 ms (usually short dots)	Any shape	Ultra short	us
Frequency < 91 kHz Duration < 10 ms (all shapes)	Any shape	Short	s
USV categories with frequency < 91 kHz and duration (USV length) > 10 ms			
Frequency modulation < 5 kHz		Flat	f
Frequency modulation: a decrease in frequency > 5 kHz		Down	d
Frequency modulation: an increase in frequency > 5 kHz		Up	up
Frequency modulation: decrease + increase in frequency (each) > 5 kHz		U-shaped	u
Frequency modulation: an increase + decrease in frequency (each) > 5 kHz		U-shaped inverted	ui
Frequency modulation: ≥ 2 directional changes in frequency > 5 kHz		Complex	c
Frequency modulation: 2 elements with 1 frequency jump and no time separation		Complex 2	c2
Frequency modulation: 3 elements with 2 frequency jump (no time separation)		Complex 3	c3
Frequency modulation: 4 elements with 3 frequency jump (no time separation)		Complex 4	c4
Frequency modulation: 5 elements with 4 frequency jump (no time separation)		Complex 5	c5
Frequency modulation: syllables with 1 harmonic element		Harmonic	h
Not belonging to any USV category for frequency or duration (not defined or measurable spectrographic features)		Unclassified	uc

Figure 1. Ultrasonic vocalization (USV) repertoire classification. Examples of 15 types of USV syllables emitted by male mice, defined by duration and frequency parameters (adapted from Zala et al., 2019).

(Zala et al., 2017a). We also calculated the mean frequency bandwidth as the difference between the maximum and the minimum frequency for each syllable. For the stimulation period (phases 1 and 2 and the entire 10 min), we computed the intercall interval (ICI) which we defined as the temporal gap between two vocalizations within one recording. We divided the ICI into two separate parameters: the mean intersyllable interval (ISI), defined as the gap between two consecutive USVs within a bout (<300 ms), and the mean interbout interval (IBI) which we defined as the gap between one sequence of vocalizations (bout) and the following one (>300 ms, Chabout et al., 2015; Castellucci, Calbick, & McCormick, 2018). We also computed the total USV length (time the mouse spent vocalizing during one phase and trial), and the latency to the first USV as the time from the start of each recording to the first USV emitted by the mouse. Each parameter was computed separately for each mouse, each phase and trial.

Statistical Analyses

All variables were inspected for normal distribution and homogeneity of variance using the Kolmogorov–Smirnov test and the Levene's test. We used a square-root transformation for the USV count (sqrt(USV count+0.5)) and the repertoire composition for making correlations and other comparisons. We used percentages of the syllable types to correct for mice that emitted few USVs, unless stated otherwise. Most of our data had a negative binomial distribution and were highly right skewed, which do not meet the assumptions of parametric models. There is no consensus regarding the most appropriate method for analysing individual signatures when data are highly skewed (Linhart et al., 2019). Thus, when the assumptions for parametric tests were not met, we performed nonparametric statistical tests. Results are presented as mean ± SD, unless specified otherwise. We conducted two-tailed tests and

statistical significance was set at $\alpha \leq 0.05$. All tests were corrected for multiple testing with the false discovery rate (FDR) and post hoc tests were calculated with the Benjamini–Hochberg procedure. We performed statistical tests using SPSS (IBM Statistics 25, IBM, Armonk, NY, U.S.A.) and RStudio (R version 3.5.1, R Core Team, 2018). From R, we ran the package 'vegan' (Oksanen et al., 2018), with the functions 'vegdist', 'anosim' and 'metaMDS' for the nonparametric analysis of similarities (ANOSIM) and the nonmetric multidimensional scaling (NMDS) graphs. Spearman rank correlations were computed to remove highly correlated parameters in both the t-distributed stochastic neighbour embedding (t-SNE; van der Maaten & Hinton, 2008; van der Maaten, 2009) and the recursive function elimination based on the random forest classifier (RFE). Correlations were run and graphed with the package 'corrplot' (Wei & Simko, 2017). The t-SNE was run with the package 'tsne' (Donaldson, 2016). The random forest classifier required the packages 'randomForest' (Liaw & Wiener, 2002), 'caret' (Kuhn et al., 2018) for parameter selection by RFE, 'reshape2' (Wickham, 2007) and 'ggplot' (Wickham, 2006) for data preparation and graphs.

To describe how male mice modulated their USV emission upon encountering female odour, we analysed the variation in USV count, repertoire size and composition between stimulation phases and sampling weeks using Friedman tests. We used the mean percentages of all individuals for each syllable type for each phase and trial to control for differences between individuals and create pie charts. We angular transformed the repertoire composition (proportion of USVs/syllable type) to run Friedman tests for each syllable type within and between phases and weeks. We present the repertoire composition in each phase and trial using pie charts for all mice, and bar charts for examples of three vocalizing mice to show the individual variation during each phase and trial. Since USV count and repertoire size showed similar patterns between phases and over time, we investigated the relationship between the square-root-transformed USV count and repertoire size for each phase and sampling week, including the full 10 min stimulation period, by performing Spearman rank correlations. To test for interindividual variation in each phase and over time, we compared the distribution of both USV count and repertoire size before and during males' stimulation and we tested them for kurtosis and skewness in each recording week (for the USV count distribution see also Zala et al. (2017a) and Zala et al. (2019)). We then tested intraindividual consistency in USV count and repertoire size through correlations between the 3 weeks for each phase with Spearman correlation tests. Thus, we determined whether these parameters might be candidates for individual signatures by performing univariate tests (Friedman tests) on the entire 10 min stimulation period. Since we found higher variation between than within individuals in USV count and repertoire size, we ranked the mice according to their variation in USV emission (from the lowest to the highest SD of the USV count) assigning them the code 1–22 and we kept this order for all our models.

To quantify the individual variation between males and determine which parameters provide candidates for individual signatures, we calculated the PIC (Favaro et al., 2015). We examined 27 USV parameters and used four different models. The first two models included 13 mice vocalizing in phases 1 and 2 in all 3 weeks and had equal numbers of USVs ($N = 1872$, model 1) and recordings ($N = 39$, model 2), respectively. The other two models sampled 19 mice vocalizing at least during phase 1 or 2 and included all USVs ($N = 23\,414$, model 3) and all recordings ($N = 57$ recordings, model 4; see Appendix). We performed Friedman tests for each parameter when we had equal sample sizes (i.e. number of USVs/mouse in the first model and number of recordings/mouse in the second and fourth models), to compare individual variation between weeks and correct with FDR (Benjamini–Hochberg procedure) for multiple testing.

To obtain more robust assumptions about the candidate vocalizations for individual signatures, we performed ANOSIM and used a machine learning method (random forest). We examined separately the repertoire composition for each mouse and recording day (3 weeks) with ANOSIM (Clarke & Warwick, 2001) and graphed our results using NMDS (Clarke & Warwick, 2001; see Appendix). Then, we used machine learning and ran the RFE to select the potential individual signatures and the random forest classifier to assign the recordings to the individuals (Breiman, 2001). Before implementing machine learning methods (RFE and random forest classifier) and visualizing the data (t-SNE), we ran Spearman correlation matrices to remove highly correlated parameters and kept one of them from each group to run the models (Murdoch & Chow, 1996; Friendly, 2002). We graphically presented the data run with the remaining parameters using the t-SNE. Then, we implemented the RFE including all 26 USV parameters previously computed with the PIC. The RFE sequentially reduces the number of parameters and selects the candidates as individual signatures. Finally, we trained the random forest classifier to learn how to correctly identify the individuals using the relationship between the selected acoustic parameters and the ground truths (correct individual identities). Thus, we report the accuracy in correctly identifying individuals (random forest classifier). We computed three different models with different numbers of mice ($N = 13$ males vocalizing in both phases or $N = 19$ males vocalizing at least in one stimulation phase) and recordings from the stimulation period (i.e. three recordings when the entire 10 min of stimulation was used and six recordings when phase 1 and phase 2 were added separately; see Appendix). To assess possible mechanisms for individual signalling, we also tested whether higher percentages of some syllable types in the repertoire composition influenced the spectrotemporal features chosen by the RFE (e.g. a high percentage of short USVs would affect the USV length or a high percentage of vocalizations with a small difference between the start and the end frequency of the vocalization might influence the frequency bandwidth). We correlated each spectrotemporal feature (grand mean frequency, frequency bandwidth and mean USV length) with the percentage of each syllable type in each week and corrected our significant results with the FDR for multiple testing ($N = 19$). We also inspected the candidates for individual signatures detected by the RFE, for intraindividual consistency and tested whether they experienced a decline or an increase over the 3 weeks. Thus, we performed correlations between consecutive weeks for each spectrotemporal feature and percentage of specific syllable types chosen by the RFE and also included the percentage of 'up' since it was the most common syllable type used by males during our recordings ($N = 19$).

Ethical Note

The study was conducted in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All experiments were carried out at the Konrad Lorenz Institute of Ethology (Vienna, Austria) and all protocols have been approved and followed guidelines and ethical standards in the care and use of experimental animals of the Ethical and Animal Welfare Commission of the University of Veterinary Medicine (ETK-17/04/2015, Vienna, Austria). We did not kill any of the mice used for our study, and they were returned to our colony.

RESULTS

Male Response to Female Odour Stimulus

The 22 males rarely vocalized during prestimulation (phase 0) and they emitted a grand mean of 10 ± 21 USVs/5 min over 3

weeks. Once presented with the odour stimulus, males showed a 17× increase in USV count (phase 1: grand mean 173 ± 184 USVs; phase 2: grand mean 186 ± 207 USVs; see Fig. 2a–c, Appendix Tables A1, A2). USV emission remained high and there was no difference for the USV count or repertoire size between the first and second half of the stimulation period in each week (Appendix Table A1) or between the 3 weeks (Appendix Table A2). Repertoire size showed a similar pattern (grand mean 2 ± 1 syllable types in phase 0 versus 7 ± 5 syllable types in phase 1 and phase 2; see Fig. 2d–f, Appendix Table A2). Phase 0 was always significantly different from the other phases for USV count, repertoire size and repertoire composition (see Fig. 3, Appendix Fig. A1; Table A3). USV count and repertoire size increased over the 3 weeks during phase 0, but not during the stimulation phases, although this difference was not statistically significant (see Fig. 2, Appendix Table A4).

When we compared the stimulus phases within each of the 3 weeks we found that syllable repertoire had greater spectral diversity and complexity during odour stimulation than prestimulation, and that this short-term increase occurred each week (see Fig. 3, Appendix Fig. A1; Table A3). Over the 3 weeks, males increased the repertoire composition between phases. In the first week, phase 0 differed by 8/15 syllable types from phase 1, but by the third week, it differed by only 4/15 syllable types; thus, the two phases were more similar in week 3. Males did not emit significantly more diverse or spectrally complex syllable types between the 3 weeks and the three stimulation phases after FDR correction, but our power to detect differences is low. During prestimulation, the percentage of 'up' increased from week 1 to week 3 although not significantly, and during stimulation more than 50% of the total number of USVs were 'up' vocalizations emitted by all males (see Fig. 3, Appendix Fig. A1; Table A5).

Relationship Between USV Count and Repertoire Size

We investigated the relationship between male USV count and repertoire size during the stimulation period since they showed the same pattern between phases and over time (Fig. 2). As expected, USV count and repertoire size were positively correlated in a nonlinear, asymptotic curve (Fig. 4). These nonlinear correlations

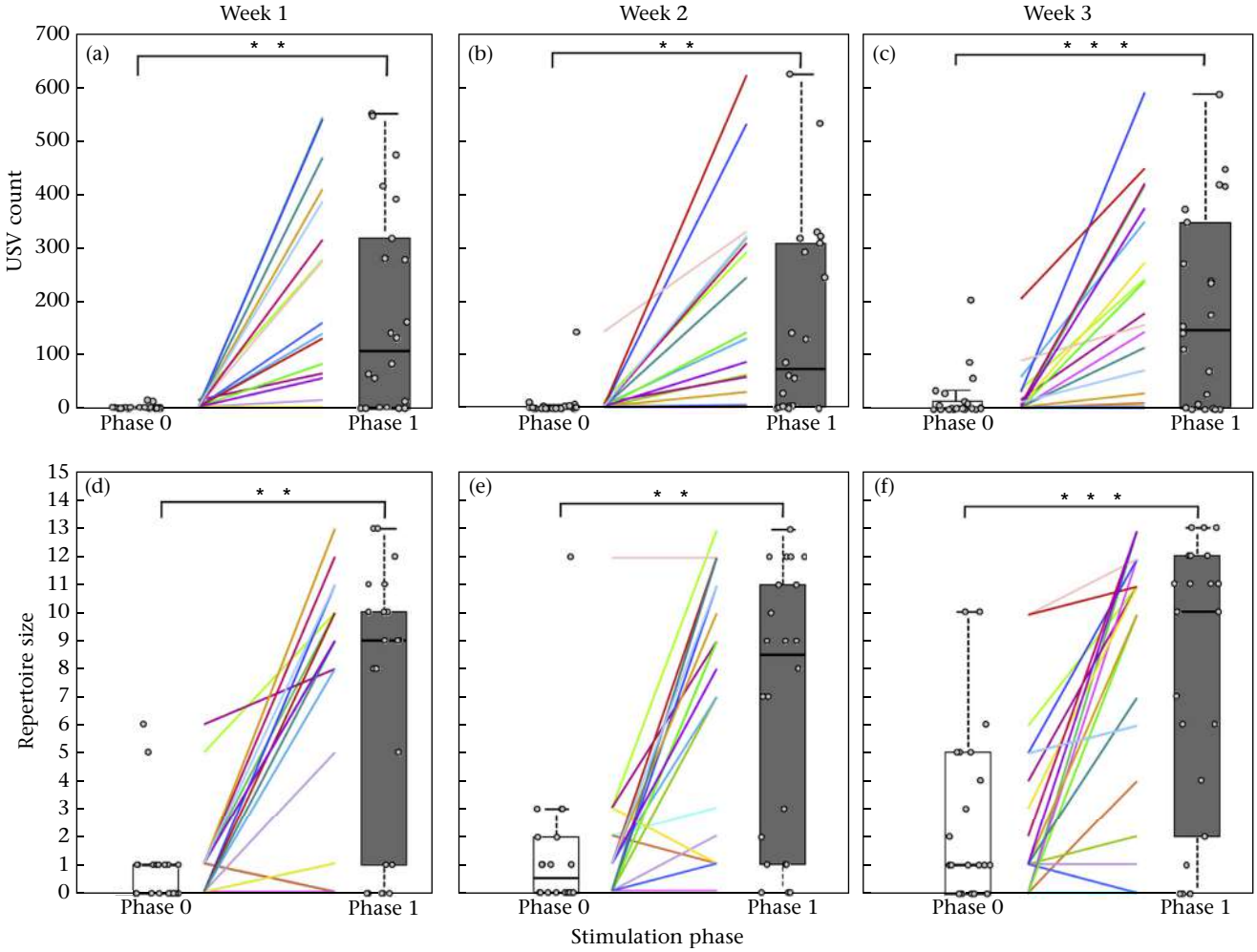

Figure 2. Individual male USV emission before and during sexual odour stimulation. (a, b, c) USV count and (d, e, f) repertoire size during the 5 min prestimulation phase (phase 0, white bars) and during the first 5 min of the sexual stimulation period (phase 1, grey bars) for (a, d) week 1, (b, e) week 2 and (c, f) week 3. Box plots display the median, interquartile ranges (25th and 75th percentile) and 95% confidence interval (whiskers). Circles represent individual males. $^{**}P < 0.01$; $^{***}P < 0.001$. Each colour in the line graphs (shown between the two box plots) represents an individual male's USV count or repertoire size. See Appendix Tables A1 and A2 for *P* values.

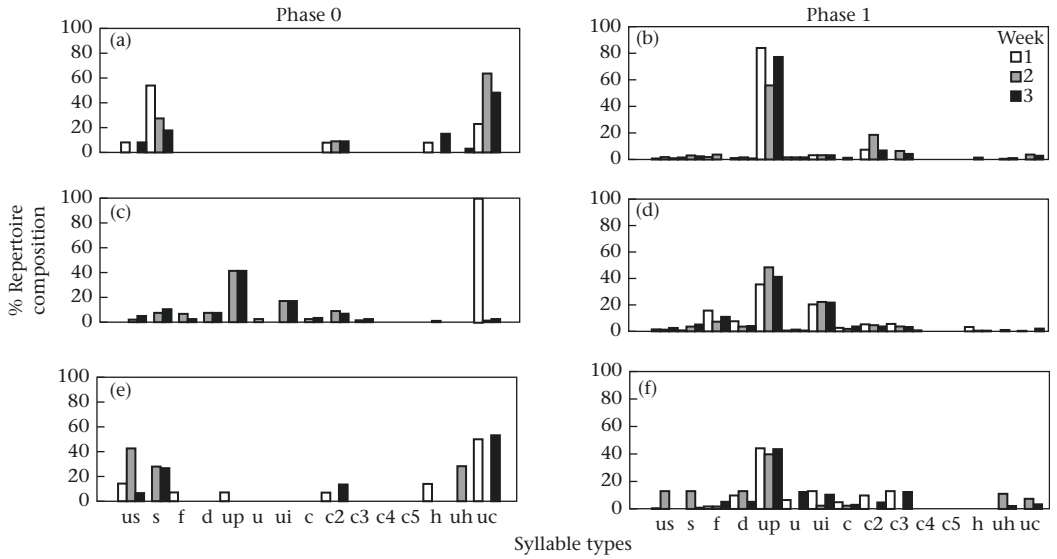

Figure 3. Repertoire composition before and during stimulation (three exemplars). (a, b) Male 1, (c, d) male 2 and (e, f) male 3. (a, c, e) Prestimulation phase 0 and (b, d, e) phase 1 (the first 5 min of the stimulation period). The 15 syllable types in the repertoire are shown on the *x* axis. The *y* axis shows the percentage of vocalizations for each syllable type and in each week. The three mice vocalized 1–142 USVs in phase 0 (a: 20 ± 11; c: 12 ± 4; e: 77 ± 58 USVs) and 56–330 USVs in phase 1 (b: 271 ± 22; d: 99 ± 55; f: 254 ± 73 USVs) over 3 weeks.

were significant in all 3 weeks and phases separately (all phases: $P < 0.001$) and the results did not change by adding phase 1 and phase 2 (10 min stimulation period; see Fig. 4, Appendix Table A6).

Interindividual Variation

We quantified interindividual variation in USV count (Fig. 5) and repertoire size (Appendix Fig. A2). During the prestimulation phase (phase 0), males emitted 0–205 USVs and 0–12 different syllable types. Nearly half of the males (41%) did not vocalize whatsoever during this phase, and although vocalizing males tended to show an increase in USV count and repertoire size during this phase over the 3 weeks, this difference was not statistically significant. During the first 5 min of the stimulation period, males emitted 0–625 USVs. Most (15/22) males vocalized in all 3 weeks (grand mean 246 ± 179 USVs; Fig. 5) and had a repertoire size of 1–13 syllable types (Appendix Fig. A2). During this phase, male USV count was moderately and positively skewed (week 1: skewness = 0.780, kurtosis = −0.822; Shapiro–Wilk test: $W = 0.837$, $N = 22$, $P = 0.002$; week 2: skewness = 1.126, kurtosis = 0.513; Shapiro–Wilk test: $W = 0.820$, $N = 22$, $P = 0.001$; week 3: skewness = 0.655, kurtosis = −0.686; Shapiro–Wilk test: $W = 0.884$, $N = 22$, $P = 0.014$). However, repertoire size was moderately and negatively skewed (week 1: skewness = −0.468, kurtosis = −1.467; Shapiro–Wilk test: $W = 0.834$, $N = 22$, $P = 0.002$; week 2: skewness = −0.333, kurtosis = −1.605; Shapiro–Wilk test: $W = 0.858$, $N = 22$, $P = 0.005$; week 3: skewness = −0.502, kurtosis = −1.429, Shapiro–Wilk test: $W = 0.844$, $N = 22$, $P = 0.003$). USV count and repertoire size in phase 2 showed a similar pattern and were not significantly different from those in phase 1.

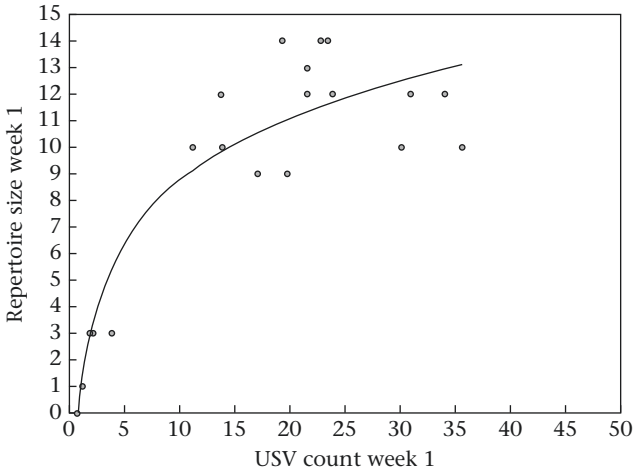

Figure 4. Relationship between individual USV count and repertoire size. The asymptotic curve shows the correlation between the USV count (sqrt(USV count + 0.5)) and the repertoire size during odour stimulation (10 min) in week 1 ($N = 22$). The curves for weeks 2 and 3 show similar patterns (graphs not shown).

Intraindividual Consistency

Most males (16/22 or 73%) did not vocalize at all during phase 0 in at least one of the 3 weeks (grand mean 10 ± 26 USVs), and only two individuals did not vocalize at all during the 10 min stimulation period in all 3 weeks. During phase 0, 59% (13/22) of the mice vocalized in at least 1 of the 3 weeks, but not consistently over time. Thus, USV count and repertoire size did not correlate between weeks. In contrast, during the first 5 min of stimulation, males emitted between 0 and 625 USVs and 15 consistent vocalizers emitted 246 ± 179 USVs over the 3 weeks (see Fig. 6, Appendix Table A7). Individual USV count was correlated between consecutive weeks (i.e. week 1 versus 2, and week 2 versus 3; $P_{\text{FDRcorrected}} \leq 0.001$, phases 1 and 2), and the correlation was more similar than between nonconsecutive weeks (i.e. week 1 versus week 3: $P_{\text{FDRcorrected}} = 0.016$ in phase 1 and $= 0.012$ in phase 2; see Fig. 6a, Appendix Table A7). Repertoire size in phase 1 was correlated between weeks (i.e. week 1 versus 2: $P_{\text{FDRcorrected}} < 0.001$; week 2 versus 3 and week 1 versus 3: $P_{\text{FDRcorrected}} = 0.012$). However, in phase 2 it was only correlated between weeks 2 and 3 ($P_{\text{FDRcorrected}} < 0.001$), but not between the other 2 weeks (see Fig. 6c, Appendix Table A7). Mice also showed high consistency in their USV count and repertoire size within the same sound file

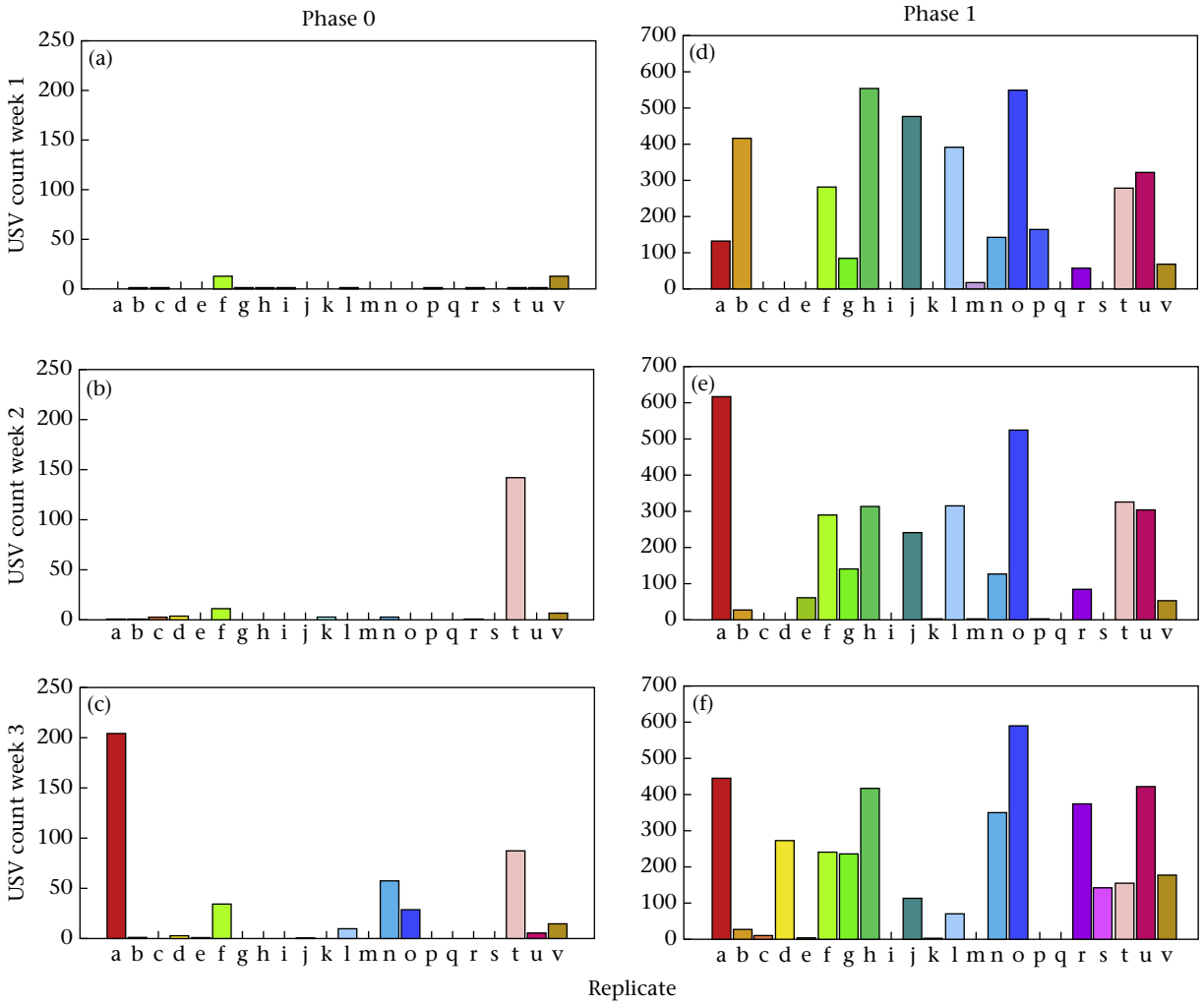

Figure 5. Interindividual variation in USV count before and during sexual stimulation. USV counts are shown during (a, b, c) phase 0 and (d, e, f) phase 1 over (a, d) week 1, (b, e) week 2 and (c, f) week 3 ($N = 22$). Note that graphs for phase 0 and phase 1 have different y axes. The same letter and colour correspond to the same individual (replicate).

when the first part of the recording was correlated with the second part of the recording during stimulation (phase 1 versus phase 2) in all 3 weeks (see Appendix Fig. A3; Table A8).

Comparing Inter- versus Intraindividual Variation

For the male USV count (using data from the entire 10 min stimulation period, we found significantly greater differences between than within individuals (Friedman test: $N = 22$, $\chi^2 = 51.465$, $P < 0.001$; Fig. 7a). Some individuals had a more distinctive USV count than others, once again indicating that some mice were consistently low or high vocalizers.

For male repertoire size, we again found significantly greater differences between than within individuals (Friedman test: $N = 22$, $\chi^2 = 54.222$, $P < 0.001$; Fig. 7b). Some individuals had a more distinctive repertoire size than others. Thus, individuals appeared to cluster into several different groups according to the USV count and repertoire size (see Fig. 8). We could graphically and statistically detect five different groups (Kruskal–Wallis test: $H_4 = 19.89$, $N = 22$, $P = 0.001$): (1) consistent non and low vocalizers ($N = 6$, having consistently low USV count and repertoire size); (2) inconsistent low-level vocalizers ($N = 4$, with low USV count and high variation in repertoire size); (3) consistent intermediate-level vocalizers ($N = 4$, with intermediate USV count and repertoire size) whose USV count was close to the mean and median USV number (e.g. mean range for four mice: 294–444 USVs where grand mean = 359 USVs and mean of the median = 323 USVs over 22 males); (4) inconsistent high-level vocalizers ($N = 3$, high variation in USV count but small variation in repertoire size); and (5) consistent high-level vocalizers ($N = 5$, high consistency for both parameters).

Candidates for Individual Signatures

To be a good candidate for an individual signature, a parameter should have a PIC > 1 (see Appendix). Model 1, which had the same USV count for each individual ($N = 1872$ USVs) and was run on a subset of mice ($N = 13$) over 3 weeks (Appendix Table A9), included six parameters of which the frequency bandwidth, the ICI and the USV length all had a PIC > 1. The minimum, mean and maximum frequencies had a PIC close to 1 (0.78 < PIC > 0.81). Friedman tests showed that all parameters were significantly different between individuals after FDR correction for multiple testing (Appendix Table A9). Model 2, run on three recordings per mouse and week ($N = 39$ recordings and 20 parameters), showed that all percentages of syllable types had a PIC > 1. In more detail, the percentage of less common USVs ('c5', 'uh', 'c4', 'uc' and 'h'), 'c3' and 'up' had a PIC > 2. However, we found a significant difference only for the

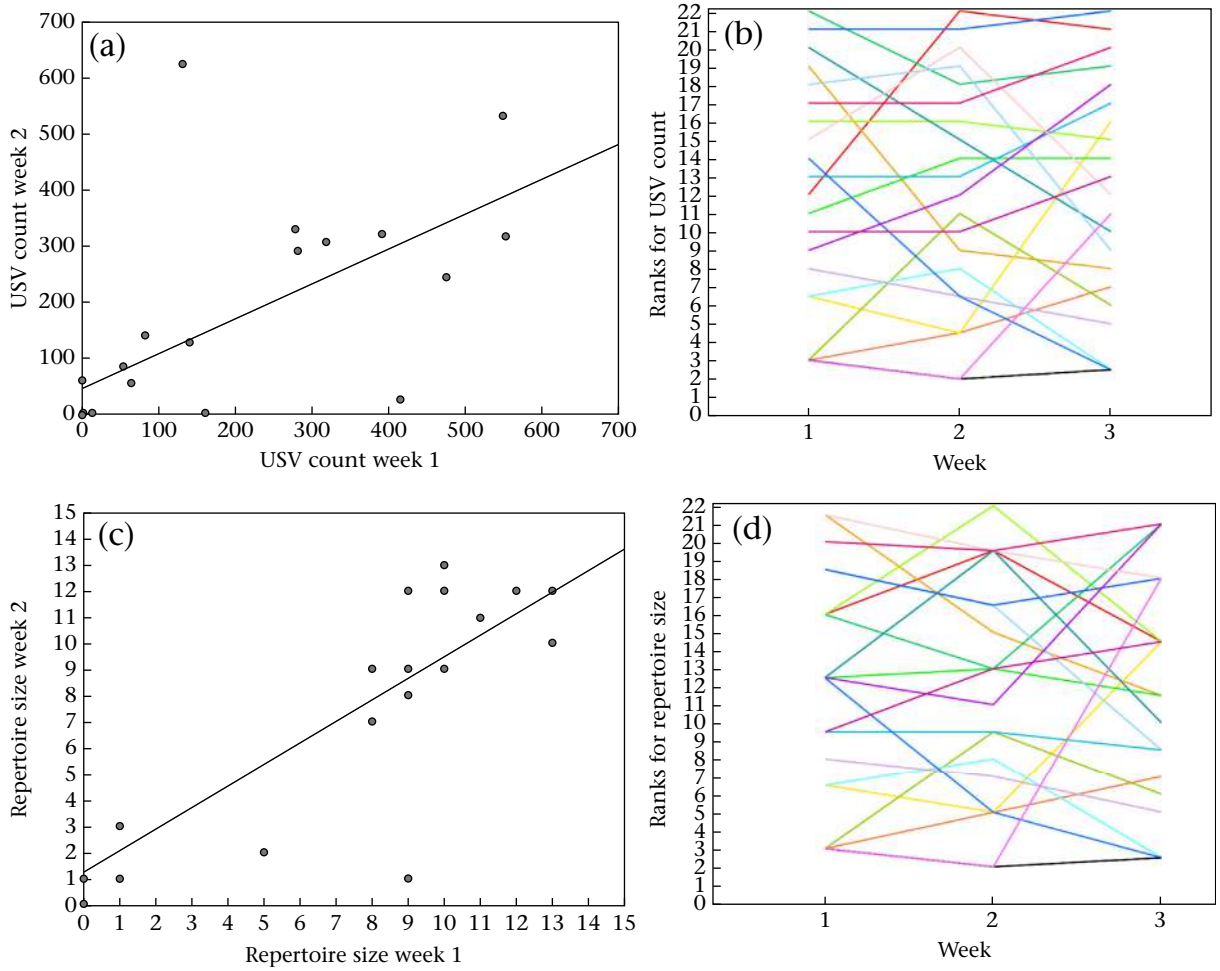

Figure 6. Intraindividual consistency in male USV emission during sexual stimulation. (a) Correlation of USV counts between weeks 1 and 2 for each male. (b) Ranks of USV counts for each male over the 3 weeks. (c) Correlation of repertoire sizes between weeks 1 and 2 for each male. (d) Ranks of repertoire sizes for each male over the 3 weeks. $N = 22$. Results are shown from the first 5 min of the stimulation period (phase 1). (b, d) Individuals (represented by lines of different colours) are ranked according to their USV count or repertoire size from the lowest to the highest in each week.

percentage of more common syllable types ('c3', 'up', 'ui', 'd', 'u' and 'f') between individuals recorded over the 3 weeks. The IBI, ISI, repertoire size and total USV length had a PIC > 2 and latency to the first USV had a PIC > 1 and were still significantly different between weeks and individuals after FDR correction (Appendix Table A9). Model 3 included six parameters and all vocalizing mice ($N = 19$) and all recorded USVs ($N = 23\ 414$). In contrast to the results of model 1, the maximum, mean and minimum frequencies had a PIC > 1, but the frequency bandwidth was only close to 1 (PIC = 0.80). We still found the ICI had a PIC > 1 but the USV length had a PIC just close to 1 (PIC = 0.95; Appendix Table A9). Model 4 included three recordings per mouse and week ($N = 57$ recordings and 21 parameters). As for model 2, all the syllable types had a PIC > 1. The percentage of less common types ('h', 'c5', 'us', 'uh', 'c' and 'c4') had a PIC > 2 and showed significant differences (only for 'h', 'uh' and 'c') after FDR correction for multiple testing. However, unlike model 2, the percentage of 'c' had a PIC > 2, whereas the percentage of 'c3' and 'uc' had a lower PIC (PIC = 1.7 and PIC = 1.9, respectively) and the percentage of 'uc' was not significantly different between mice and weeks. As previously found, the percentage of more common syllable types ('c3', 'd', 'f', 'c2', 'ui', 'u' and up') were still significantly different between mice and weeks after FDR correction for multiple testing. Moreover, the IBI had a PIC > 2 and the latency to the first USV had a PIC higher than in the previous model (PIC = 2.2 versus PIC = 1.7). The ISI, repertoire size, total USV length and USV count had a PIC > 1 and together with the IBI and the latency to the first USV were significantly different between individuals and over the 3 weeks after FDR correction for multiple testing (Appendix Table A9). Mice emitted USVs with a grand mean ISI of 139 ± 26 ms and a grand mean IBI of 34 ± 60 s. In summary, all models confirmed most of the parameters as individual signatures. Models 1 and 3, which used parameters related to each USV, selected different candidates with PIC >1 but the other candidates were close to 1. Models 2 and 4, run on parameters calculated for each recording, confirmed most of the parameters as individual signatures. We observed a general decrease in the PIC values from model 2 to 4, possibly due to an increasing number of USVs (i.e. from 1872 to 23 414) and an increase in PIC values for the IBI and latency to the first USV.

Repertoire Composition

Individual males showed moderate, but significant clustering according to differences in their repertoire composition (ANOSIM: $R = 0.562$, $N = 19$, $P = 0.001$), and the NMDS presented a moderate separation of the individuals (stress value = 0.17; Fig. 9).

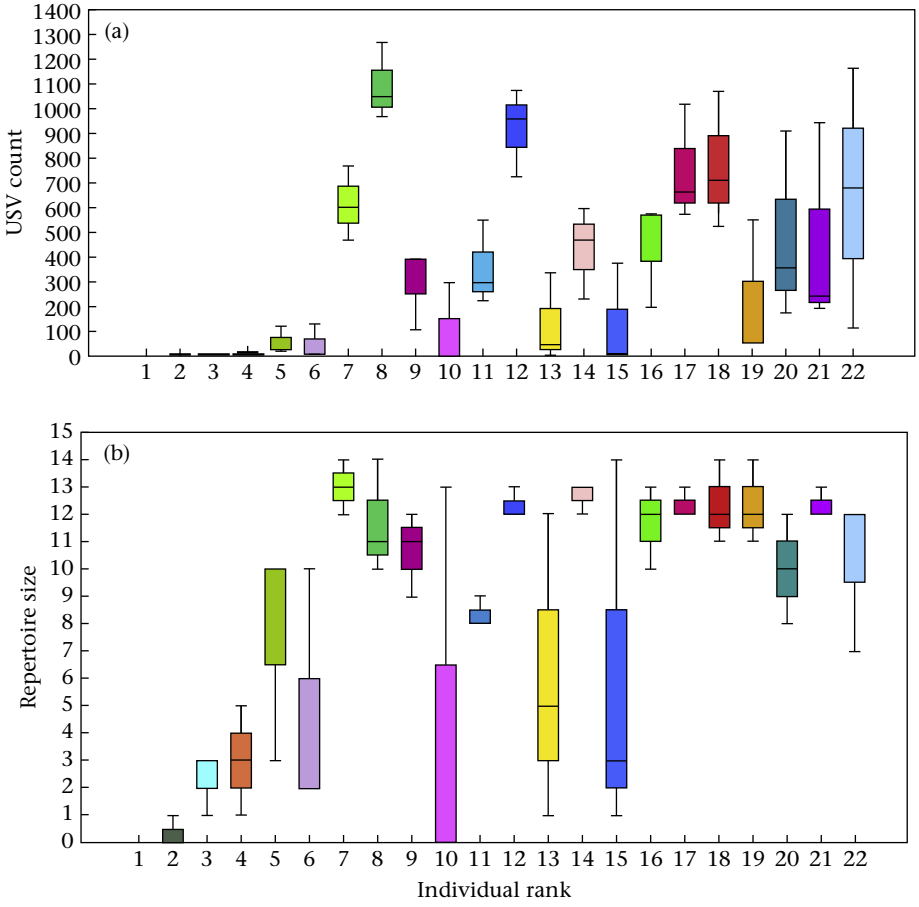

Figure 7. Individual variation in USV emission during sexual stimulation. (a) USV count and (b) repertoire size comparing between- versus within-individual differences during the 10 min odour stimulation period over 3 weeks ($N = 22$). Box plots display the median, interquartile ranges (25th and 75th percentile) and 95% confidence interval (whiskers). The same number and colour correspond to the same individual. Males are sorted according to their individual variation in USV count (i.e. from the lowest to the highest standard deviation) over the 3 weeks (see Methods).

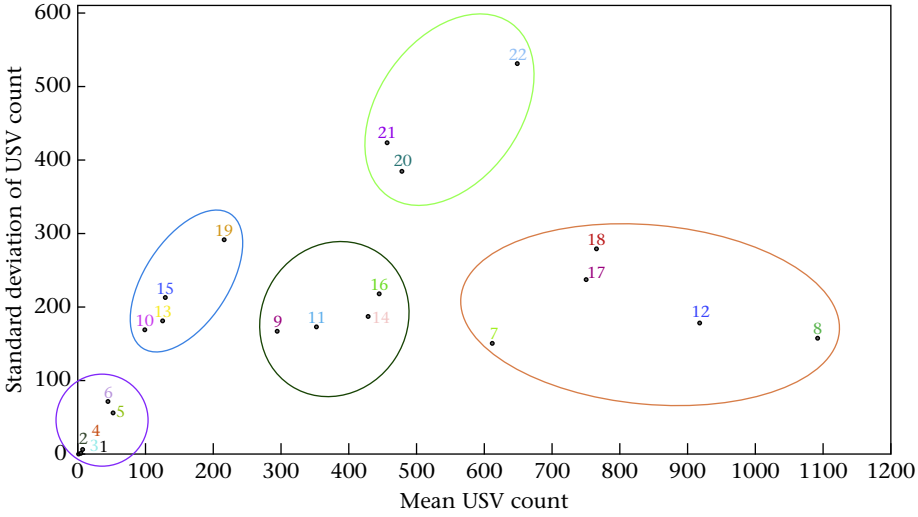

Figure 8. Relationship between the mean and variance in the number of USVs emitted over 3 weeks. Individuals clustered within ellipses: purple: consistent non and low vocalizers; light blue: inconsistent low-level vocalizers; dark green: consistent intermediate-level vocalizers; light green: inconsistent high-level vocalizers; orange: consistent high-level vocalizers. The same number and colour correspond to the same individual. Numbers are assigned according to individual variation in USV count (i.e. from the lowest to the highest standard deviation) over the 3 weeks (see Methods).

Spectrographic Features

Similar to the PIC models, the first machine learning model included mice that vocalized in both phase 1 and phase 2 during 10 min of stimulation in all 3 weeks ($N = 13$ and 23 parameters) and we correctly assigned 86% of recordings to the corresponding individuals (Fig. 10).

The grand mean frequency and mean frequency bandwidth, the ISI and the repertoire composition with the percentage of simple and complex USVs ('f' and 'c2') were selected in all iterations to predict individuals (Appendix Table A10). In the second model, run on data sets including separately phase 1 and phase 2 (6 data points/male and $N = 13$), mice were correctly identified by their recordings with 89% accuracy (24 parameters). The grand mean frequency, the mean frequency bandwidth, the mean USV length, and some syllable types with their percentages ('up', 'f', 'd', 'u', 'ui', 'c2' and 'c3') were chosen in 100% of the iterations (Appendix Table A10). The third model included all the mice that vocalized over the 3 weeks ($N = 19$) during at least one of the stimulation phases (i.e. 3 recordings/male). The random forest classifier did not perform as well as for the previous data sets, assigning 56% of recordings to the correct individuals (22 parameters). The grand mean frequency, the percentage of 'd' and 'f' were selected in all iterations (Appendix Table A10). In summary, the grand mean frequency and the percentage of 'f' appear to provide good candidates for individual signatures (chosen in all three models) together with the mean frequency bandwidth and the percentage of 'd' and 'c2' (chosen in two models) and the ISI (chosen in one model) in all iterations. We confirmed the results obtained with the RFE by performing univariate analyses on our candidates as individual signatures. We found that the variation in frequency parameters, temporal parameters and percentage of 'f' was significantly higher between than within individuals for both the set of 13 and the set of 19 sampled mice (see Appendix Tables A9 and A11). However, the percentage of 'c2' was only significantly different between individuals in the set of 19 males and we found only a trend when sampling 13 males. This discrepancy might suggest a higher between-individual variation in repertoire composition when including low and high vocalizers (19 mice) in the analyses, than when inspecting the repertoire composition in more vocal mice (13 mice) uttering USVs in both stimulation phases (see Appendix Figs A4 and A5; Tables A9 and A10). To test whether repertoire composition influenced the spectrotemporal features identified as individual signatures with the RFE, we computed correlations between percentages of syllable types and spectrotemporal features. We found that syllables with changes in frequency or with frequency jumps (i.e. percentage of 'u', 'c' and 'c3') were correlated with the mean USV length in all 3 weeks; however, in week 1 we found a trend for the percentage of 'c3' (Appendix Table A12). We did not find correlations between the grand mean frequency or the mean frequency bandwidth and the percentage of different syllable types ($N = 19$). Then, we tested whether mice showed high intra-individual consistency (high correlation between consecutive weeks for each parameter) for the potential individual signatures found with the RFE ($N = 19$ and 3 recordings/week/mouse). The grand mean frequency, the average frequency bandwidth, the percentage of 'd' and the mean USV length were highly correlated between all 3 weeks. The percentage of 'c2' was correlated between consecutive weeks (i.e. week 1 versus 2 and week 2 versus 3). However, the percentage of 'up' correlated only between weeks 2 and 3 and week 1 versus 3. The percentage of 'f' correlated only between weeks 1 and 3 and the ISI only between weeks 1 and 2 (Appendix Table A13). In summary, mice did not modulate their repertoire composition, frequency or temporal parameters during stimulation over the 3 weeks. Moreover, the relationship between 2 consecutive weeks for these parameters was not more similar than between nonconsecutive weeks as we previously reported for the USV count and the repertoire size. However, we detected the strongest intraindividual consistency, and thus correlation, between weeks 2 and 3 for most of the parameters (5/7 parameters; see Appendix Table A13).

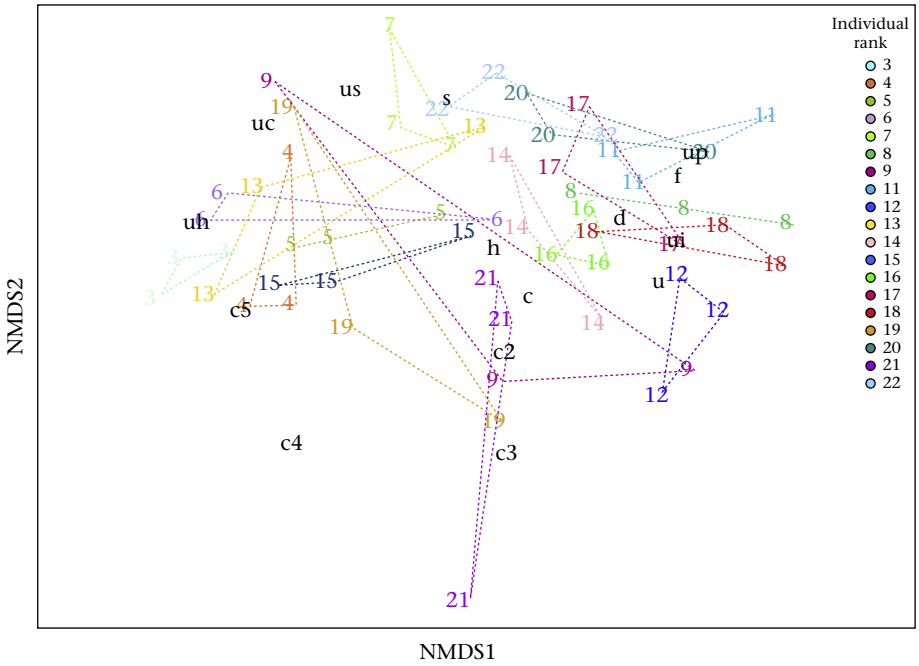

Figure 9. Individual clustering according to the repertoire composition using nonmetric multidimensional scaling (NMDS) for pairwise dissimilarity of the 15 syllable types (black letters). The same number and colour correspond to the same individual ($N = 19$ mice, 3 sampling points/male). Numbers correspond to males sorted according to their individual variation in USV count over the 3 weeks.

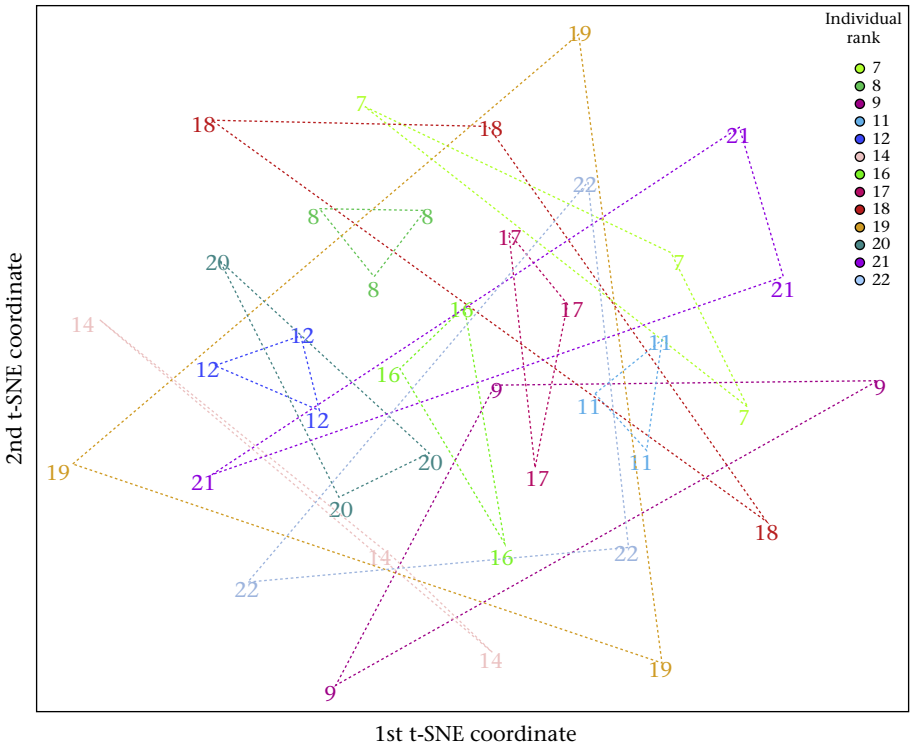

Figure 10. Individual visualization based on 23 parameters using t-distributed stochastic neighbour embedding (t-SNE) with Kullback–Leibler divergence loss of 0.49 (Kullback & Leibler, 1951). It includes mice that vocalized in both phases 1 and 2 during 10 min of stimulation in all 3 weeks ($N = 13$). The same number and colour correspond to the same individual (numbers correspond to males sorted according to their individual variation in USV count). Individuals that are more consistent have shorter distances for their three recordings.

DISCUSSION

Our results confirm high interindividual variation in the vocal responses of male mice to female urinary odour, and provide the first evidence that male mice increase the types of USVs, as well as the number of vocalizations, upon encountering female scent. We found individual consistency in USV emission over time and across trials on different weeks, and greater inter- than intraindividual variation in both the number (USV count) and types of vocalizations (repertoire size). Most males showed individually distinctive USV count and repertoire size, although some males emitted more distinctive vocalizations than others. Our results extend the findings of a previous study on wild-derived male house mice (Hoffmann et al., 2012a), by showing stability in individual vocalizations over 3 weeks. Below we address our main findings in more detail and the implications of our results.

Effects of Female Scent on Male Vocalizations

Our analysis of male vocalizations before and during the presentation of female urinary odour gave the following results. First, during the prestimulation phase, the males emitted very few, if any, USVs (e.g. in week 1, 10/22 of the males did not vocalize at all, and the rest on average emitted 2 ± 4 USVs and 1 ± 2 syllable types/5 min during this phase). Our results are consistent with another study on wild-derived mice (Musolf et al., 2010), and with some previous studies with laboratory strains, which found that males emitted few USVs without a social or sexual stimulus. For example, males emitted no USVs over 3 min in Gourbal et al.'s (2004) study, and over 10 min in Rieger and Dougherty's (2016) study. However, Chabout et al. (2015) found that males spontaneously emitted 26 USVs/5 min. These differences could be due to genetics or environmental conditions. A study on C57BL/6J mice found that males spontaneously emitted ca. 25 USVs (median) in a novel environment under dim light versus only ca. 1–10 USVs/5 min in other conditions (Mun, Lipina, & Roder, 2015). The mice in our study were also recorded in a novel environment and under dim light, and yet none of the unstimulated males vocalized at such high rates, at least in the first week. On the other hand, 12 of 22 males increased the number of different syllable types produced during the prestimulation phase over the 3 weeks. This change was not statistically significant, but it supports evidence that males increase spontaneous USV emission after previous experience in sociosexual interactions (Burke, Screven, & Dent, 2018). Thus, spontaneous USV emission might provide an index of a male's previous sexual interactions or current state of arousal, and domesticated male mice might vocalize when socially isolated because they are easily aroused (and because vocalizing might no longer have negative fitness consequences).

Second, after the presentation of female urine, half of the males began vocalizing within 1 min and they significantly increased both the number and types of USVs emitted (Fig. 2). In detail, 95% of the males started vocalizing within ca. 30 s after first sniffing the urine, whereas in only three of 57 of the recordings over 3 weeks did males vocalize before sniffing urine. Males showed an 89 times increase in the USV count and a seven times increase in the repertoire size during the stimulation period (mean in week 1: 178 ± 190 USVs and 7 ± 5 syllable types) compared to prestimulation. USV count and repertoire size were highly correlated before and during the stimulation, as previously shown (Zala et al., 2019). The males also increased the spectral complexity of their USVs upon sexual stimulation: before stimulus presentation males mainly emitted simple USVs (short duration and without frequency jumps), whereas during

odour stimulation, they emitted USVs with greater spectral complexity (long duration, more than one element, harmonic elements). During stimulation (first 5 min), 54% of the USVs emitted were 'up' syllables, whereas the other syllables were < 10% of the total USVs. Laboratory males emit many 'up' USVs just after the introduction of a female into a male's cage (C57BL/6Nc, Matsumoto & Okanoya, 2016), during direct interactions with a female without mounting (strains B6 and BTBR, Scattoni et al., 2011) and after a female is removed from the test arena (strains B6 and FVB, Yang, Loureiro, Kalikhman, & Crawley, 2013). These findings suggest that males emit a high proportion of 'up' USVs during courtship, but not during mounting or mating. Males might also emit 'up' syllables to attract a female's attention after she leaves the area (Yang et al., 2013). Future studies are needed with wild-derived mice to examine the temporal changes in syllable sequences used during courtship to test whether different syllable types have different functions.

Males in the present study were surprisingly more vocal than those in our previous studies, and in other studies on wild-derived mice, but not as vocal as laboratory mice. The mice in our study vocalized at five times higher rates (number of USVs/min) than those in a previous study using a thawed sample of frozen urine, and at 1.7 times higher rates than when the stimulus was fresh urine (Hoffmann et al., 2009, 2012b; Musolf et al., 2015). Males vocalized at three times higher rates in the present study than when stimulus females were presented to a male on the opposite side of a divider (Zala et al., 2017a; Nicolakis et al., 2020), and the males' USV counts are comparable to recordings of males directly interacting with a female (Nicolakis et al., 2020). These results could be due to genetics, as our mice originated from a different population than in these previous studies (Hoffmann et al., 2009, 2012b; Musolf et al., 2015), and wild populations of mice show differences in USV emission (Musolf et al., 2015). They might also be due to subtle differences in rearing or handling, even though our methods for rearing and handling animals were identical to these studies, and they were conducted in the same rooms. In contrast, studies on laboratory mice have found that males vocalize at even higher rates when presented with female urine (e.g. male B6D2F1/J mice presented with fresh urine emitted four times more USVs than we found (Chabout et al., 2015) and Yang et al. (2013) found males emitted six times more USVs during direct interaction with an oestrous female, i.e. phase 1 for strains B6 and FVB). It is misleading to compare our results with many previous USV studies on laboratory mice, however, because they often excluded nonvocalizing mice. If we had arbitrarily excluded the six lowest vocalizers in our study (27% of males), for example, then this would have resulted in a 1.4 times increase in the mean USVs emitted (see more below). Future studies should avoid or at least report the practice of selectively recording only the most vocal mice.

Our results suggest that male mice begin to alter the quantity and quality of USV emission as soon as they detect female scent, as well as during the latter stages of courtship and mating (Matsumoto & Okanoya, 2016). Our findings suggest that males emit USVs at a comparable rate and complexity when exposed to an odour versus during direct interactions with a female (Nicolakis et al., 2020). In contrast, a previous study on laboratory mice found that males emit more complex syllables, although at a similar rate, when exposed to female odour versus the presence of a female (Chabout et al., 2015). Male mice may emit USVs in response to detecting female scent to attract potential mates or to keep them in the vicinity, and subsequently increase the spectral complexity of vocalizations once they begin to interact (Nicolakis et al., 2020). Studies are needed to test whether it is the changes in male USV emission over time that attract and influence female mice.

Relationship Between USV Count and Repertoire Size

Mouse USVs are often classified into 3–15 syllable types, according to the absence or presence of frequency jumps and other spectrotemporal features (reviewed in Heckman et al., 2016; Ehret, 2018). However, only one other study, to our knowledge, has examined the relationship between USV number and repertoire size in mice (Zala et al., 2019). Both that study and ours showed that increasing the number of USVs uttered increases the number of spectrally different syllable types. If all the USV types have the same probability of occurring and only 15 vocalizations are sufficient to obtain the full vocal repertoire, the graphical representation should be a straight line with a perfect correlation between the USV count and the repertoire size. However, if USVs do not all have the same probability of being emitted, the relationship between USV count and repertoire size should be different. Our data showed that the USV count and repertoire size have a positive nonlinear, asymptotic relationship. Producing a high diversity of USVs requires producing many USVs (e.g. if a male only emits five USVs, it is not possible to emit more than five types). This inescapable constraint explains the lack of USVs to the left of the line and the region with >15 USV types (see Fig. 4). Repertoire size need not necessarily increase with increasing USV numbers (males could potentially emit a large number of calls using one type of USV), and yet numbers of USVs and types of USVs are positively correlated. It is unclear why males also produce a large USV repertoire when they emit a large number of USVs. It might be that repertoire size and number are both indicators of sexual motivation or of sexual attraction (often called arousal, see Nyby, 1983; Hanson & Hurley, 2012; Chabout et al., 2015; Matsumoto & Okanoya, 2016; Nicolakis et al., 2020). Individuals with larger repertoires might emit specific USV types with specific spectrotemporal features (e.g. higher mean length for USVs with multiple frequency jumps) than other individuals, but this hypothesis needs to be tested.

Interindividual Variation

Our results show that there is high interindividual variation in male USV emission, but mainly during sexual stimulation. Before sexual stimulation, males produced few if any USVs, as previously mentioned, and USV count and repertoire size showed little individual variation (even though the mean and variation in USVs and syllable types emitted during the prestimulation phase tended to increase over the 3 weeks). In contrast, during sexual stimulation, we found very high interindividual variation in USV count (e.g. in the first 5 min, males emitted 0–110, 0–125 and 0–118 USVs/min over weeks 1, 2 and 3, respectively). Thus, the average and variance in USV emission were higher in the stimulation than the prestimulation phase, which can be explained by the low USV count in the latter (i.e. a floor effect). Some males vocalized at low rates during the stimulation period (e.g. in week 1, six mice emitted 0–16 USVs). The repertoire size for these males was 9–12 syllable types, indicating that males can have a large repertoire size, despite uttering few USVs. We also found that some high vocalizers (5/22 males) over 3 weeks showed both a large USV count > 334 USVs/5 min (e.g. week 1: median = 107, interquartile range = 1–336/5 min) and a large repertoire size of 9–13 different syllable types (e.g. week 1: median = 9, interquartile range = 1–10/5 min). Two of 22 males did not vocalize at all in the stimulation period, as reported in previous studies. For example, in studies on laboratory mice during direct interactions with a female, one study excluded four of 24 males that were nonvocalizing (Scattoni et al., 2011), whereas another study excluded six of 16 males that emitted < 10 USVs (Matsumoto & Okanoya, 2016). Two comparable studies on wild-derived mice found that three of 15 males did not vocalize

during exposure to female odour (Hoffmann et al., 2012a), and another reported that two of 14 males did not vocalize during social interaction with females or males (von Merten et al., 2014). One study found that nonvocalizing individuals were more common among wild-derived than inbred laboratory mice, and that wild-derived laboratory mice from four different strains emitted < 10 USVs/15 min (Sugimoto et al., 2011). Studies are needed to investigate the possibility that nonvocalizing functions as an informative signal.

Intraindividual Consistency

During stimulus presentation, we found little consistency in nonvocalizing behaviour over time, whereas among vocalizers, low vocalizers tended to remain low and high vocalizers remained high vocalizers. We first investigated whether nonvocalizers, which did not vocalize in the first week, remained silent over the next 2 weeks. During the prestimulation phase, nearly half of the males did not vocalize at all (during any week), but only two of 22 individuals were consistent nonvocalizers over 3 weeks; most males instead alternated between vocalizing and nonvocalizing. During the stimulation period (first half), most males vocalized, but again only two of 22 individuals were consistent nonvocalizers (one male did not vocalize before or after stimulation). Instead, four of 22 males alternated between vocalizing and nonvocalizing during the 3 weeks, indicating that nonvocalizing is rare and consistent nonvocalizing is even rarer, contrary to what has been suggested (Hoffmann et al., 2012a; von Merten et al., 2014). To determine why some males were consistent nonvocalizers, we examined the rearing conditions of these 'silent types': one male was reared with only one sibling (a brother) and the other two males were reared in large litters (five pups) and sex ratios, neither of which is unusual. Two of the three nonvocalizing males had lower body mass (22–24 g) than the average for this population (grand mean body mass over 3 weeks: 26.8 ± 2.9 g), suggesting that nonvocalizers were in poorer condition than other mice. We observed no qualitative differences in the behaviour of these individuals during the experiments (although detailed quantitative behavioural analyses were not performed). Yet, we noted that one of the consistent and small nonvocalizing males was difficult to capture (i.e. high handling time) during our 3 weeks of recording, and he was very active in his home cage. The other nonvocalizing male spent an unusual amount of time self-grooming and was relatively immobile in his cage. Thus, these results suggest that male mice do not emit courtship vocalizations when they experience prenatal stress, or they are distressed or in poor health or condition (Lopes & König, 2016).

Among the males that vocalized, many consistently produced USVs over all 3 weeks (despite the presentation of scent from different females each week). We found no evidence that the males habituated to the presentation of novel female odour, or that they increased vocalizing rates during the stimulation period over the 3 weeks (unlike the prestimulation phase). During the prestimulation phase, only six of 22 males consistently vocalized over all 3 weeks, whereas, during sexual stimulation (first 5 min), 15 males vocalized during all 3 weeks (although two mice never vocalized at all and one vocalized only in the third week). Only one of the nonvocalizers was a consistent nonvocalizer for both the prestimulation phase and stimulation period and over the 3 weeks. The other four mice (18%) did not consistently vocalize during the three trials. Overall, six mice showed consistency in vocalizing before and during stimulus presentation and over the 3 weeks. Thus, during odour stimulation, most mice were vocalizers and remained vocalizers. Some mice were more consistent than others in their USV count and repertoire size, and most individuals did not change the rank order of USV emission over the 3 weeks (3/22 males increased and 5/22 decreased; Fig. 6b, d). Low vocalizers remained low and high vocalizers remained high vocalizers over time (e.g. in week 1, $r^2 = 0.422$ and see correlations in Fig. 6, Appendix Table A7). Individual consistency in USV emission was higher between 2 consecutive weeks, and still consistent over the 3 weeks. The repertoire size during the stimulation period was significantly correlated only in the second 5 min between weeks 2 and 3. This result might be driven by inconsistent males that did not vocalize over all 3 weeks. In fact, inconsistent males might use a similar repertoire size but with fewer USVs compared to the repertoire size of more consistent mice with a larger USV count. We also found that during the 10 min stimulation period, the USV count and the repertoire size were highly correlated in phases 1 and 2 in all 3 weeks (see Appendix Fig. A3, Table A8). Our findings provide support for a previous study on individual signatures in USVs emitted by wild-derived mice recorded during a single session (Hoffmann et al., 2012a).

We found high interindividual variation at each time point, and individual consistency over time, i.e. consistent individual variation, vocal individuality or vocal personality, which might reveal other personality traits (Friel, Kunc, Griffin, Asher, & Collins, 2016; Naguib, van Rooij, Snijders, & van Oers, 2016; Špinka, Syrová, Policht, & Linhart, 2018). We also found variation in consistency, i.e. rank order did not remain consistent. These results are not surprising, but studies might find consistency in rank order when mice are recorded in more natural conditions, and males become dominant or subordinate, for example.

Individual Signatures: Inter- versus Intraindividual Variation

We found significantly more variation in USV emission between than within individuals, and that some individuals were more distinctive in their USV count and repertoire size than others. When we compared the prestimulation phase with the first 5 min of the stimulation period, we found greater differences between than within individuals in male USV count and repertoire composition over the 3 weeks (Fig. 7). We found low, intermediate and high vocalizers, as well as consistent and inconsistent vocalizers for both the USV count and the repertoire size. Males seemed to cluster into these five different groups, separated depending upon their individual variation and consistency for both parameters (see Fig. 8). Less consistent males could be either low vocalizers with high variation in their repertoire size or high vocalizers with high variation in their USV count but not in repertoire size.

We assessed the strength of individual signatures using the PIC models. We found that all parameters provided individual signatures for models based on three recordings per mouse (models 2 and 4) and three to four of the six parameters provided potential individual signatures for models based on large data sets (models 1 and 3) that included all USVs (Appendix Table A9). In more detail, all parameters from model 1 and 11–16 of the 20 parameters from models 2 and 4 differed significantly between males (Appendix Table A9). In model 1 the frequency bandwidth, the ICI and the USV length were candidates but in model 3, only the other frequency parameters (grand mean frequency, mean minimum and maximum frequency) and the ICI were. For models 2 and 4, the percentage of syllable types with lower occurrence (USVs with more than one frequency modulation and multiple frequency jumps) and high-frequency USVs (>91 kHz) were candidates. Moreover, the ISI, the IBI and the latency to the first USV had a high PIC. Thus, these findings suggest a possible role of spectrotemporal features and repertoire composition as candidates for individual recognition, even if they were not completely repeatable using different models. Differences in parameter selection between

models might be due to the different sample sizes and different approaches to extract the parameters used for the analyses.

We found that males had consistent repertoire sizes and repertoire compositions. A consistent male for both the repertoire size and composition should utter the same number of different syllable types and the same proportion of USVs for each syllable type (e.g. a male always has three syllable types over the 3 weeks which are always 'up', 'd' and 'f'). However, mice could also be consistent in their repertoire size, but not in their repertoire composition (e.g. a male uses three syllable types in week 1 that are 'up', 'd' and 'f' but in week 2 he utters 'u', 'ui' and 's' and in week 3 'up', 'c2' and 'uh'). Thus, repertoire size and composition provide different information, but we still do not know the function of the different syllable types. Hence, we showed for the first time in wild-derived mice that some males had a distinct vocal repertoire composition compared to other mice, and that this trait was consistent over the 3 weeks (and across recording trials), thus providing a distinctive individual signature potentially used for individual recognition (see ANOSIM, Fig. 9). More vocal individuals were clustered according to spectrally simple USVs ('f', 'up', 'd', 'u' and 'ui') and short USVs (<10 ms USV length), whereas some of the less vocal individuals were grouped around less spectrally distinct or less common syllable types ('c4', 'c5', 'uh', 'uc' and 'us'; Fig. 9). This result is consistent with the results from the high PIC scores assigned to the rarer syllable types chosen as candidates for individual signatures.

The RFE supported our findings and selected seven different parameters from three different models: the grand mean frequency, the mean frequency bandwidth, the percentage of 'd', 'f' and 'c2', the mean USV length and the ISI. The accuracy score interval between models was from 56% (19 males sampled over 22 parameters) to 89% (13 males sampled over 24 percentage parameters) and the models selected similar parameters as candidates for acoustic signatures. We obtained the highest accuracy score for the RFE (89%) when we included in the model phases 1 and 2 separately. There are at least two explanations for this result, which are nonmutually exclusive: increasing the number of sampling points increased the accuracy score, or there might be consistency in the spectrotemporal parameters within the same recording.

In more detail, the grand mean frequency (66.57 ± 12.64 kHz) and the percentage of 'f' were the most important parameters for individual signatures, as they were chosen in all three models, the frequency bandwidth and the percentage of 'd' and 'c2' in two models and the mean USV length and the ISI in one model during the 10 iterations run by the RFE (Appendix Table A10). Thus, the RFE confirmed some candidates as individual signatures already found with the PIC, but not all of them were confirmed by both models (i.e. latency to the first USV, IBI and percentage of less common syllable types). Previous studies have investigated possible functions of these acoustic parameters. The mean frequency and duration (USV length) were parameters previously indicated as individual signatures in wild-derived house mice (Hoffmann et al., 2012a). The mean frequency is also repeatedly used by other species for individual discrimination (e.g. giggling vocalizations in hyaenas, *Crocuta crocuta*, Mathevon, Koralek, Weldele, Glickman, & Theunissen, 2010). In sac-winged bats, *Saccopteryx bilineata*, males are individually distinguished by two frequency parameters (the fundamental frequency and the end frequency of the trill), which are emitted only during courtship (Behr & von Helversen, 2004), and the fundamental frequency and a high song rate are correlated with reproductive success (Behr et al., 2006). In human voices, the fundamental frequency is paired with formants for individual recognition (Latinus & Belin, 2011; Xu, Homae, Hashimoto, & Hagiwara, 2013).

The frequency bandwidth has been reported to be used for individual and species recognition. For example, a study on laboratory mice, trained to recognize differences in USVs, showed that both males and females could distinguish USVs from artificial sweeps when the bandwidth was lower or higher than 60–80 kHz (35–80 kHz and 55–80 kHz) or the USV was longer than 200 ms (Screven & Dent, 2016). Another study reported that male mice modify their frequency bandwidth when another male is in the same cage and they are both provided with a female odour stimulus (Seagraves, Arthur, & Egnor, 2016). However, it is not known whether mice can discriminate between most of the syllable types, defined in our study or others, and whether they can perceive differences in the acoustic parameters in more spectrally similar vocalizations. In some species, however, a displacement in the frequency bandwidth can avoid the acoustic overlap between vocalizations from heterospecifics or conspecifics (e.g. birdsong in the forest or echolocation calls in bats; see Planqué & Slabbekoorn, 2008; Hase, Miyamoto, Kobayasi, & Hiryu, 2016).

We are unable to explain why the number and percentage of 'd' and 'f' were parameters chosen by our models as individual signatures, since more than 50% of the USV count included 'up' vocalizations. Some studies report that 'up' vocalizations are one of the most represented syllable types in the repertoire of mice and are emitted in different social contexts and behavioural conditions (Hanson & Hurley, 2012; Heckman et al., 2016; Hurley & Kalcounis-Rueppell, 2018; Matsumoto & Okanoya, 2018; Nicolakis et al., 2020; Sangiamo, Warren, & Neunuebel, 2020). Other studies have suggested that 'd' syllable types are involved in highly social interactions in juvenile B6 mice (Panksepp et al., 2007) or in pups separated from their mothers (Lahvis, Alleva, & Scattoni, 2011). In adult laboratory mice, males emit many 'f' and other syllable types ('up', 'u', 'ui') before mounting (Wang et al., 2008) and during direct interaction with a female (Yang et al., 2013) whereas they reduce the number of 'f' calls when the female is removed from the cage after directly interacting with the male (Hanson & Hurley, 2012; Yang et al., 2013). However, the RFE also selected the percentage of 'c2' used by males mainly during sexual contexts (as cited in Ehret, 2018) and by females during direct interactions (Scattoni et al., 2011; Hoier, Pfeifle, von Merten, & Linnenbrink, 2016; Matsumoto & Okanoya, 2018). Vocalizations with multiple frequency jumps might facilitate individual recognition. Trained laboratory mice can distinguish between vocalizations that are more spectrally distinct (e.g. USVs with frequency jumps from spectrally simpler USVs as 'f' or 'up', Neilans, Holfoth, Radziwon, Portfors, & Dent, 2014). In other studies, complex vocalizations have been suggested to increase male mating and reproductive success. Some studies, for example, reported that females are attracted to playbacks of spectrally complex USVs (Chabout et al., 2015). Opposite-sex pairs of mice emitted more 'up' and 'one-jump' syllables in the early phase of courtship (and long syllables with multiple jumps were rare), whereas they emitted more 'harmonics' after 15 min, the phase during which more mountings were observed (Matsumoto & Okanoya, 2016). Studies on bird vocalizations found that individuals can use different syllable types for individual recognition (Elie & Theunissen, 2018) and a higher spectral complexity in vocalizations relates to higher quality and reproductive success for males (Catchpole, 1987; Ballentine, 2004; Garamszegi, 2005).

The ISI, the silence between syllables, and the IBI, the silence between isolated vocalizations or sequences of vocalizations (Sueur, 2002; Jain, Diwakar, Bahuleyan, Deb, & Balakrishnan, 2013), might provide information about the structure and rhythm of the USV emission (Ravignani et al., 2019). We predicted that mice emitting few USVs might show higher variability in both parameters than more vocal males, since more USVs are often grouped within a bout and low vocalizers produced few long bouts in their

recordings. A study on laboratory mice suggested that the IBI discriminates more vocal from less vocal males and adult males from pups (Liu, Miller, Merzenich, & Schreiner, 2003). As expected, we found that males emitting fewer USVs had more variable ISIs and IBIs than other males. Previous studies on other species showed that ISI and IBI carry information about individual identity (cited in Kershenbaum et al., 2016) and are involved in turn-taking communication between conspecifics (Pika, Wilkinson, Kendrick, & Vernes, 2018). Together with other time intervals, the ISI is also reported in studies on mice and other species to determine rhythm in vocalizations (Castellucci et al., 2018; Ravignani et al., 2019) and it appears that the intervals between vocalizations are determined by specific breathing cycles with sequences of vocalizations separated by silences of more than 500 ms in mice (Castellucci et al., 2018).

We also found that the latency to emit the first USV and the most common syllable types, with both simple and complex spectral parameters, were often chosen in more than 50% of the trials by the RFE suggesting that they might be additional candidates for individual signatures as also detected with the PIC. The latency to the first vocalization might be an additional indicator of sexual motivation or arousal. The initiation of male vocalizations is triggered by female stimulation, at least in wild, outbred mice, and as previously mentioned, high variation in male responsiveness might be due to differences in sexual motivation. However, we have no evidence to support these speculations (see Zala et al., 2019). Since the number of data points we had was similar to the number of variables that we analysed, we need more recordings to confirm our findings. More data will improve our ability to identify syllable types that are most relevant for individual recognition.

We did not find a correlation between the frequency parameters and the percentage of syllable types suggesting that the repertoire composition does not influence the frequency parameters in our recordings. However, syllables with one or more changes in frequency or with frequency jumps (i.e. percentage of 'u', 'c' and 'c3') correlated with the mean USV length, suggesting that more complex vocalizations are also longer than simple vocalizations. We also found that five of seven individual signatures showed a strong correlation between weeks 2 and 3 and that four of seven individual signatures correlated in all 3 weeks. In more detail, the frequency and temporal parameters (i.e. mean USV length) and the percentage of 'd' correlated between the 3 weeks. Thus, males did not change their acoustic parameters over time, but they might have slightly modulated their repertoire composition and still showed some intraindividual consistency (Appendix Table A13).

Our results raise several questions. First, we showed that male courtship USVs potentially mediate individual recognition in house mice, but further studies are needed to test this hypothesis and to test whether male vocalizations or other secondary sexual traits mediate individual recognition in other species. In mice, USVs might mediate individual recognition in social contexts other than courtship, such as parent–offspring interactions. For our study, we arbitrarily chose 3 weeks as a sampling time, which is sufficient to mediate social and sexual interactions involving individual recognition in house mice in different social contexts: territory formation by males, courtship and mating (over one to three oestrous cycles), rearing and weaning offspring are all important social interactions known to be critical for fitness (e.g. inter- and intrasexual interactions, parent–offspring recognition, see Musolf & Penn, 2012; Asaba, Hattori, Mogi, & Kikusui, 2014; Hoier et al., 2016; Varholick, Bailoo, Palme, & Wurbel, 2018). Second, our results raise questions about the proximate mechanisms controlling individual variation and consistency in USV emission (e.g. laryngeal morphology Riede, Borgard, & Pasch, 2017), neuroendocrine control (Asaba et al., 2014; Mills, Dhillo, & Comninos, 2018) and their heritability (Spence, Aslam, Hofer, Brunelli, & Shair, 2016). Third, our results provide four spectrotemporal features and three candidate vocalizations to explain individual signatures; however, studies are needed to investigate other parameters, and especially sequences of syllables or 'syntax', which vary between geographical populations of house mice (von Merten et al., 2014). The order of USVs within a sequence might be random or not, and with a recording of 100 USVs and 15 syllable types, we obtain 3^{17} possible combinations of sequences ignoring sequential order, and if we consider the sequential order of the same USVs, we obtain 4^{117} possible permutations. A previous study with laboratory mice found a similar temporal order of syllable types within a song (USV sequence with ISI < 250 ms) in the presence of female odour and during interaction with a female, suggesting that songs in the presence of a female or her scent might produce similar signals (Chabout et al., 2015). Fourth, studies are needed to test whether USVs together with odour (or other cues) provide more reliable individual signatures than either alone. Many species use multimodal signalling during courtship (Mitoyen, Quigley, & Fusani, 2019), and if multiple sensory modalities provide particularly robust individual signatures (as suggested for biometric recognition in humans; Jain, 2005; Kumar & Tiwari, 2019), then they should facilitate individual recognition (Kulahci, Drea, Rubenstein, & Ghazanfar, 2014). Finally, our results suggest that USVs are potentially useful for researchers aiming to identify individual mice during social interactions (Warren et al., 2018), i.e. individual voice recognition (Latinus & Belin, 2011; Xu et al., 2013; Budka, Wojas, & Osiejuk, 2015); however, studies are needed to investigate the stability of individual signatures in the USVs of mice over longer periods of time and across different social contexts.

Conclusions

Our results provide the following main novel findings. (1) Most wild-derived male mice greatly increased the types of syllables, as well as the numbers of USVs that they emitted, upon encountering female urinary scent. (2) There was high variation in males' propensity to vocalize or not, as some males vocalized very little, if at all, upon encountering female odour (Hoffmann et al., 2012b), and we found that nonvocalizing was rarely a consistent trait, contrary to what has been suggested (Hoffmann et al., 2012a; von Merten et al., 2014). It is unclear why males sometimes remain silent, but our results suggest that the lack of vocal behaviour and responsiveness to female odour might be due to poor health or condition. (3) A striking result from our study and other studies on wild mice, regardless of the sex, stimulus or social context, is the high interindividual variation in USV count (Hoffmann et al., 2012a; von Merten et al., 2014; Musolf et al., 2015; Zala et al., 2017a), and among vocalizers, we found that there was high variation in the number and types of vocalizations emitted. (4) We found individual consistency in vocalizing rate and syllable types, i.e. most low vocalizers remained low using a smaller vocal repertoire, and high vocalizers usually remained high with a larger vocal repertoire. (5) Finally, we showed that the vocalizations of wild-derived males contain individual signatures over 3 weeks. We found greater interthan intraindividual variation in USV number and types, and most of the spectrotemporal features we analysed provide candidates for individual recognition. The vocalizations of individual males could be clustered according to their USV count and repertoire size and composition, and their spectrotemporal features showed individual signatures using different analytical approaches. Studies are needed to test whether USVs of house mice are used for individual

and mate recognition, and whether individual signatures are long lasting over time and across different contexts.

Author Contributions

M.A.M., D.N., S.M.Z. and D.J.P. designed the experiment. M.A.M. and D.N. conducted the experiments and M.A.M., D.N. and R.A. processed and analysed the data. M.A.M., D.J.P. and S.M.Z. wrote the paper, and all authors reviewed and approved the final manuscript.

Data Availability

All data generated or analysed during this study are included in this article and its supplementary material files.

Declaration of Interest

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by a grant (FWF P 28141-B25) of the Austrian Science Fund (http://www.fwf.ac.at). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank U. Madlsperger and H. Sasse for animal care; T. Klaus, E. Straßer, B. Wernisch, S. Bartsch, M. Grijincu for their help during data collection and processing; P. Balazs and A. Noll for technical support with the STx program; P. Linhart and two anonymous referees whose insightful comments greatly improved the manuscript; and R. Hengsberger for her help with formatting figures and the manuscript.

Supplementary Material

Supplementary material associated with this article is available at https://doi.org/10.1016/j.anbehav.2020.09.006.

References

Asaba, A., Hattori, T., Mogi, K., & Kikusui, T. (2014). Sexual attractiveness of male chemicals and vocalizations in mice. *Frontiers in Neuroscience, 8*, 231. https://doi.org/10.3389/fnins.2014.00231

Asaba, A., Osakada, T., Touhara, K., Kato, M., Mogi, K., & Kikusui, T. (2017). Male mice ultrasonic vocalizations enhance female sexual approach and hypothalamic kisspeptin neuron activity. *Hormones and Behavior, 94*, 53–60. https://doi.org/10.1016/j.yhbeh.2017.06.006

Ballentine, B. (2004). Vocal performance influences female response to male bird song: An experimental test. *Behavioral Ecology, 15*, 163–168. https://doi.org/10.1093/beheco/arg090

Behr, O., & von Helversen, O. (2004). Bat serenades—complex courtship songs of the sac-winged bat (*Saccopteryx bilineata*). *Behavioral Ecology and Sociobiology, 56*, 106–115. https://doi.org/10.1007/s00265-004-0768-7

Behr, O., von Helversen, O., Heckel, G., Nagy, M., Voigt, C. C., & Mayer, F. (2006). Territorial songs indicate male quality in the sac-winged bat *Saccopteryx bilineata* (Chiroptera, Emballonuridae). *Behavioral Ecology, 17*, 810–817. https://doi.org/10.1093/beheco/arl013

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Budka, M., Wojas, L., & Osiejuk, T. S. (2015). Is it possible to acoustically identify individuals within a population? *Journal of Ornithology, 156*, 481–488. https://doi.org/10.1007/s10336-014-1149-2

Burke, K., Screven, L. A., & Dent, M. L. (2018). CBA/CaJ mouse ultrasonic vocalizations depend on prior social experience. *PloS One, 13*, Article e0197774. https://doi.org/10.1371/journal.pone.0197774

Byers, S. L., Wiles, M. V., Dunn, S. L., & Taft, R. A. (2012). Mouse estrous cycle identification tool and Images. *PloS One, 7*, Article e35538. https://doi.org/10.1371/journal.pone.0035538

Castellucci, G. A., Calbick, D., & McCormick, D. (2018). The temporal organization of mouse ultrasonic vocalizations. *PloS One, 13*, Article e0199929. https://doi.org/10.1371/journal.pone.0199929

Castellucci, G. A., McGinley, M. J., & McCormick, D. A. (2016). Knockout of *Foxp2* disrupts vocal development in mice. *Scientific Reports, 6*, 23305. https://doi.org/10.1038/srep23305

Catchpole, C. K. (1987). Bird song, sexual selection and female choice. *Trends in Ecology & Evolution, 2*, 94–97. https://doi.org/10.1016/0169-5347(87)90165-0

Chabout, J., Sarkar, A., Dunson, D. B., & Jarvis, E. D. (2015). Male mice song syntax depends on social contexts and influences female preferences. *Frontiers in Behavioral Neuroscience, 9*, 76. https://doi.org/10.3389/fnbh.2015.00076

Clarke, K. R., & Warwick, R. M. (2001). *Change in marine communities: An approach to statistical analysis and interpretation* (2nd ed.). Primer-E Ltd.

Donaldson, J. (2016). *tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE)*. R Package Version 0.1-3 https://github.com/jdonaldson/rtsne/.

Ehret, G. (2018). Characteristics of vocalization in adult mice. In S. M. Brudzynski (Ed.), *Handbook of behavioral neuroscience* (pp. 187–195). Elsevier.

Elie, J. E., & Theunissen, F. E. (2018). Zebra finches identify individuals using vocal signatures unique to each call type. *Nature Communications, 9*, 4026. https://doi.org/10.1038/s41467-018-06394-9

Favaro, L., Gamba, M., Alfieri, C., Pessani, D., & McElligott, A. G. (2015). Vocal individuality cues in the African penguin (*Spheniscus demersus*): A source-filter theory approach. *Scientific Reports, 5*, 17255. https://doi.org/10.1038/srep17255

Friel, M., Kunc, H. P., Griffin, K., Asher, L., & Collins, L. M. (2016). Acoustic signalling reflects personality in a social mammal. *Royal Society Open Science, 3*, 160178. https://doi.org/10.1098/rsos.160178

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *American Statistician, 56*, 316–324. https://doi.org/10.1198/000313002533

Garamszegi, L. Z. (2005). Bird song and parasites. *Behavioral Ecology and Sociobiology, 59*, 167–180. https://doi.org/10.1007/s00265-005-0041-8

Gourbal, B. E. F., Barthelemy, M., Petit, G., & Gabrion, C. (2004). Spectrographic analysis of the ultrasonic vocalisations of adult male and female BALB/c mice. *Naturwissenschaften, 91*, 381–385. https://doi.org/10.1007/s00114-004-0543-7

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*, 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Grimsley, J. M. S., Monaghan, J. J. M., & Wenstrup, J. J. (2011). Development of social vocalizations in mice. *PloS One, 6*, Article e17460. https://doi.org/10.1371/journal.pone.0017460

Hammerschmidt, K., Radyushkin, K., Ehrenreich, H., & Fischer, J. (2009). Female mice respond to male ultrasonic 'songs' with approach behaviour. *Biology Letters, 5*, 589–592. https://doi.org/10.1098/rsbl.2009.0317

Hanson, J. L., & Hurley, L. M. (2012). Female presence and estrous state influence mouse ultrasonic courtship vocalizations. *PloS One, 7*, Article e40782. https://doi.org/10.1371/journal.pone.0040782

Hase, K., Miyamoto, T., Kobayasi, K. I., & Hiryu, S. (2016). Rapid frequency control of sonar sounds by the FM bat, *Miniopterus fuliginosus*, in response to spectral overlap. *Behavioural Processes, 128*, 126–133. https://doi.org/10.1016/j.beproc.2016.04.017

Heckman, J., McGuinness, B., Celikel, T., & Englitz, B. (2016). Determinants of the mouse ultrasonic vocal structure and repertoire. *Neuroscience and Biobehavioral Reviews, 65*, 313–325. https://doi.org/10.1016/j.neubiorev.2016.03.029

Heckman, J. J., Proville, R., Heckman, G. J., Azarfar, A., Celikel, T., & Englitz, B. (2017). High-precision spatial localization of mouse vocalizations during social interaction. *Scientific Reports, 7*, 3017. https://doi.org/10.1038/s41598-017-02954-z

Hoffmann, F., Musolf, K., & Penn, D. J. (2009). Freezing urine reduces its efficacy for eliciting ultrasonic vocalizations from male mice. *Physiology and Behavior, 96*, 602–605. https://doi.org/10.1016/j.physbeh.2008.12.014

Hoffmann, F., Musolf, K., & Penn, D. J. (2012a). Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. *Physiology and Behavior, 105*, 766–771. https://doi.org/10.1016/j.physbeh.2011.10.011

Hoffmann, F., Musolf, K., & Penn, D. J. (2012b). Ultrasonic courtship vocalizations in wild house mice: Spectrographic analyses. *Journal of Ethology, 30*, 173–180. https://doi.org/10.1007/s10164-011-0312-y

Hoier, S., Pfeifle, C., von Merten, S., & Linnenbrink, M. (2016). Communication at the Garden Fence—context dependent vocalization in female house mice. *PloS One, 11*, Article e0152255. https://doi.org/10.1371/journal.pone.0152255

Holy, T. E., & Guo, Z. (2005). Ultrasonic songs of male mice. *PLoS Biology, 3*, e386. https://doi.org/10.1371/journal.pbio.0030386

Hurley, L. M., & Kalcounis-Rueppell, M. C. (2018). State and context in vocal communication of rodents. In M. L. Dent, R. R. Fay, & A. N. Popper (Eds.), *Rodent bioacoustics* (pp. 191–221). Springer International.

Jain, A. K. (2005). Biometric recognition: How do I know Who You are? In F. Roli, & S. Vitulano (Eds.), *Image analysis and processing* (pp. 19–26). Springer.

Jain, M., Diwakar, S., Bahuleyan, J., Deb, R., & Balakrishnan, R. (2013). A rain forest dusk chorus: Cacophony or sounds of silence? *Evolutionary Ecology, 28*, 1–22. https://doi.org/10.1007/s10682-013-9658-7

Jornod, M., & Roche, D. G. (2015). Inter- vs intra-individual variation and temporal repeatability of escape responses in the coral reef fish *Amblyglyphidodon curacao*. *Biology Open, 4*, 1395–1399. https://doi.org/10.1242/bio.013508

Kanno, K., & Kikusui, T. (2018). Effect of sociosexual experience and aging on number of courtship ultrasonic vocalizations in male mice. *Zoological Science, 35*, 208–214. https://doi.org/10.2108/zs170175

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., et al. (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews, 91*, 13–52. https://doi.org/10.1111/brv.12160

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. (2018). . In *caret: Classification and regression training*. R Package Version 6.0-79 https://CRAN.R-project.org/package=caret.

Kulahci, I. G., Drea, C. M., Rubenstein, D. I., & Ghazanfar, A. A. (2014). Individual recognition through olfactory-auditory matching in lemurs. *Proceedings of the Royal Society B, 281*, 20140071. https://doi.org/10.1098/rspb.2014.0071

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86. www.jstor.org/stable/2236703.

Kumar, M., & Tiwari, A. K. (2019). Computational intelligence techniques for biometric recognition: A review. In *Proceedings of the 2nd international conference on advanced computing and software engineering (ICACSE-2019)* (pp. 230–236).

Lahvis, G. P., Alleva, E., & Scattoni, M. L. (2011). Translating mouse vocalizations: prosody and frequency modulation. *Genes, Brain, and Behavior, 10*(1), 4–16. https://doi.org/10.1111/j.1601-183X.2010.00603.x

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*, R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*, 18–22. http://CRAN.R-project.org/doc/Rnews/.

Linhart, P., Osiejuk, T. S., Budka, M., Šálek, M., Špinka, M., Policht, R., et al. (2019). Measuring individual identity information in animal signals: Overview and performance of available identity metrics. *Methods in Ecology and Evolution, 10*, 1558–1570. https://doi.org/10.1111/2041-210x.13238

Liu, R. C., Miller, K. D., Merzenich, M. M., & Schreiner, C. E. (2003). Acoustic variability and distinguishability among mouse ultrasound vocalizations. *Journal of the Acoustical Society of America, 114*, 3412–3422. https://doi.org/10.1121/1.1623787

Lopes, P. C., & König, B. (2016). Choosing a healthy mate: Sexually attractive traits as reliable indicators of current disease status in house mice. *Animal Behaviour, 111*, 119–126. https://doi.org/10.1016/j.anbehav.2015.10.011

van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the 12th international conference on artificial intelligence and statistics (AISTATS) 2009* (Vol. 5, pp. 384–391). Clearwater Beach, Florida, USA.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Mahrt, E. J., Perkel, D. J., Tong, L., Rubel, E. W., & Portfors, C. V. (2013). Engineered deafness reveals that mouse courtship vocalizations do not require auditory experience. *Journal of Neuroscience, 33*, 5573–5583. https://doi.org/10.1523/Jneurosci.5054-12.2013

Mathevon, N., Koralek, A., Weldele, M., Glickman, S. E., & Theunissen, F. E. (2010). What the hyena's laugh tells: Sex, age, dominance and individual signature in the giggling call of *Crocuta crocuta*. *BMC Ecology, 10*, 9. https://doi.org/10.1186/1472-6785-10-9

Matsumoto, Y. K., & Okanoya, K. (2016). Phase-specific vocalizations of male mice at the initial encounter during the courtship sequence. *PloS One, 11*, Article e0147102. https://doi.org/10.1371/journal.pone.0147102

Matsumoto, Y. K., & Okanoya, K. (2018). Mice modulate ultrasonic calling bouts according to sociosexual context. *Royal Society Open Science, 5*. https://doi.org/10.1098/rsos.180378

von Merten, S., Hoier, S., Pfeifle, C., & Tautz, D. (2014). A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (*Mus musculus domesticus*). *PloS One, 9*, Article e97244. https://doi.org/10.1371/journal.pone.0097244

Mills, E. G. A., Dhillo, W. S., & Comninos, A. N. (2018). Kisspeptin and the control of emotions, mood and reproductive behaviour. *Journal of Endocrinology, 239*, R1–R12. https://doi.org/10.1530/JOE-18-0269

Mitoyen, C., Quigley, C., & Fusani, L. (2019). Evolution and function of multimodal courtship displays. *Ethology, 1215*, 503–515. https://doi.org/10.1111/eth.12882

Mun, H.-S., Lipina, T. V., & Roder, J. C. (2015). Ultrasonic vocalizations in mice during exploratory behavior are context-dependent. *Frontiers in Behavioral Neuroscience, 9*, 316. https://doi.org/10.3389/fnbeh.2015.00316

Murdoch, D. J., & Chow, E. D. (1996). A graphical display of large correlation matrices. *American Statistician, 50*, 178–180. https://doi.org/10.1080/00031305.1996.10474371

Musolf, K., Hoffmann, F., & Penn, D. J. (2010). Ultrasonic courtship vocalizations in wild house mice, *Mus musculus*. *Animal Behaviour, 79*, 757–764. https://doi.org/10.1016/j.anbehav.2009.12.034

Musolf, K., Meindl, S., Larsen, A. L., Kalcounis-Rueppell, M. C., & Penn, D. J. (2015). Ultrasonic vocalizations of male mice differ among species and females show assortative preferences for male calls. *PloS One, 10*, Article e0134123. https://doi.org/10.1371/journal.pone.0134123

Musolf, K., & Penn, D. J. (2012). Ultrasonic vocalizations in house mice: A cryptic mode of acoustic communication. In M. Macholán, S. J. E. Baird, P. Munclinger, & J. Piálek (Eds.), *Evolution of the house mouse* (pp. 253–277). Cambridge University Press.

Naguib, M., van Rooij, E. P., Snijders, L., & van Oers, K. (2016). To sing or not to sing: Seasonal changes in singing vary with personality in wild great tits. *Behavioral Ecology, 27*, 932–938. https://doi.org/10.1093/beheco/arv235

Neilans, E. G., Holfoth, D. P., Radziwon, K. E., Portfors, C. V., & Dent, M. L. (2014). Discrimination of ultrasonic vocalizations by CBA/CaJ mice (*Mus musculus*) is related to spectrotemporal dissimilarity of vocalizations. *PloS One, 9*, Article e85405. https://doi.org/10.1371/journal.pone.0085405

Neunuebel, J. P., Taylor, A. L., Arthur, B. J., & Egnor, S. E. R. (2015). Female mice ultrasonically interact with males during courtship displays. *eLIFE, 4*, Article e06203. https://doi.org/10.7554/eLife.06203

Nicolakis, D., Marconi, M. A., Zala, S. M., & Penn, D. J. (2020). Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success. *Frontiers in Zoology, 17*, 10. https://doi.org/10.1186/s12983-020-00353-1

Nomoto, K., Ikumi, M., Otsuka, M., Asaba, A., Kato, M., Koshida, N., et al. (2018). Female mice exhibit both sexual and social partner preferences for vocalizing males. *Integrative Zoology, 13*, 735–744. https://doi.org/10.1111/1749-4877.12357

Nyby, J. (1983). Ultrasonic vocalizations during sex behavior of male house mice (*Mus musculus*): A description. *Behavioral and Neural Biology, 39*, 128–134. https://doi.org/10.1016/S0163-1047(83)90722-7

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018). *vegan: Community ecology package*. R package version 2.5-2 http://CRAN.R-project.org/package=vegan.

Panksepp, J. B., Jochman, K. A., Kim, J. U., Koy, J. J., Wilson, E. D., Chen, Q., et al. (2007). Affiliative behavior, ultrasonic communication and social reward are influenced by genetic variation in adolescent mice. *PloS One, 2*, e351. https://doi.org/10.1371/journal.pone.0000351

Parma, V., Gordon, A. R., Cecchetto, C., Cavazzana, A., Lundström, J. N., & Olsson, M. J. (2017). Processing of human body odors. In A. Buettner (Ed.), *Springer handbook of odor* (pp. 963–986). Springer.

Penn, D. J., Oberzaucher, E., Grammer, K., Fischer, G., Soini, H. A., Wiesler, D., et al. (2007). Individual and gender fingerprints in human body odour. *Journal of The Royal Society Interface, 4*, 331–340. https://doi.org/10.1098/rsif.2006.0182

Petrusková, T., Pišvejcová, I., Kinštová, A., Brinke, T., Petrusek, A., & Börger, L. (2016). Repertoire-based individual acoustic monitoring of a migratory passerine bird with complex song as an efficient tool for tracking territorial dynamics and annual return rates. *Methods in Ecology and Evolution, 7*, 274–284. https://doi.org/10.1111/2041-210x.12496

Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B, 285*, 20180598. https://doi.org/10.1098/rspb.2018.0598

Planqué, R., & Slabbekoorn, H. (2008). Spectral overlap in songs and temporal avoidance in a Peruvian bird assemblage. *Ethology, 114*, 262–271. https://doi.org/10.1111/j.1439-0310.2007.01461.x

Pomerantz, S. M., Nunez, A. A., & Bean, N. J. (1983). Female behavior is affected by male ultrasonic vocalizations in house mice. *Physiology and Behavior, 31*, 91–96. https://doi.org/10.1016/0031-9384(83)90101-4

R Core Team. (2018). *R: A language and environment for statistical computing*. R. Foundation for Statistical Computing. http://www.R-project.org/.

Ravignani, A., Dalla Bella, S., Falk, S., Kello, C. T., Noriega, F., & Kotz, S. A. (2019). Rhythm in speech and animal vocalizations: A cross-species perspective. *Annals of the New York Academy of Sciences, 1453*, 79–98. https://doi.org/10.1111/nyas.14166

Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour, 78*, 233–240. https://doi.org/10.1016/j.anbehav.2009.06.007

Riede, T., Borgard, H. L., & Pasch, B. (2017). Laryngeal airway reconstruction indicates that rodent ultrasonic vocalizations are produced by an edge-tone mechanism. *Royal Society Open Science, 4*, 170976. https://doi.org/10.1098/rsos.170976

Rieger, M. A., & Dougherty, J. D. (2016). Analysis of within subjects variability in mouse ultrasonic vocalization: Pups exhibit inconsistent, state-like patterns of call production. *Frontiers in Behavioral Neuroscience, 10*, 182. https://doi.org/10.3389/fnbeh.2016.00182

Sangiamo, D. T., Warren, M. R., & Neunuebel, J. P. (2020). Ultrasonic signals associated with different types of social behavior of mice. *Nature Neuroscience*. https://doi.org/10.1038/s41593-020-0584-z

Scattoni, M. L., Gandhy, S. U., Ricceri, L., & Crawley, J. N. (2008). Unusual repertoire of vocalizations in the BTBR T+tf/J mouse model of autism. *PloS One, 3*, Article e3067. https://doi.org/10.1371/journal.pone.0003067

Scattoni, M. L., Ricceri, L., & Crawley, J. N. (2011). Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. *Genes, Brain and Behavior, 10*, 44–56. https://doi.org/10.1111/j.1601-183X.2010.00623.x

Screven, L. A., & Dent, M. L. (2016). Discrimination of frequency modulated sweeps by mice. *Journal of the Acoustical Society of America, 140*, 1481–1487. https://doi.org/10.1121/1.4962223

Seagraves, K. M., Arthur, B. J., & Egnor, S. E. R. (2016). Evidence for an audience effect in mice: Male social partners alter the male vocal response to female cues. *Journal of Experimental Biology, 219*, 1437–1448. https://doi.org/10.1242/jeb.129361

Shapiro, A. D. (2010). Recognition of individuals within the social group: Signature vocalizations. In S. M. Brudzynski (Ed.), *Handbook of mammalian vocalization: An integrative neuroscience approach* (pp. 495–503). Academic Press.

Spence, H. R., Aslam, A. M., Hofer, M. A., Brunelli, S. A., & Shair, H. N. (2016). Vocal coselection in rat pup ultrasonic vocalizations. *Ecology and Evolution, 6*, 1922–1929. https://doi.org/10.1002/ece3.1907

Špinka, M., Syrová, M., Policht, R., & Linhart, P. (2018). Individual stability in vocalization rates of preweaning piglets. *Journal of Animal Science, 96*, 11–16. https://doi.org/10.1093/jas/skx014

Stowell, D., Petrusková, T., Šálek, M., & Linhart, P. (2019). Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *Journal of The Royal Society Interface, 16*, 20180940. https://doi.org/10.1098/rsif.2018.0940

Sueur, J. (2002). Cicada acoustic communication: Potential sound partitioning in a multispecies community from Mexico (Hemiptera: Cicadomorpha: Cicadidae). *Biological Journal of the Linnean Society, 75*, 379–394.

Sugimoto, H., Okabe, S., Kato, M., Koshida, N., Shiroishi, T., Mogi, K., et al. (2011). A role for strain differences in waveforms of ultrasonic vocalizations during male-female interaction. *PloS One, 6*, Article e22093. https://doi.org/10.1371/journal.pone.0022093

Tibbetts, E. A., & Dale, J. (2007). Individual recognition: It is good to be different. *Trends in Ecology & Evolution, 22*, 529–537. https://doi.org/10.1016/j.tree.2007.09.001

Varholick, J. A., Bailoo, J. D., Palme, R., & Wurbel, H. (2018). Phenotypic variability between social dominance ranks in laboratory mice. *Scientific Reports, 8*, 6593. https://doi.org/10.1038/s41598-018-24624-4

Wang, H., Liang, S., Burgdorf, J., Wess, J., & Yeomans, J. (2008). Ultrasonic vocalizations induced by sex and amphetamine in M2, M4, M5 muscarinic and D2 dopamine receptor knockout mice. *PloS One, 3*, Article e1893. https://doi.org/10.1371/journal.pone.0001893

Warren, M. R., Spurrier, M. S., Roth, E. D., & Neunuebel, J. P. (2018). Sex differences in vocal communication of freely interacting adult mice depend upon behavioral context. *PloS One, 13*, Article e0204527. https://doi.org/10.1371/journal.pone.0204527

Wei, T., & Simko, V. (2017). *corrplot: Visualization of a correlation matrix*. R Package Version 0.84 https://github.com/taiyun/corrplot.

Wickham, B. (2006). An implementation of the grammar of graphics in R: ggplot. In *Book of abstracts, Second International R User Conference June 15-17 2006, Vienna, Austria* (p. 183).

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*, 1–20. https://doi.org/10.18637/jss.v021.i12

Wiley, R. H. (2013). Specificity and multiplicity in the recognition of individuals: Implications for the evolution of social behaviour. *Biological Reviews, 88*, 179–195. https://doi.org/10.1111/j.1469-185X.2012.00246.x

Xu, M., Homae, F., Hashimoto, R., & Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Frontiers in Psychology, 4*, 735. https://doi.org/10.3389/fpsyg.2013.00735

Yang, M., Loureiro, D., Kalikhman, D., & Crawley, J. N. (2013). Male mice emit distinct ultrasonic vocalizations when the female leaves the social interaction arena. *Frontiers in Behavioral Neuroscience, 7*, 159. https://doi.org/10.3389/fnbeh.2013.00159

Yorzinski, J. L. (2017). The cognitive basis of individual recognition. *Current Opinion in Behavioral Sciences, 16*, 53–57. https://doi.org/10.1016/j.cobeha.2017.03.009

Zala, S. M., Nicolakis, D., Marconi, M. A., Noll, A., Balazs, P., & Penn, D. J. (2019). *Primed to vocalize: wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female* (submitted for publication).

Zala, S. M., Reitschmidt, D., Noll, A., Balazs, P., & Penn, D. J. (2017a). Sex-dependent modulation of ultrasonic vocalizations in house mice (*Mus musculus musculus*). *PloS One, 12*. https://doi.org/10.1371/journal.pone.0188647. e0188647.

Zala, S. M., Reitschmidt, D., Noll, A., Balazs, P., & Penn, D. J. (2017b). Automatic mouse ultrasound detector (A-MUD): A new tool for processing rodent vocalizations. *PloS One, 12*(7). https://doi.org/10.1371/journal.pone.0181200. e0181200.

APPENDIX

Detection of Individual Signatures: PIC

We conducted an analysis to quantify individual variation and to determine which parameters provide candidates for individual signatures, called PIC (potential for individual coding; Favaro et al., 2015). PIC compares variation within individuals (coefficient of variation within individuals, CVw) to between-individual variation (coefficient of variation between individuals, CVb) for each parameter, and all the parameters with a PIC > 1 are considered to be candidates for individual signatures. Unlike multivariate analyses, each parameter was analysed separately and assigned a distinct PIC score. We could only analyse vocalizing males and therefore excluded one mouse that never vocalized, one that vocalized only once and another from which we had USVs only for week 3. We examined 27 USV parameters and used four different models. The first and second PIC models sampled 13 mice vocalizing in both phases 1 and 2 in all 3 weeks (model 1 and model 2). We selected the same number of USVs per week and male to have equal sample sizes. We analysed the first 48 USVs from each mouse and week ($N = 144$ USVs per mouse and 48 USVs/week), which was the minimum number of USVs emitted by the lowest vocalizer in a week. Model 1 was run with six parameters (minimum frequency, mean frequency, maximum frequency, frequency bandwidth, ICI and USV length) calculated separately from each USV. Model 2 was run with 'summary' parameters ($N = 39$ recordings, 13 mice recorded in 3 weeks) and using 20 parameters, including repertoire size, percentage of the repertoire composition (i.e. 15 syllable types), mean ISI and IBI, latency to emit the first USV and total USV length. We excluded 'USV count', as it was the same for all mice ($N = 144$). Since we had equal sample sizes in both models (USV count/mouse and number of recordings/mouse, respectively), we performed Friedman tests for each parameter to detect individual variation between weeks and corrected for multiple testing with FDR using the Benjamini–Hochberg procedure. The third and fourth models included all 19 males that vocalized in at least one of the stimulation phases in all 3 weeks (model 3 and model 4). Model 3 included six parameters calculated separately from each of all recorded USVs ($N = 23\,414$). Model 4 was run on 'summary' parameters from each individual and week ($N = 57$ recordings, 19 mice recorded in 3 weeks) and 21 parameters (the same previous 20 parameters and the USV count). Thus, we performed Friedman tests for each parameter since we had equal sample sizes (i.e. number of recordings/mouse), to compare individual variation between weeks and correct ed with FDR (Benjamini–Hochberg procedure) for multiple testing. For the 'summary' parameters in models 2 and 4 we calculated the percentages of each syllable type, the repertoire size, the USV count and the grand mean of the ISI, IBI and the total USV length for each recording. The 'nonsummary' parameters in models 1 and 3 did not require these computational steps and each USV with its own parameters was included in the PIC calculation as such.

We ran four different models to present both continuous parameters related to each USV and the parameters extracted after performing additional calculations for each recording, which involved some trade-offs. We could compare models 1 and 3 since both included 'nonsummary' parameters (e.g. minimum, mean and maximum frequency) and models 2 and 4 which included 'summary' parameters (e.g. USV count, repertoire size, percentage of syllable types). However, we could not directly compare models that had the same number of individuals (i.e. models 1 and 2 sampling 13 mice and models 3 and 4 sampling 19 males) since they belonged to data sets computed in a different way. Models 1 and 3 included thousands of USVs (1872 USVs in model 1 and 23 414 USVs in model 3) and models 2 and 4 included parameters computed from all the USVs included in each recording (39 recordings in model 1 and 57 recordings in model 4).

Individual Variation in Repertoire Composition: ANOSIM

To examine the repertoire composition for each mouse and recording day (3 weeks) separately, we conducted an ANOSIM, comparing the similarity for some variables in one recording within one mouse with the similarity in the same variables between different individuals and the other two recordings. We ran 999 permutations and selected the Bray–Curtis dissimilarity as a distance measure based on ranks. We report the R value to show the degree of separation between groups ($-1 < R > 1$; Clarke & Warwick, 2001). The Bray–Curtis dissimilarity index was also included in the NMDS. We report the stress coefficient that relates the distance matrix to its graphical configuration in two dimensions (stress < 0.05 is excellent, but a stress > 0.3 is not reliable; Clarke & Warwick, 2001). Closer distances between samples represent more similarity in the parameters measured (i.e. a male

with the same USV emission in his three trials should have overlapping recordings in the NMDS representation).

Spectrographic Features for Individual Detection: Random Forest

To test all 26 USV parameters previously used in the PIC, we conducted a random forest analysis and graphed the results using t-SNE. This algorithm embeds high-dimensional data in a low-dimensional space (two dimensions in our study) creating a probability distribution between pairwise multidimensional objects. Similar multidimensional objects have a higher probability of being closer to each other than dissimilar objects. A perplexity value was set at 30 ($5 < \text{perplexity} \leq 50$), a fixed value to standardize the scaling density of the points that are reduced in dimensions when multidimensional data are transferred in a two-dimensional space (van der Maaten & Hinton, 2008; van der Maaten, 2009). Before running the models, we ran Spearman correlation matrices to remove highly correlated parameters for three subsets. Within all correlated variables, we always kept one of them in the model (Murdoch & Chow, 1996; Friendly, 2002). For each model, we implemented the RFE for parameter selection for 10 iterations and we ran the random forest classifier for another 10 iterations (for each subset of features selected by the RFE) based on the acoustic parameters (individual signatures) selected by the RFE (100 iterations in total). Thus, we calculated the accuracy in correctly identifying the males based on comparing the random forest output with the identities of real mice (real labels or ground truths). For all three models, we selected the grand mean frequency which was highly correlated with the mean minimum and maximum frequency. We then ran three different models with different sample sizes and number of recordings from the stimulation period (i.e. three recordings when the entire 10 min of stimulation was used and six recordings when phase 1 and phase 2 were added separately). The first model included only males vocalizing in both phases 1 and 2 during the 10 min odour stimulation and in all 3 weeks ($N = 13$). We selected the grand mean frequency and the USV count correlated with the total USV length (23 parameters). A second model included data from phase 1 and phase 2 for the same males ($N = 13$) separately. We increased the number of data points for each individual (from three to six) and improved the accuracy (by adding more data to the training and test sets in the random forest classifier); however, this also introduced pseudoreplication in the data set. As a correlated parameter, we selected the grand mean frequency (24 parameters). The third model included all the vocalizing mice ($N = 19$) in all 3 weeks, and in at least one of the two phases, of the 10 min stimulation period (i.e. 3 recordings/male). We selected the grand mean frequency, the USV count correlated with the total USV length and the percentage of 'uc' correlated with the percentage of 'us' (22 parameters). After data reduction, the males' recordings with the respective parameters were implemented in the training and test sets. The 'training set' is the data used to construct a model according to fixed parameters (e.g. individual IDs and their USV features), whereas the 'test set' is another set of data, excluded from the training test, that is used to test the model (e.g. to determine whether the machine accurately classifies individuals). The training set was run on 67% (two-thirds) of the data for the 10 min audio files, and on 83% (five-sixths) of the data for the same recordings treated separately for phase 1 and phase 2. The remaining files, 33% (one-third) for the 10 min audio files and 17% (one-sixth) for the 5 min recordings, were used for the test set to determine whether these files were correctly assigned to the corresponding individual according to the parameters established during the training set. In the whole classification model, the random forest classifier assigned the sound files to the mice (classification process). At the end of the test, the accuracy values showed the model performance. To calculate the model accuracy, RFE was applied using 10 iterations. At each iteration, the random forest classifier was applied 10 times and the final accuracy was calculated as the average over all accuracies at all iterations ($10 \times 10 = 100$ iterations). Thus, the accuracy value was based on the random forest classifier output compared with the ground truths (the real data).

Results

Table A1
Comparison of USV emission between the different phases of each trial within each week

H_0: phases $0 = 1 = 2$		Friedman test		Friedman pairwise comparison	FDR (BH)
		χ^2	P	Phases	P adjusted
Week 1	USV count	20.88	< **0.001**	0 versus 1	**0.003**
				1 versus 2	0.366
				0 versus 2	< **0.001**
	Repertoire size	22.03	< **0.001**	0 versus 1	**0.002**
				1 versus 2	0.451
				0 versus 2	< **0.001**
Week 2	USV count	16.91	< **0.001**	0 versus 1	**0.001**
				1 versus 2	0.498
				0 versus 2	**0.005**
	Repertoire size	14.00	**0.001**	0 versus 1	**0.006**
				1 versus 2	0.763
				0 versus 2	**0.008**
Week 3	USV count	20.68	< **0.001**	0 versus 1	< **0.001**
				1 versus 2	0.940
				0 versus 2	**0.001**
	Repertoire size	22.14	< **0.001**	0 versus 1	< **0.001**
				1 versus 2	0.821
				0 versus 2	< **0.001**

'Phases' indicates which phases are used for post hoc comparisons ($N = 22$). Results are corrected for multiple testing (P adjusted) with false discovery rate (FDR) using the Benjamini–Hochberg procedure (BH). All significant P values are in bold.

Table A2
Comparison of USV emission between the phases of each trial

H_0: phases 0 = 1 = 2	Friedman test		Friedman pairwise Comparison	FDR (BH)
	χ^2	*P*	Phases	*P* adjusted
USV count	28.99	< **0.001**	0 versus 1	< **0.001**
			1 versus 2	0.940
			0 versus 2	< **0.001**
Repertoire size	28.61	< **0.001**	0 versus 1	< **0.001**
			1 versus 2	0.546
			0 versus 2	< **0.001**

Friedman tests were performed on the grand mean of each variable (USV count and repertoire size) for each individual over 3 weeks. 'Phases' indicates which phases are used for post hoc comparisons ($N = 22$) and FDR (false discovery rate) was used to correct for multiple testing (*P* adjusted) using the Benjamini–Hochberg procedure (BH). All significant *P* values are in bold.

Table A4
Comparison of USV emission between the weeks and within each phase

H_0: weeks 1 = 2 = 3	Phase	Friedman test	
		χ^2	*P*
USV count	**0**	5.45	**0.065**
	1	0.10	0.950
	2	0.33	0.850
Repertoire size	**0**	4.95	**0.084**
	1	0.59	0.744
	2	0.18	0.912

$N = 22$. *P* values <0.1 for USV count and repertoire size during phase 0 are in bold.

Table A3
Comparison of repertoire composition between the phases of each trial and within each week

Week	Syllable type	Friedman test		FDR (BH)	Friedman pairwise comparison	Post hoc comparing phases
		χ^2	*P*	*P* adjusted	Phases	*P* BH corrected
1	s	8.00	**0.018**	**0.018**		NS
	f	17.16	< **0.001**	< **0.001**	0 versus 1	**0.015**
					0 versus 2	**0.003**
	d	9.71	**0.008**	**0.009**	0 versus 1	**0.03**
	up	23.76	< **0.001**	< **0.001**	0 versus 1	< **0.001**
					0 versus 2	**0.002**
	u	21.34	< **0.001**	< **0.001**	0 versus 1	**0.045**
					0 versus 2	< **0.001**
	ui	20.41	< **0.001**	< **0.001**	0 versus 1	**0.003**
					0 versus 2	**0.005**
	c	18.03	< **0.001**	< **0.001**	0 versus 1	**0.012**
					0 versus 2	**0.003**
	c2	18.33	< **0.001**	< **0.001**	0 versus 1	**0.008**
					0 versus 2	**0.003**
	c3	17.23	< **0.001**	< **0.001**	0 versus 1	**0.024**
					0 versus 2	**0.009**
	uh	5.36	0.069	NS		
2	f	13.90	**0.001**	**0.002**	0 versus 1	**0.015**
					0 versus 2	**0.024**
	d	15.16	**0.001**	**0.002**	0 versus 1	**0.009**
					0 versus 2	**0.036**
	up	13.37	**0.001**	**0.002**	0 versus 1	**0.009**
					0 versus 2	**0.024**
	u	15.26	< **0.001**	< **0.001**	0 versus 2	**0.009**
	ui	16.59	**0.001**	**0.002**	0 versus 1	**0.009**
					0 versus 2	**0.024**
	c	8.34	**0.015**	**0.015**		NS
	c2	12.04	**0.002**	**0.003**	0 versus 1	**0.044**
					0 versus 2	**0.044**
	c3	12.33	**0.002**	**0.003**	0 versus 2	**0.048**
	h	9.91	**0.007**	**0.008**		NS
	uc	5.28	0.071	NS		
3	f	14.80	**0.001**	**0.002**	0 versus 1	**0.015**
					0 versus 2	**0.006**
	d	5.54	0.063	NS		
	u	10.92	**0.004**	**0.005**	0 versus 1	**0.036**
					0 versus 2	**0.036**
	ui	13.21	**0.001**	**0.002**	0 versus 1	**0.009**
	c	7.09	**0.029**	**0.029**		NS
	c3	14.33	**0.001**	**0.002**	0 versus 1	**0.021**
	c4	4.93	0.085	NS		
	h	4.65	0.098	NS		

Only syllable types showing significant differences or trends between phases are reported. 'Phases' indicates which phases are used for post hoc comparisons ($N = 22$) and FDR (false discovery rate) was used to correct for multiple testing (*P* adjusted) using the Benjamini–Hochberg procedure (BH). All significant *P* values are in bold.

Table A5
Comparison of repertoire composition between the weeks and within each phase

H_0: weeks 1 = 2 = 3	Syllable type	Friedman test		FDR (BH)	Post hoc comparing weeks
		χ^2	P	P adjusted	P BH corrected
Phase 0	up	6.93	**0.031**	**0.047**	NS
	c2	10.40	**0.006**	**0.018**	NS
	ui	5.60	0.061		
Phase 1	uc	6.19	**0.045**	NS	NS
	c2	5.18	0.075		
Phase 2	up	7.43	**0.024**	NS	NS
	us	4.62	0.099		
	s	4.78	0.092		
	c	5.45	0.065		

Results are reported only for significant values and trends ($N = 22$). Post hoc comparisons between weeks were run on significant values (BH corrected). False discovery rate (FDR) was used to correct for multiple testing (P adjusted) using the Benjamini–Hochberg procedure (BH). Significant P values are in bold.

Table A6
Relationship between USV count and repertoire size within each phase and week

Phase	Week	Spearman rank correlation			FDR (BH)
		R^2	r_S	P	P adjusted
0	1	0.93	1.00	< **0.001**	< **0.001**
	2	0.86	0.99	< **0.001**	< **0.001**
	3	0.92	0.99	< **0.001**	< **0.001**
1	1	0.95	0.88	< **0.001**	< **0.001**
	2	0.92	0.90	< **0.001**	< **0.001**
	3	0.87	0.86	< **0.001**	< **0.001**
2	1	0.89	0.75	< **0.001**	< **0.001**
	2	0.94	0.93	< **0.001**	< **0.001**
	3	0.84	0.92	< **0.001**	< **0.001**
1+2	1	0.89	0.74	< **0.001**	< **0.001**
	2	0.90	0.82	< **0.001**	< **0.001**
	3	0.87	0.89	< **0.001**	< **0.001**

$N = 22$. Phase 1 + 2 indicates the entire 10 min stimulation period. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied to each phase (P adjusted). Significant P values are in bold.

Table A7
Relationship between weeks in each phase to show individual consistency in USV count and repertoire size over time

Parameter	Phase	Weeks	Spearman rank correlation			FDR (BH)
			R^2	r_S	P	P adjusted
USV count	0	1 versus 2	0.00	0.32	0.157	
		2 versus 3	0.11	0.50	**0.018**	NS
		1 versus 3	< 0.001	0.07	0.768	
	1	1 versus 2	0.42	0.76	< **0.001**	< **0.001**
		2 versus 3	0.51	0.68	< **0.001**	< **0.001**
		1 versus 3	0.20	0.51	**0.016**	**0.016**
	2	1 versus 2	0.42	0.65	**0.001**	**0.001**
		2 versus 3	0.47	0.78	< **0.001**	< **0.001**
		1 versus 3	0.16	0.52	**0.012**	**0.012**
Repertoire size	0	1 versus 2	0.01	0.32	0.147	
		2 versus 3	0.38	0.48	**0.023**	NS
		1 versus 3	0.04	0.09	0.707	
	1	1 versus 2	0.70	0.84	< **0.001**	< **0.001**
		2 versus 3	0.39	0.53	**0.012**	**0.012**
		1 versus 3	0.33	0.55	**0.009**	**0.012**
	2	1 versus 2	0.31	0.39	0.071	
		2 versus 3	0.66	0.72	< **0.001**	< **0.001**
		1 versus 3	0.15	0.25	0.255	

$N = 22$. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied to each phase (P adjusted). Significant P values are in bold.

Table A8
Relationship of USV emission between the first and second half of the stimulation period (phases 1 and 2) in each trial

Parameter	Week	Spearman rank correlation			FDR (BH)
		R^2	r_S	P	P adjusted
USV count	1	0.60	0.86	< **0.001**	< **0.001**
	2	0.55	0.88	< **0.001**	< **0.001**
	3	0.73	0.92	< **0.001**	< **0.001**
Repertoire size	1	0.82	0.79	< **0.001**	< **0.001**
	2	0.83	0.82	< **0.001**	< **0.001**
	3	0.80	0.84	< **0.001**	< **0.001**

$N = 22$. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied to each phase (P adjusted). Significant P values are in bold.

Table A9
Potential individual coding (PIC) for USV parameters over 3 weeks (10 min of stimulation)

Model No. of mice (no. of parameters)	Data set No. of USVs or recordings	Parameter	Mean CVw ± SD	CVb	PIC	Friedman test		FDR (BH)
						χ^2	*P*	*P* adjusted
Model 1								
$N = 13$	$N = 1872$ USVs	**Frequency parameters**						
(6)		Frequency bandwidth (kHz)	78.03 ± 20.02	80.35	1.03	329.42	< **0.001**	< **0.001**
		Maximum frequency (kHz)	19.76 ± 22.50	16.09	0.81	416.19	< **0.001**	< **0.001**
		Minimum frequency (kHz)	20.10 ± 22.53	15.97	0.79	356.38	< **0.001**	< **0.001**
		Mean frequency (kHz)	18.26 ± 22.91	14.26	0.78	503.19	< **0.001**	< **0.001**
		General parameters						
		ICI (ms)	295.87 ± 127.78	457.79	1.55	129.61	< **0.001**	< **0.001**
		USV length (ms)	50.46 ± 18.40	50.79	1.01	339.46	< **0.001**	< **0.001**
Model 2								
$N = 13$	$N = 39$ recordings	**Percentage of syllable types**						
(20)	(3 recordings/mouse)	Perc.c5	141.67 ± 0.00	616.44	4.35	12.00	0.446	
		Perc.uh	118.06 ± 33.39	392.17	3.32	16.51	0.169	
		Perc.c4	125.68 ± 22.61	330.74	2.63	16.68	0.162	
		Perc.uc	127.41 ± 24.44	324.11	2.54	15.95	0.193	
		Perc.h	119.60 ± 28.79	256.58	2.15	16.20	0.182	
		Perc.c3	96.39 ± 36.45	194.70	2.02	24.96	**0.015**	**0.022**
		Perc.up	18.22 ± 24.31	36.49	2.00	31.18	**0.002**	**0.015**
		Perc.ui	57.03 ± 28.40	102.58	1.80	27.28	**0.007**	**0.019**
		Perc.us	120.60 ± 30.44	209.63	1.74	11.41	0.494	
		Perc.c2	81.85 ± 40.93	133.92	1.64	20.37	0.060	
		Perc.c	107.40 ± 42.96	170.45	1.59	19.50	0.077	
		Perc.d	63.41 ± 32.34	90.92	1.43	24.84	**0.016**	**0.022**
		Perc.u	79.23 ± 41.59	109.14	1.38	23.04	**0.027**	**0.033**
		Perc.s	94.00 ± 44.13	125.80	1.34	18.44	0.103	
		Perc.f	59.53 ± 38.26	78.07	1.31	21.51	**0.043**	**0.043**
		General parameters						
		IBI (ms)	33.08 ± 14.87	93.99	2.84	29.28	**0.004**	**0.015**
		ISI (ms)	6.54 ± 2.85	16.77	2.57	26.29	**0.010**	**0.022**
		Repertoire size	12.27 ± 9.76	29.29	2.39	29.96	**0.003**	**0.015**
		Total USV length (ms)	10.68 ± 7.59	24.82	2.32	25.36	**0.013**	**0.022**
		Latency to first USV (s)	30.87 ± 14.79	51.69	1.67	22.33	**0.034**	**0.037**
Model 3								
$N = 19$	$N = 23\,414$ USVs	**Frequency parameters**						
(6)		Maximum frequency (kHz)	14.50 ± 3.39	16.26	1.12	Highly unequal sample sizes		
		Mean frequency (kHz)	12.76 ± 3.82	13.92	1.09	(not possible to compute Friedman		
		Minimum frequency (kHz)	15.30 ± 4.50	15.69	1.03	tests for validation)		
		Frequency bandwidth (kHz)	94.99 ± 57.44	76.20	0.80			
		General parameters						
		ICI (ms)	361.86 ± 137.78	517.53	1.43			
		USV length (ms)	48.56 ± 13.83	46.37	0.95			
Model 4								
$N = 19$	$N = 57$ recordings	**Percentage of syllable types**						
(21)	(3 recordings/mouse)	Perc.h	121.12 ± 50.21	494.96	4.09	29.99	**0.038**	**0.038**
		Perc.c5	180.28 ± 10.41	514.80	2.86	25.89	0.102	
		Perc.us	108.02 ± 53.25	306.62	2.84	26.35	0.092	
		Perc.uh	132.02 ± 44.38	344.90	2.61	33.24	**0.016**	**0.017**
		Perc.c	103.49 ± 71.22	227.15	2.20	42.93	**0.001**	**0.003**
		Perc.c4	174.17 ± 31.70	362.02	2.08	25.37	0.115	
		Perc.uc	152.09 ± 41.68	284.11	1.87	22.88	0.195	
		Perc.s	89.64 ± 49.96	165.02	1.84	19.02	0.390	
		Perc.c3	105.62 ± 51.08	180.75	1.71	38.12	**0.004**	**0.008**
		Perc.d	67.40 ± 60.64	111.13	1.65	45.30	**<0.001**	**<0.001**
		Perc.f	66.29 ± 64.43	104.98	1.58	37.30	**0.005**	**0.009**
		Perc.c2	94.05 ± 56.42	148.37	1.58	36.53	**0.006**	**0.009**
		Perc.ui	73.34 ± 57.98	108.46	1.48	35.46	**0.008**	**0.011**
		Perc.u	68.94 ± 54.89	96.22	1.40	40.51	**0.002**	**0.005**
		Perc.up	48.29 ± 56.96	51.11	1.06	36.29	**0.006**	**0.009**
		General parameters						
		IBI (ms)	68.44 ± 41.58	228.82	3.34	38.78	**0.003**	**0.007**
		Latency to first USV (s)	49.24 ± 22.99	105.60	2.14	41.22	**0.001**	**0.003**
		ISI (ms)	19.15 ± 27.34	34.95	1.83	33.28	**0.015**	**0.017**
		Repertoire size	35.84 ± 37.92	43.49	1.21	33.76	**0.013**	**0.016**
		Total USV length (ms)	85.23 ± 55.07	102.18	1.20	42.46	**0.001**	**0.003**
		USV count	81.75 ± 52.31	90.54	1.11	42.34	**0.001**	**0.003**

We analysed 26 parameters for a subset of mice vocalizing in phases 1 and 2 ($N = 13$) and 27 parameters for all vocalizing mice ($N = 19$). The coefficients for within-individual variation (CVw) and between-individual variation (CVb) and the PIC (CVw/CVb) are shown for each variable and for the acoustic parameters in each category (i.e. frequency parameters, percentage of syllable types and general parameters) from the highest to the lowest PIC. ICI = intercall interval; IBI = interbout interval; ISI = intersyllable interval. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied after Friedman tests (χ^2, *P*) to analyse differences between individuals for significant parameters (*P* adjusted). We were not able to run any Friedman test for model 3 due to highly unequal sample sizes in USV counts between mice (11–3276 USVs) in 3 weeks. Significant *P* values are in bold.

Table A10
Candidates for individual signatures after RFE (recursive feature elimination based on the random forest classifier) with different sample sizes

File duration (min) (no. of individuals)	No. of parameters	Parameters 100% trials	Parameters > 50% trials	Parameters 50% trials	Accuracy (%)
10 (13)	23	GrandFmean MeanFband ISI Percentage of f and c2	MeanLength Percentage of up, d, u, ui, c and c3	Latency to first USV	86
5 + 5 (13)	24	GrandFmean MeanFband MeanLength Percentage of up, f, d, u, ui, c2 and c3	ISI Percentage of c, h and uc	USV count totalLength IBI percentage of c4	89
10 (19)	22	GrandFmean Percentage of d and f	MeanFband MeanLength ISI Latency to first USV Percentage of u, ui, c, c2 and c3	USV count Percentage of up	56

The acoustic parameters chosen during all 10 iterations (100%), in more than five iterations (>50%) or in five iterations (50%) and the respective accuracy values are shown. The RFE and the random forest classifier were run on three different subsets of mice and sound files (see Methods). GrandFmean = grand mean frequency, meanFband = mean frequency bandwidth, meanLength = mean USV length, total Length = sum of the lengths of all USVs in one sound file, ISI = intersyllable interval, IBI = interbout interval.

Table A11
Interindividual variation for frequency parameters over 3 weeks

Group	Parameter	Mean ± SD	Friedman test	
			χ^2	P
Usually vocal	Grand mean frequency	60.74 ± 13.29	36.27	**0.007**
	Mean frequency bandwidth	10.82 ± 5.55	42.13	**0.001**
Always vocal	Grand mean frequency	57.77 ± 4.79	33.58	**0.001**
	Mean frequency bandwidth	12.18 ± 4.05	31.69	**0.002**

Comparison of the frequency parameters (grand mean frequency and mean frequency bandwidth) between individuals calculated on 3 recordings/mouse. Usually vocal: all mice vocalizing during stimulation in at least one of the two phases ($N = 19$); always vocal: mice that vocalized in both phases 1 and 2 during stimulation and over the 3 weeks ($N = 13$). Significant P values are in bold.

Table A12
Relationship between the percentage of syllable type and the mean USV length within each week

Parameter (% syllable type)	Week	Spearman rank correlation			FDR (BH)
		R^2	r_S	P	P adjusted
u	1	0.26	0.48	**0.038**	**0.038**
	2	0.32	0.65	**0.003**	**0.009**
	3	0.15	0.52	**0.022**	**0.033**
c	1	0.07	0.48	**0.039**	**0.039**
	2	0.35	0.65	**0.003**	**0.009**
	3	0.13	0.59	**0.008**	**0.012**
c3	1	0.11	0.41	0.078	
	2	0.09	0.54	**0.017**	**0.017**
	3	0.25	0.71	**0.001**	**0.002**

The correlation coefficient (R^2) between the repertoire composition (percentage of syllable types) and the USV length is shown for each significant syllable type or trend. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied to each parameter (P adjusted). Significant P values are in bold.

Table A13
Relationship between weeks to show individual consistency in individual signatures

Parameter	Correlation Method	Weeks	R^2	Correlation coefficient	P	FDR (BH)
						P adjusted
Grand mean frequency	Spearman	1 versus 2	0.01	0.46	**0.047**	**0.047**
		2 versus 3	0.44	0.65	**0.003**	**0.005**
		1 versus 3	0.02	0.72	**< 0.001**	**< 0.001**
Mean frequency bandwidth	Pearson	1 versus 2	0.05	0.21	0.378	
		2 versus 3	0.65	0.80	**< 0.001**	**< 0.001**
		1 versus 3	0.08	0.28	0.247	
	Spearman	1 versus 2	0.05	0.57	**0.011**	**0.011**
		2 versus 3	0.65	0.85	**< 0.001**	**< 0.001**
		1 versus 3	0.08	0.60	**0.006**	**0.009**
Percentage of f	Pearson	1 versus 2	0.09	0.29	0.222	
		2 versus 3	0.02	0.13	0.605	
		1 versus 3	0.37	0.61	**0.005**	**0.015**
	Spearman	1 versus 2	0.09	0.37	0.120	
		2 versus 3	0.02	0.41	0.080	
		1 versus 3	0.37	0.83	**< 0.001**	**< 0.001**
Percentage of d	Spearman	1 versus 2	0.50	0.70	**0.001**	**0.002**
		2 versus 3	0.71	0.93	**< 0.001**	**< 0.001**
		1 versus 3	0.54	0.64	**0.003**	**0.003**
Percentage of c2	Spearman	1 versus 2	0.23	0.47	**0.042**	**0.042**
		2 versus 3	0.62	0.79	**< 0.001**	**< 0.001**
		1 versus 3	0.32	0.29	0.237	
Percentage of up	Spearman	1 versus 2	0.14	0.36	0.132	
		2 versus 3	0.56	0.67	**0.002**	**0.006**
		1 versus 3	0.2	0.49	**0.032**	**0.048**
ISI	Spearman	1 versus 2	0.14	0.62	**0.005**	**0.015**
		2 versus 3	0.18	0.23	0.355	
		1 versus 3	0.09	0.43	0.065	
Mean USV length	Pearson	1 versus 2	0.41	0.64	**0.003**	**0.005**
		2 versus 3	0.72	0.85	**< 0.001**	**< 0.001**
		1 versus 3	0.34	0.58	**0.009**	**0.009**
	Spearman	1 versus 2	0.41	0.60	**0.007**	**0.007**
		2 versus 3	0.72	0.85	**< 0.001**	**< 0.001**
		1 versus 3	0.34	0.66	**0.002**	**0.003**

The table includes parameters chosen as individual signatures by the recursive feature elimination based on the random forest classifier ($N = 19$) and the percentage of 'up'. Correlations between weeks are shown for each parameter and phase. Spearman rank correlation coefficients: r_S and R^2; Pearson correlation coefficients: r and R^2. False discovery rate (FDR) correction for multiple testing using the Benjamini–Hochberg procedure (BH) was applied to each parameter (P adjusted). Significant P values are in bold.

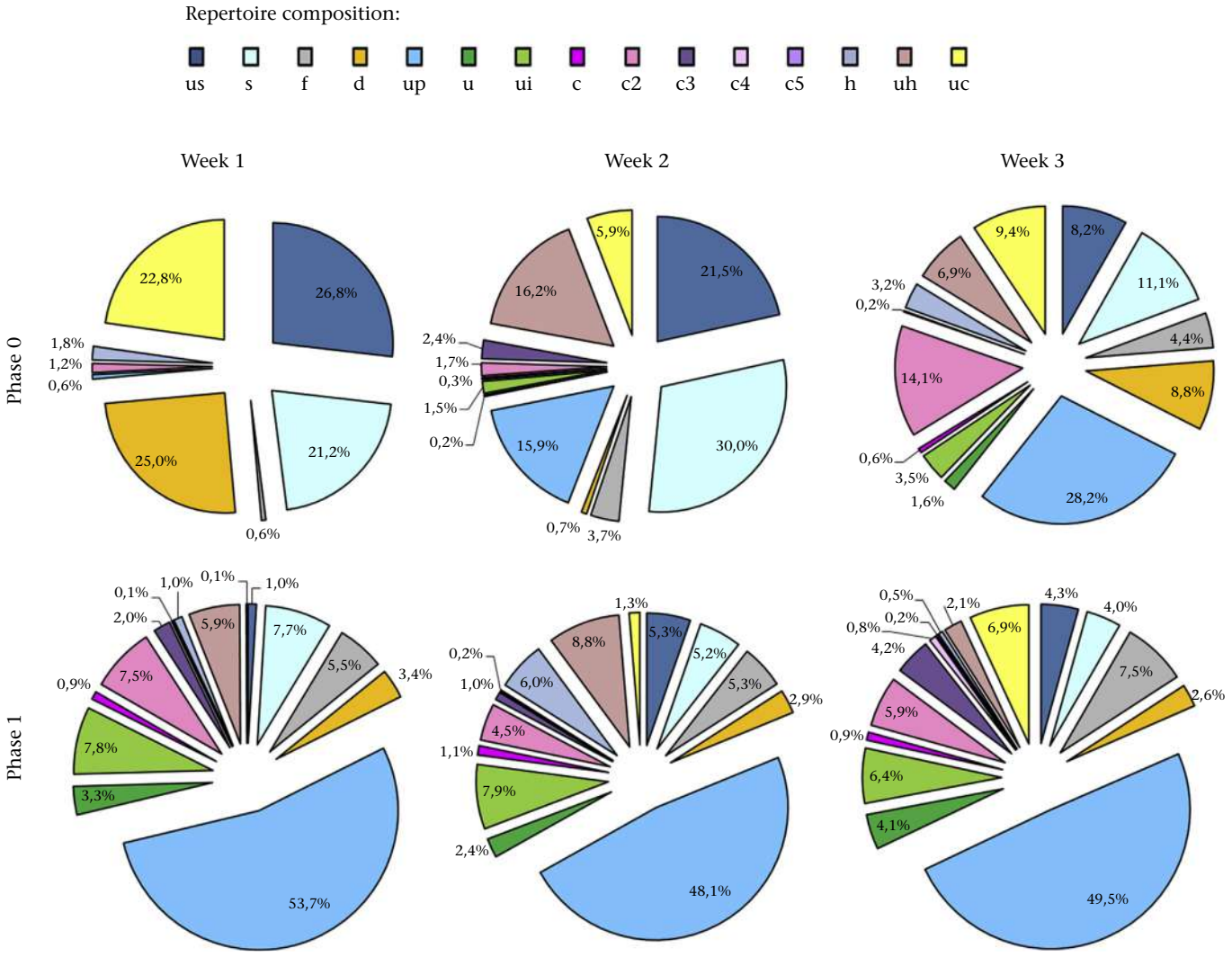

Figure A1. Repertoire composition before and during sexual stimulation and over 3 weeks. The pie charts show differences in repertoire composition of USVs to compare the different types of vocalizations emitted before versus during odour stimulation (phase 0 versus phase 1) and over the 3 weeks ($N = 22$ individual males). The syllable types are presented on the pie graphs in the same order and colour as shown in the legend (e.g. 'd' is shown in orange and is always followed by 'up' in light blue). Results from phase 2 are not shown, because they do not differ from those of phase 1. See Appendix Table A3 for statistical analyses.

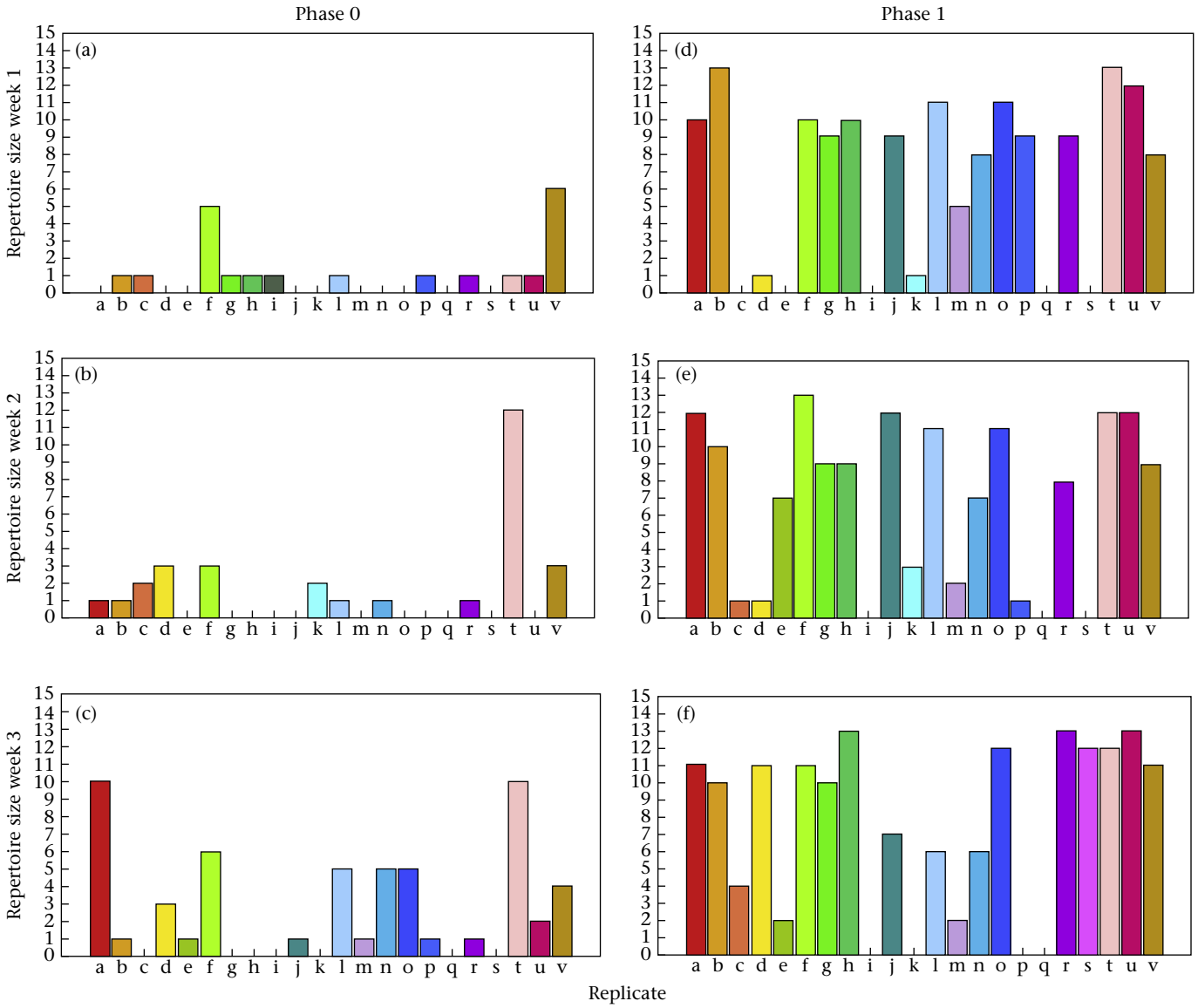

Figure A2. Interindividual variation in repertoire size before and during sexual stimulation. Repertoire size is shown during (a, b, c) phase 0 and (d, e, f) phase 1 in (a, d) week 1, (b, e) week 2 and (c, f) week 3 ($N = 22$). The same letter and colour correspond to the same individual (replicate). Two mice did not vocalize at all over the 3 weeks (replicates i and q).

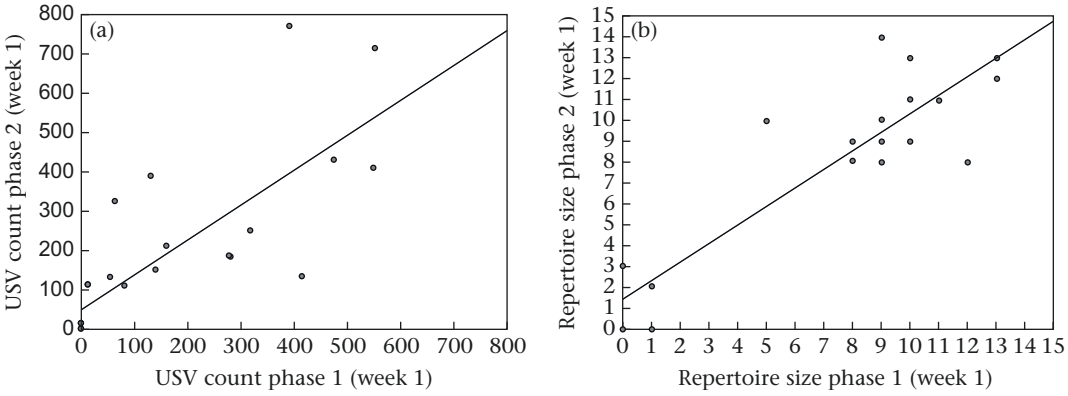

Figure A3. Relationship of male USV emission between stimulation phases. (a) USV count and (b) repertoire size are shown for week 1 and phase 1 versus phase 2 (5 min each, $N = 22$). Correlations for the other 2 weeks are not shown. See Appendix Table A8 for statistical analyses.

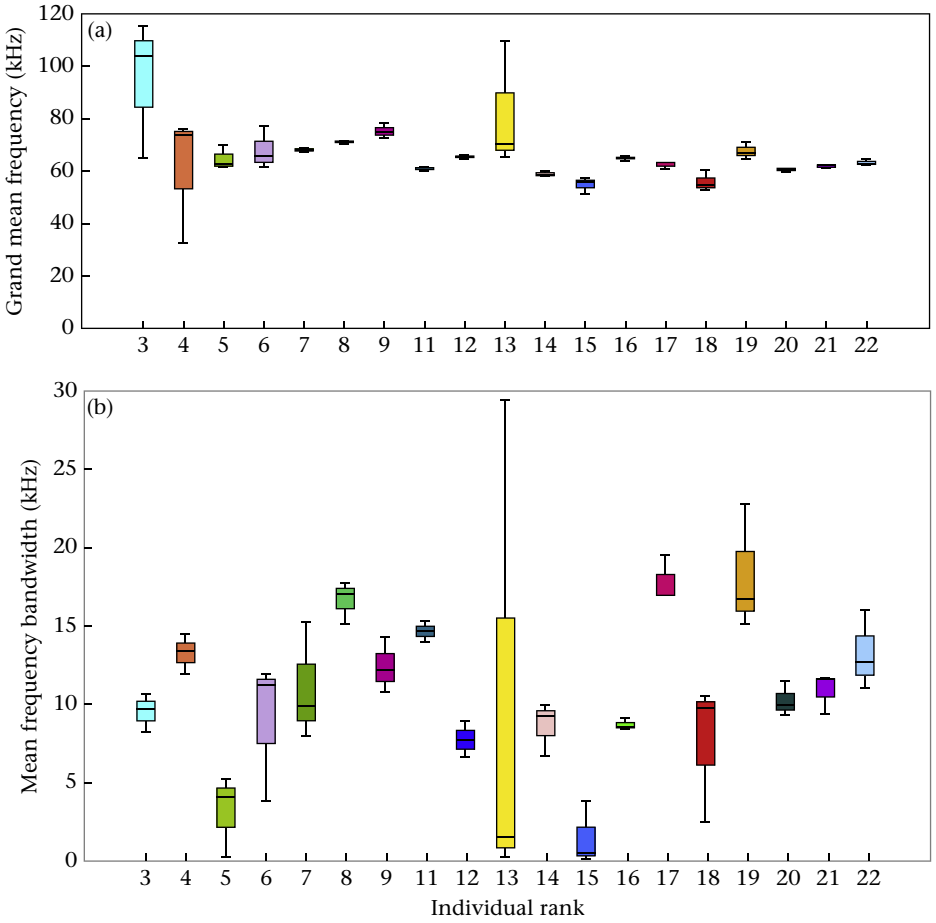

Figure A4. Individual variation in USV grand mean frequency and frequency bandwidth during stimulation. (a) The grand mean frequency and (b) the mean frequency bandwidth during the 10 min stimulation period from 3 recordings/mouse ($N = 19$). Note that (a) and (b) have different *y*-axes. Box plots display the median, interquartile ranges (25th and 75th percentile) and 95% confidence interval (whiskers). The same number and colour correspond to the same individual. Individuals are ranked according to the interindividual variation (standard deviation) in the USV count. See Appendix Table A11 for statistical analyses.

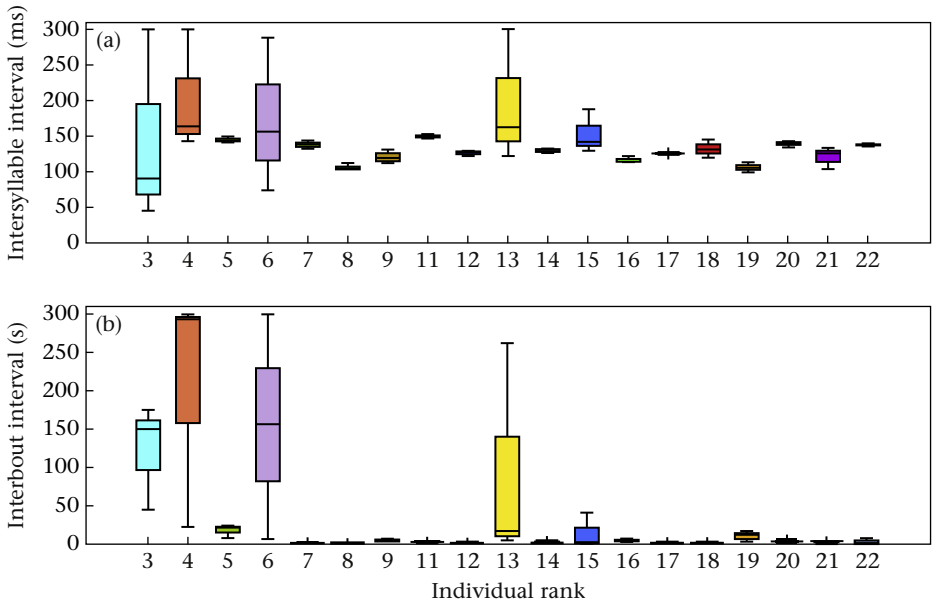

Figure A5. Individual variation in intersyllable and interbout interval. (a) The mean intersyllable interval and (b) the mean interbout interval for each male during the 10 min stimulation period and over 3 weeks ($N = 19$). Box plots display the median, interquartile ranges (25th and 75th percentile) and 95% confidence interval (whiskers). The same number and colour correspond to the same individual. Individuals are ranked according to the interindividual variation (standard deviation) in the USV count. See Appendix Table A9 for statistical analyses.

CHAPTER 5: Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences

Caruso, Angela; Marconi, Maria Adelaide;
Scattoni, Maria Luisa; Ricceri, Laura

Genes, Brain & Behavior, impact factor: 3.7; **published**

Received: 29 January 2022 | Revised: 29 April 2022 | Accepted: 4 May 2022

DOI: 10.1111/gbb.12815

ORIGINAL ARTICLE

Genes, Brain and Behavior

Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences

Angela Caruso[1] | Maria Adelaide Marconi[2] | Maria Luisa Scattoni[1] | Laura Ricceri[3]

[1]Research Coordination and Support Service, Istituto Superiore di Sanità, Rome, Italy

[2]Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine, Vienna, Austria

[3]Center for Behavioral Sciences and Mental Health, Istituto Superiore di Sanità, Rome, Italy

Correspondence
Maria Luisa Scattoni, Research Coordination and Support Service, Istituto Superiore di Sanità, Rome, Italy.
Email: marialuisa.scattoni@iss.it

Funding information
2020-2022 Ricerca Corrente ISS, Centro per le Scienze Comportamentali e la salute mentale "Modelli sperimentali di disturbi del neurosviluppo e neuropsichiatrici"

Abstract

Mice produce ultrasonic vocalizations (USVs) in different social contexts across lifespan. There is ethological evidence that pup USVs elicit maternal retrieval and adult USVs facilitate social interaction with a conspecific. Analysis of mouse vocal and social repertoire across strains, sex and contexts remains not well explored. To address these issues, in inbred (C57BL/6, FVB) and outbred (CD-1) mouse strains, we recorded and evaluated USVs as neonates and during adult social encounters (male–female and female–female social interaction). We showed significant strain differences in the quantitative (call rate and duration of USVs) and qualitative vocal analysis (spectrographic characterization) from early stage to adulthood, in line with specific patterns of social behaviors. Inbred C57BL/6 mice produced a lower number of calls with less internal changes and shorter duration; inbred FVB mice displayed more social behaviors and produced more syllables with repeated internal changes; outbred CD-1 mice had an intermediate profile. Our results suggest specific vocal signatures in each mouse strain, thus helping to better define socio-communicative profiles of mouse strains and to guide the choice of an appropriate strain according to the experimental settings.

KEYWORDS
development, mouse communication, neonatal behavior, social interaction, vocal repertoire

1 | INTRODUCTION

Ultrasonic vocalizations (USVs) by laboratory mice have been collected and deeply analyzed in different contexts during the early phases of postnatal development and at adulthood.

During the neonatal stages, USVs are emitted when pups are isolated from the nest to gain their mother's attention.[1] These calls are defined "isolation-induced USVs" and have been extensively characterized. The rate of emission follows a clear ontogenetic profile, peaking during the first postnatal week and then decreasing to zero when pups are 2-week old.[2,3] Since their first description it was suggested that neonatal USVs played an important role in vocal communication.[4] Although functional significance of such vocalizations has been debated,[5,6] there is sound ethological evidence that pup USVs elicit maternal orientation/approach and retrieval.[3,7–9] Previous data suggested that mother's genotype or strain and maternal responsiveness (an index of mother's solicitude towards pups in a potentially dangerous situation) can affect neonatal USV emission.[10] It is likely that pups' behavioral changes are expressed in parallel to mother's behavior.

USVs have been also detected in adolescent mice of both sexes after weaning, during a social interaction paradigm (consisting of

5-days social housing and one-day social isolation between behavioral tests[11]), as well as in mice exposed to an anesthetized female.[12]

In adult mice, emission of USVs has been primarily reported in reproductive contexts, with males being responsible of most of the calls,[13,14] although females have been recently shown to actively participate to the vocal interaction, but with a lower number of USVs when compared with males' vocalization rate.[15–17] A short or long exposure to either a female partner (previous socio-sexual or reproductive experience) or to female urine induces a clear USV response in adult male mice.[13,14,18–23] The quantitative analysis performed by Holy and Guo in 2005 illustrated for the first time that the male USVs are characterized by temporal sequences and that they are specific for each individual.[19] More recent studies based on analyses of spectral parameters and temporal sequences on USVs highlighted the individual signatures in both pup and adult mice.[24–27] The production of USVs during adult female–female mouse encounters is also a sound phenomenon[28]: resident female mice during encounters with a female intruder emit a large number of USVs, at rates comparable to those of the male–female interaction.[29] These female calls, that only occur during resident-intruder interactions in laboratory conditions, contribute to the establishment of female social dominance hierarchies,[28] but may also serve to enhance physical proximity and enable social information gathering.[17,28,30,31] Both pup and adult mice emit USVs to communicate with each other and to convey their emotional state.[1]

Whereas the earlier studies initially provided only quantitative data (primarily rate of USV emission and duration), in the last years qualitative analyses had been also carried out. Several categorizations of the spectrographic appearances of the calls have been proposed, some of which share some basic principles.[11,32,33] Following pivotal works, USV categorization has been used as a biomarker to identify qualitative alterations of the vocal repertoire in different mouse models of neurodevelopmental disorders. USV categorizations have been explored in both neonates and adult subjects modeling socio-communicative deficits, including autism spectrum disorders.[1,34] Recent USV data from different mouse strains are also available, so far limited to early phases of postnatal development (first 2–3 weeks of postnatal life).[11,35,36] Together with previous data on cross fostering at birth[37] or embryo transfer procedure,[38] these data indicate that USV production and their acoustic variations are subjected to genetic and background control.[39,40] Crucially, differences in USV emission have been detected even between mouse substrains, such as C57BL/6N and C57BL/6J.[38] The embryo transfer study showed that the difference between C57BL/6J OlaHsd and C57BL/6N Crl in USV rate was primarily dependent on the dyadic interaction between mother and pup.[38]

Aim of our study was to evaluate vocal repertoire in three common mouse strains (C57BL/6, FVB and CD-1) in both males and females, at two developmental stages (neonatal and adult), as for adult subjects in two different social contexts (male–female and female–female interaction) known to elicit maximal vocalization rates in laboratory settings. In adult testing, the social investigation was also recorded to obtain a more complete picture of the social responses in these three strains. We selected the inbred strain C57BL/6 since it is the commonest background for genetically modified lines and widely used in mouse phenotyping studies; the inbred strain FVB since it is considered highly social and some features (i.e., albinism, litter and body size) render it an ideal control for outbred strains; and CD1 as the commonest mouse outbred strain, extensively used in neuroscience, neuropharmacology and neurotoxicology studies. The choice of C57BL/6 substrains is a critical issue in experimental designs dealing with genetically modified mouse lines, since several behavioral phenotypic differences have been reported among mouse substrains.[38,41,42] We selected the C57BL/6N substrain that has been less characterized in previous studies and it is becoming increasingly popular because of large initiatives like the International Knockout Mouse Consortium (IKMC, https://www.mousephenotype.org/).

We hypothesized differences between inbred and outbred mouse in: 1) USV emission in both pups and adults; 2) social responses in adults; 3) qualitative USV patterns with the effect of age and social context.

2 | MATERIALS AND METHODS

2.1 | Animals and housing

C57BL/6N (inbred, hereinafter B6), FVB/NHan™Hsd (inbred, hereinafter FVB) and CD-1 (outbred) breeding pairs were purchased from Harlan Laboratories (S. Pietro al Natisone, Italy) and bred in our mouse facility. Mice were housed on a reversed 12:12 h light: dark cycle (lights on at 19:00 h) in standard wire-topped polycarbonate cages ages (33 cm × 13 cm × 14 cm) with sawdust bedding and water and food (DP/1000, Altromin-Rieper, Vandoies-BZ, Italy) ad libitum. Temperature was maintained at 21 ± 1°C, and relative humidity at 60 ± 10%. Females were individually housed and subsequently daily inspected for pregnancy and delivery 10 days after mating. The day of birth was considered as postnatal day (pnd) 0. Pups were tattooed on the paw with animal tattoo ink (Ketchum permanent Tattoo Inks green paste, Ketchum Manufacturing Inc., Brockville ON Canada) by subcutaneous injection (30G needle) into paw plantar surface. The procedure was performed at 2 days of age, immediately after behavioral testing.

Subject mice for adult social interaction tests were weaned into cages of same sex pairs. After weaning on postnatal day 21, each animal was socially housed with two same-sex partners per cage. Mice were 2-month-old B6 ($N = 12$ males and 12 females), FVB ($N = 12$ males and 12 females) and CD1 ($N = 12$ males and 12 females) when tested for social interaction tests. Behavioral testing was always conducted between 9.30 and 13.30 h, during the dark phase of the circadian cycle, under red light.

All procedures were conducted in compliance with the European Communities guidelines (EC Council Directive 63/2010), Italian legislation on animal experimentation (DL 26/2014).

2.2 | Ultrasonic vocalizations in pups

Tested litters contained more than seven pups. Within each litter, one male and one female underwent behavioral testing: B6 ($N = 8$ males and 8 females), FVB ($N = 10$ males and 10 females) and CD-1 ($N = 10$

males and 10 females). The remaining pups (not tested as neonates) were pooled at weaning and assigned to adult social interaction tests (described in the following sections), as well as to other experimental designs (i.e., USV playback studies). Ultrasonic vocalization, body weight, and body temperature of pups were measured at pnd 2, 4, 6, 8, and 12. These pnds were chosen to be in accordance with previous studies focused on the ontogenetic profile of USV emission in inbred and outbred mouse strains.[36,43,44] On each day of testing, the pup was placed into an empty glass container (diameter, 5 cm; height 10 cm), located inside a sound-attenuating styrofoam box, in a room under red light, and assessed for ultrasonic vocalizations during a 3-min test. At the end of the recording session, each pup was weighed, and its axillary temperature measured by gentle insertion of the thermal probe in the skin pocket between upper foreleg and chest of the animal for about 30 s (Microprobe digital thermometer with mouse probe, Stoelting Co., IL). When the pup was returned to the nest, the mother and littermates were present.

An Ultrasound Microphone (Avisoft Ultra Sound Gate condenser microphone capsule CM16, Avisoft Bioacoustics, Berlin, Germany) sensitive to frequencies of 10–180 kHz was placed through a hole in the middle of the cover of the styrofoam sound-attenuating box, about 20 cm above the pup in its plastic container. Room temperature was maintained at 22 ± 1°C. Vocalizations were recorded using Avisoft Recorder software (Version 3.2). Settings included sampling rate at 250 kHz; format 16 bit.

2.3 | Adult social interaction tests

Within each strain, 2-month-old mice (not previously tested as neonates, to exclude any potential confounders on adult behavior) were evaluated in two different social interactions: 1) male–female ($N = 12$); 2) female–female ($N = 12$). Male and female mice were weighed the same day of the test (mean ± *SD* of B6, FVB and CD-1 males are respectively: 21.93 ± 1.44; 25.58 ± 1.39; 33.51 ± 1.64; mean ± *SD* of B6, FVB and CD-1 females are respectively: 19.25 ± 0.62, 20.0 ± 0.98, 24.55 ± 1.43). Behavioral tests were conducted under red light, videotaped using a Panasonic monochrome CCD camera and subsequently analyzed with Observer 10XT software (Noldus Information Technology, NL). The cage contained sawdust (1.5-cm deep) and the lid was removed during the test.

For a 3-min session of female–female interaction test, an unfamiliar stimulus mouse was placed into the home cage of a subject mouse who had resided in the cage for the previous 5 days without enrichment materials. In the male–female interaction test, a group-housed male was used as subject mouse and the 3-minute test session was conducted in a clean cage with clean bedding (1.5-cm deep sawdust layer), representing a novel situation for both male subject and female partner. The videocamera was mounted facing the side of the cage to record the session for subsequent scoring of social investigation parameters. The ultrasonic microphone (same as in pup vocalization experiment) was mounted 20 cm above the floor of the cage to record the session.

Stimulus mice were matched to the subject mice by strain, sex, age, and body weight. Stimulus mice were bred in our colony as described above, and maintained in social groups of three per cage. On the day of male–female testing, the vaginal estrous condition of each stimulus female was assessed as previously described.[45] Only females in estrus were selected for the test. A total of 72 stimulus mice ($N = 24$ for each strain) were employed.

Social interactions were scored from the videotapes for the frequencies and durations of the following behavioral responses performed by the subject mouse: anogenital sniffing (direct contact with the anogenital area), body sniffing (sniffing or snout contact with the flank area), nose to nose sniffing (sniffing or snout contact with the head/neck/mouth area), locomotor activity by line crossings, rearing up against the wall of the home cage, digging in the bedding, and grooming (self-cleaning, licking any part of its own body). Vocalizations were recorded using Avisoft Recorder software version 3.2. Settings included sampling rate at 250 kHz; format 16 bit.

2.4 | Ultrasonic vocalization analysis

For acoustical analysis, recordings collected from pups and adults were transferred to Avisoft SASLab Pro (Version 4.40) and a fast Fourier transformation (FFT) was conducted as previously described.[46] Spectrograms from pup vocalizations were generated with an FFT-length of 1024 points, while adult vocalizations requested spectrograms with an FFT-length of 512 points and a time window overlap of 75% (100% Frame, Hamming window). The spectrogram was produced at a frequency resolution of 488 Hz and a time resolution of 1 ms. A lower cut-off frequency of 20 kHz was used to reduce background noise outside the relevant frequency band to 0 dB. Parameters analyzed for each test day included number of calls, duration of calls, qualitative and quantitative analyses of sound frequencies measured in terms of frequency and amplitude at the maximum of the spectrum.

Start times for the video and audio files were synchronized during social encounters. However, it was not possible to synchronize scoring of behaviors with calls using the currently available recording technology. The software used for the behavioral (Noldus, Observer X) and spectrographic (Avisoft Bioacoustics, Avisoft SASLabPro version 4.40) analyses cannot be combined on the same screen because they proceed with different speeds: behavioral events occurred in a time frame of seconds whereas vocalizations occurred in a time frame of milliseconds.

Waveform patterns of calls for pups and adults were examined in depth in the sonograms collected from each subject tested. Each call was identified as one of nine distinct categories, based on internal pitch changes, lengths, and shapes, as described below: 1) Complex calls displayed one component containing two or more directional changes in pitch, each ≥6.25 kHz; 2) Two-component calls consisted of two components: a main call (flat or downward) with an additional punctuated component towards the end; 3) Upward-modulated calls exhibited a continuous increase in pitch that was ≥12.5 kHz, with a terminal dominant frequency at least 6.25 kHz more than the pitch at

the beginning of the vocalization; 4) Downward-modulated calls exhibited a continuous decrease in pitch that was ≥12.5 kHz, with a terminal dominant frequency at least 6.25 kHz less than the pitch at the beginning of the vocalization; 5) Chevron calls resembled an "inverted-U," which was identified by a continuous increase in pitch ≥12.5 kHz followed by a decrease that was ≥6.25 kHz; 6) Short calls were punctuated and shorter than 5 ms; 7) Composite calls were formed by two harmonically independent components, emitted simultaneously; 8) Frequency steps were instantaneous frequency changes appearing as a vertically discontinuous "step" on a spectrogram, but with no interruption in time; 9) Flat calls displayed a constant beginning and the ending of the pitch frequency remained constant (≤3 kHz of each other).[46,47]

We classified pup vocalizations according to strain and sex: 1820 B6 calls ($N = 1070$ emitted by males and $N = 750$ by females); 3416 FVB calls ($N = 1490$ emitted by males and $N = 1926$ by females) and 8501 CD-1 calls ($N = 4744$ emitted by males and $N = 3757$ by females). All pups except 6 mice ($N = 2$ B6, $N = 2$ FVB, $N = 2$ CD-1) vocalized at least at one time point (pnd). Data related to pup vocalizations were subjected to three different analyses: a) strain-dependent effects on the frequency and duration of the vocalizations emitted by each subject at pnd 2, 4, 6, 8, 12; b) strain-dependent effects on the probability of producing calls (proportion of calls in each category for each subject) from each of the nine categories of USVs; c) a descriptive analysis which included strain-dependent effects on the percentage of calls emitted by each subject in each of the nine categories of USVs within and between postnatal days.

Waveform patterns of adult calls were examined in depth in the sonograms collected from each subject tested, using the classification based on nine call categories (see above pup analysis). In the female–female encounter, we classified: 2418 B6 calls ($N = 11$ subjects), 12,891 FVB calls ($N = 12$ subjects) and 8707 CD-1 calls ($N = 12$ subjects); in the male–female encounter, we classified: 4442 B6 calls ($n = 10$), 8667 FVB calls ($n = 12$) and 5343 CD-1 calls ($N = 12$). The rest of the subjects were not analyzed because they did not emit vocalizations. Inter-rater reliability was 98% between the two investigators who scored the call categories. Call category data were subjected to three different analyses: a) strain-dependent effects on the frequency and duration of the vocalizations emitted by each adult subject; b) strain-dependent effects on the probability of producing calls (proportion of calls in each category for each subject) from each of the nine categories of USVs; c) a descriptive analysis which included strain-dependent effects on the percentage of calls emitted by each subject in each of the nine categories of USVs within and between social encounters.

2.5 | Statistical analysis

A mixed-model Analysis of Variance (ANOVA) with Repeated Measures was used to analyze 1) body weight and body temperature of pups with the strain as factor and postnatal days as the repeated measures; 2) neonatal USV quantitative parameters with the strain as factor and postnatal days as the repeated measures; 3) adult social behaviors with the strain as factor and sniffing of different body areas as the repeated measures; 4) probability of vocalizations with the strain as factor and social context as the repeated measures. Probability of vocalizations within each strain was calculated as number of calls in each category for each subject/total number of calls analyzed in each subject and standardized by angular transformation. As the analysis of sonographic patterns is of an exploratory nature (and not confirmatory), we did not adjust the results for multiple testing. A one-way analysis of variance (ANOVA) was used to analyze adult USV quantitative parameters in each social context. An Analysis of Covariance (ANCOVA) on USV rate and duration with body weight and temperature as covariates were performed to investigate more deeply differences among strains.

To compare variability in inbred and outbred strains, both neonatal (pnd 8, paralleling with previous data) and adult USV data (rate and duration) were also analyzed by a nonparametric analysis of variance (Kruskal–Wallis), considering as variable of interest not measurements per se but their individual deviation from the average within each group [i.e. individual difference (in absolute value) between individual value and mean value of USV rate or duration within the experimental group considered].[48]

Post-hoc comparisons were performed using Tukey's HSD test when a significant F value was determined. For all comparisons, significance was set at $p < 0.05$.

3 | RESULTS

3.1 | Pups

B6 pups had lower body weight than FVB and CD1 pups ($p = 0.01$ and $p < 0.01$, respectively). All pups showed an increased body weight from pnd 2 to 12 ($p < 0.001$) (Figure 1A).

B6 pups had lower body temperature than FVB and CD-1 pups ($p < 0.01$). All pups increased body temperature from pnd 2 to pnd 12 ($p < 0.001$). Only B6 pups had a lower temperature at pnd 2 in comparison to other two strain pups ($p < 0.01$) (Figure 1B).

USV rate changed from pnd 2 to 12 ($p < 0.001$). B6 pups emitted a lower USV rate than FVB and CD-1 pups ($p = 0.05$ and $p < 0.01$, respectively). Posthoc comparisons (performed on the significant interaction strain × pnd) confirmed that B6 pups vocalized less than CD-1 pups at pnd 4 ($p < 0.05$), 6, 8 ($p < 0.01$) and 12 ($p < 0.05$), and less than FVB pups only at pnd 8 ($p < 0.05$). Also, FVB pups vocalized less than CD-1 pups at pnd 4, 6, 8, and 12 ($p < 0.01$) (Figure 1C). No significant strain differences were detected on peak frequency and peak amplitude of USVs.

USV duration changed from pnd 2 to 12 ($p < 0.001$). FVB and CD-1 pups emitted calls longer than B6 pups ($p < 0.01$). Posthoc comparisons (performed on the significant interaction strain x pnd) revealed that both FVB and CD-1 calls were longer than B6 calls at pnd 2, 4, 6, 8 ($p < 0.01$), and 12 ($p < 0.05$) (Figure 1D).

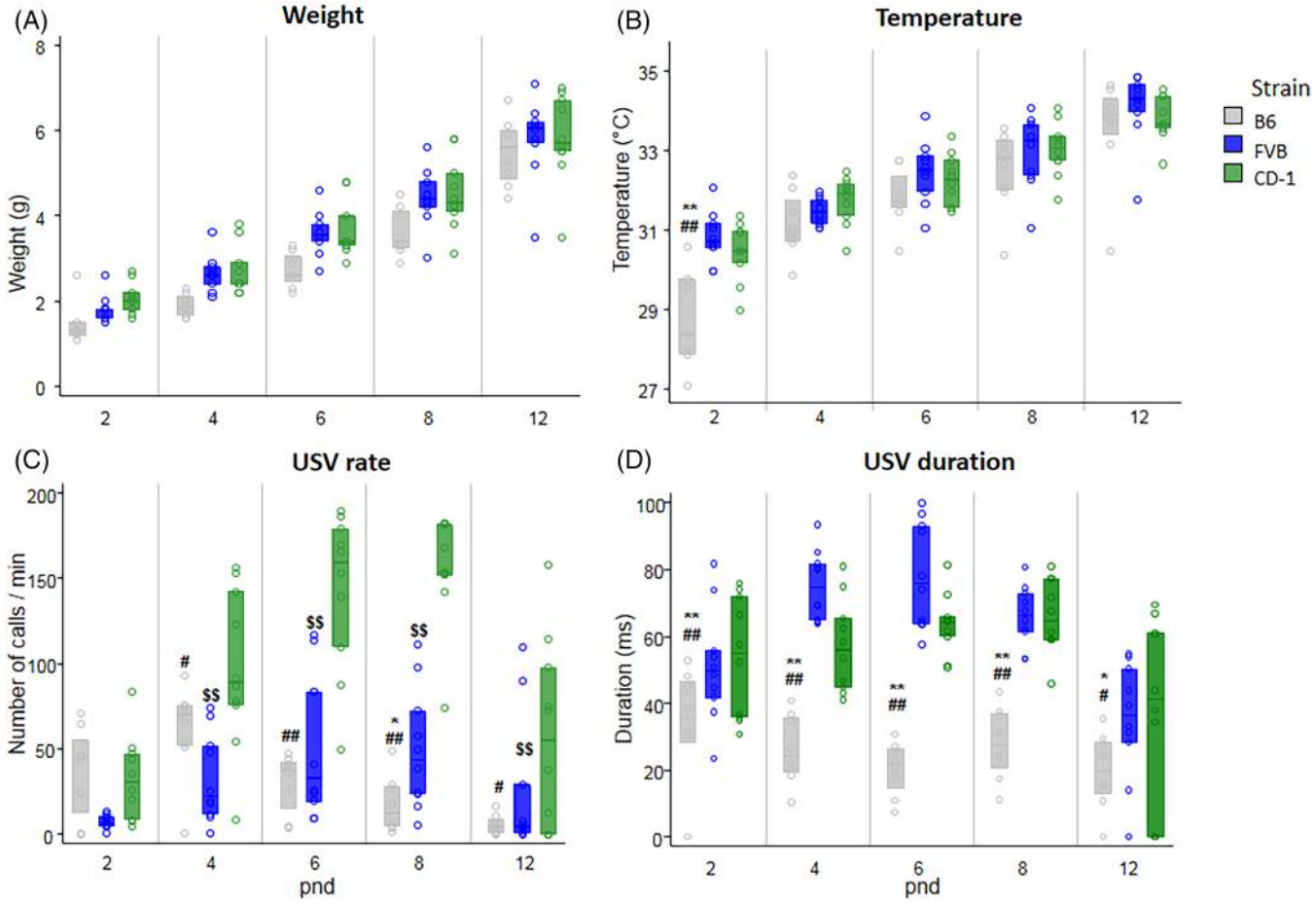

FIGURE 1 Neonatal body weight (A), body temperature (B), ultrasonic vocalization (USV) rate (C) and duration (D) in B6 ($n = 16$), FVB ($n = 20$), and CD-1 pups ($n = 20$). Neonatal USVs at postnatal day (pnd) 2, 4, 6, 8, and 12 are emitted by pups in response to social isolation (3-minute session). Significant difference between B6 and FVB ($p < 0.05^{*}$, $p < 0.01^{**}$); B6 and CD-1 ($p < 0.05^{\#}$, $p < 0.01^{\#\#}$); FVB and CD-1 ($p < 0.01^{\$\$}$). Data are expressed as median ± 1st and 3rd interquartile.

TABLE 1 Statistical analysis of neonatal body weight, body temperature and USV data (from pnd 2 to 12)

	Main effect of strain	Main effect of pnd	Interaction of strain × pnd
Body weight	$F(2, 25) = 5.26$, $p = 0.01$	$F(2, 4) = 333.58$, $p < 0.001$	$F(8, 100) = 0.473$, NS
Body temperature	$F(2, 25) = 7.33$, $p < 0.01$	$F(2, 4) = 107.99$, $p < 0.001$	$F(8, 100) = 3.76$, $p < 0.01$
USV rate	$F(2, 25) = 52.47$, $p < 0.01$	$F(2, 4) = 14.47$, $p < 0.001$	$F(8, 100) = 7.76$, $p < 0.01$
USV duration	$F(2, 25) = 109.96$, $p < 0.01$	$F(2, 4) = 17.35$, $p < 0.001$	$F(8, 100) = 8.76$, $p < 0.01$
Peak frequency of USVs	$F(2, 25) = 42.47$, NS	$F(2, 4) = 4.53$, NS	$F(8, 100) = 12.23$, NS
Peak amplitude of USVs	$F(2, 25) = 19.86$, NS	$F(2, 4) = 18.35$, NS	$F(8, 100) = 8.35$, NS

Abbreviation: NS, not statistically significant.

Analysis of covariance ruled out the possibility that the strain effects found for USV rate and duration were because of the differences in body weight or body temperature (main effect of strain and strain × pnd interaction were still significant when body weight and body temperature were used as covariates in the repeated measure design). Detailed statistical analysis are reported in Table 1.

3.2 | Adult social interaction tests

3.2.1 | Male–female

During the 3-min interaction of a male with a sexually receptive female, FVB males emitted a higher USV rate than B6 and CD-1 males ($p < 0.01$ and $p < 0.05$, respectively). FVB calls were also longer than

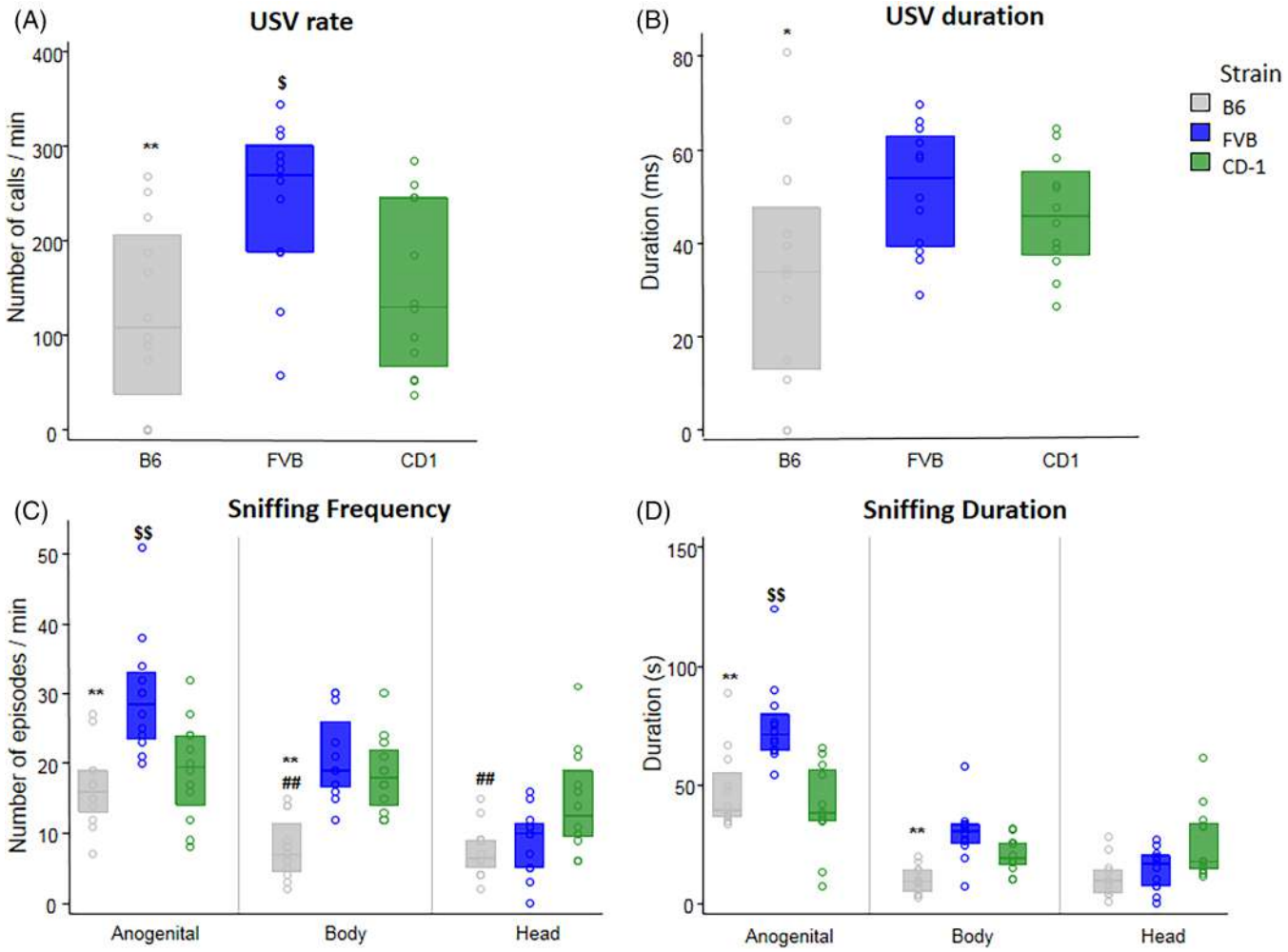

FIGURE 2 Male–female social interaction. A 3-min session measured parameters of direct interaction of a male ($n = 12$ for each strain) with a sexually receptive female of the same strain. USV rate (A) and duration (B), sniffing frequency (C), and sniffing duration (D). Significant differences between B6 and FVB ($p < 0.05^{*}$, $p < 0.01^{**}$); B6 and CD-1 ($p < 0.01^{\#\#}$); FVB and CD-1 ($p < 0.05^{\$}$, $p < 0.01^{\$\$}$). Data are expressed as median ± 1st and 3rd interquartile.

B6 ones ($p < 0.05$) (Figure 2A,B). No significant strain differences were detected on peak frequency and peak amplitude of USVs.

Analysis of the frequency of social sniffing showed that FVB and CD-1 males had higher values than B6 males ($p < 0.01$). Posthoc comparisons (performed on the significant interaction strain × body area) confirmed that B6 and FVB males sniffed the corresponding female partner more frequently in the anogenital area versus body and head area ($p < 0.01$), whereas in CD-1 males frequency of sniffing did not vary according to the different body areas. Moreover, FVB males sniffed the anogenital area of the partner more frequently than B6 and CD-1 males ($p < 0.01$). FVB and CD-1 males sniffed the body of the partner more than B6 males ($p < 0.01$), whereas CD-1 males sniffed the head of the partner more frequently than B6 males ($p < 0.01$) (Figure 2C).

Analysis of the duration of social sniffing provided a similar picture. FVB sniffing response was longer than the one of B6 and CD-1 ($p < 0.01$). Also sniffing duration was longer for anogenital area versus body and head area within each strain ($p < 0.01$). Posthoc comparisons (performed on the significant interaction strain x body area) reported that FVB males sniffed the anogenital area of the partner longer than B6 and CD-1 ($p < 0.01$) and the body area of the partner longer than B6 males ($p < 0.01$) (Figure 2D). Detailed statistical analysis are reported in Table 2.

3.2.2 | Female–Female

During the 3-min interaction of a female with a same sex conspecific, FVB females emitted a higher USV rate than B6 and CD-1 females, and CD-1 females emitted a higher USV rate than B6 ones ($p < 0.01$) (Figure 3). No significant strain differences were detected on USV duration (Figure 3B), peak frequency and peak amplitude of USVs.

Analysis of frequency of social sniffing revealed significant differences among strains and body areas. B6 females investigated the female partner less frequently than CD-1 and FVB females ($p < 0.01$). Females sniffed more frequently the anogenital area of the female

TABLE 2 Statistical analysis of adult data (USVs and social response) during social interaction tests

	Main effect of strain	Main effect of body area	Interaction of strain x body area
Male–Female			
USV rate	$F(2, 33) = 5.46$, $p < 0.01$		
USV duration	$F(2, 33) = 3.17$, $p = 0.05$		
Peak frequency of USVs	$F(2, 33) = 0.46$, NS		
Peak amplitude of USVs	$F(2, 33) = 2.54$, NS		
Frequency of social sniffing	$F(2, 33) = 15.52$, $p < 0.01$	$F(2, 66) = 44.37$, NS	$F(4, 66) = 8.15$, $p < 0.01$
Duration of social sniffing	$F(2, 33) = 17.08$, $p < 0.01$	$F(2, 66) = 88.46$, $p < 0.01$	$F(4, 66) = 9.41$, $p < 0.01$
Female–Female			
USV rate	$F(2, 33) = 45.51$, $p < 0.01$		
USV duration	$F(2, 33) = 2.78$, NS		
Peak frequency of USVs	$F(2, 33) = 2.15$, NS		
Peak amplitude of USVs	$F(2, 33) = 3.43$, NS		
Frequency of social sniffing	$F(2, 33) = 8.61$, $p < 0.01$	$F(2, 66) = 91.88$, $p < 0.01$	$F(4, 66) = 9.19$, $p < 0.01$
Duration of social sniffing	$F(2, 33) = 24.27$, $p < 0.01$	$F(2, 66) = 81.13$, $p < 0.01$	$F(4, 66) = 11.67$, $p < 0.01$

Abbreviation: NS, not statistically significant.

partner than the head area within each strain ($p < 0.01$). Posthoc analysis (performed on the significant interaction strain x body area) revealed that only CD-1 females spent more time sniffing the anogenital than the body area ($p < 0.01$). FVB and CD-1 females sniffed the anogenital area of their partner more than B6 ($p < 0.01$); FVB females also sniffed the body of their partner more frequently than B6 females ($p < 0.01$) (Figure 3C).

Analysis of duration of social sniffing response revealed significant differences among strains and body areas. B6 females spent less time investigating the female partner than CD-1 and FVB females, and CD-1 females spent less time in social investigation than FVB females ($p < 0.01$). Post hoc analysis (performed on the significant interaction strain × body area) revealed that CD-1 and FVB females spent more time sniffing the anogenital area of the female partner than the body and the head, whereas only FVB females spent more time sniffing the body area than the head ($p < 0.01$). FVB and CD-1 females also sniffed the anogenital area of their partner longer than B6 females, and FVB also longer than CD-1 females ($p < 0.01$); FVB sniffed the body of their partner longer than B6 females ($p < 0.01$) (Figure 3D). Detailed statistical analysis are reported in Table 2.

Data concerning comparison of variability of USV rate and USV duration of inbred (B6 and FVB) versus outbred (CD-1) strains were analyzed using nonparametric Kruskal–Wallis test. The variable considered for each individual data was the difference (absolute value) between individual datum and mean value for the group (e.g., for each B6 male pup, the 8/12 dots correspond to individual USV rate minus the mean value of USV rate of male B6 pups). As shown in Figure S1, only in the female–female social interaction test, variability of CD-1 females resulted significantly higher ($p < 0.01$) than B6 females; no other difference in variability emerged in the other comparisons.

3.3 | Pattern of sonographic structure among strains in different social contexts

Figure 4 describes the percentages of different call categories emitted by male and female pups and adults for each strain. B6, FVB and CD-1 mice emitted a different spectrum of call categories.

At postnatal day 8, B6 pups preferred emitting short (males 30%, females 27%) and two-component (males 21%, females 29%) calls, with a reduced proportion of complex and downward calls. By contrast, FVB displayed high prevalence in production of complex (males 35%, females 29%), frequency steps (males 40%, females 38%), and composite (males 11%, females 13%) calls. Similarly, CD-1 displayed high prevalence in production of complex (males 27%, females 31%), frequency steps (males 30%, females 28%), along with two-component (males 27%, females 28%) calls.

During adult male–female interaction, B6 preferred emitting two-component (36%) and complex (25%) calls, with a reduced proportion of upward (16%) and short (13%) calls. FVB produced a consistent number of complex (42%), frequency steps (24%), and upward calls (21%). CD-1 emitted predominantly a type of nine call categories, showing 63% of complex calls, with a small number of upward (16%) and two-component (10%) calls.

During adult female–female interaction, B6 preferred emitting two-component (35%) and complex (23%) calls, with a reduced proportion of short (16%), frequency steps (11%) and upward (8%) calls. FVB produced a consistent number of frequency steps (43%), along with a similar proportion of complex (18%), two-component (16%) and upward (18%) calls. CD-1 emitted predominantly two types of call categories: 49% of complex and 28% of frequency-steps calls.

When analyzing each USV category separately (see Figure 5), the probability of producing defined call categories differed across strains and social contexts (neonatal stage, adult male–female and female–

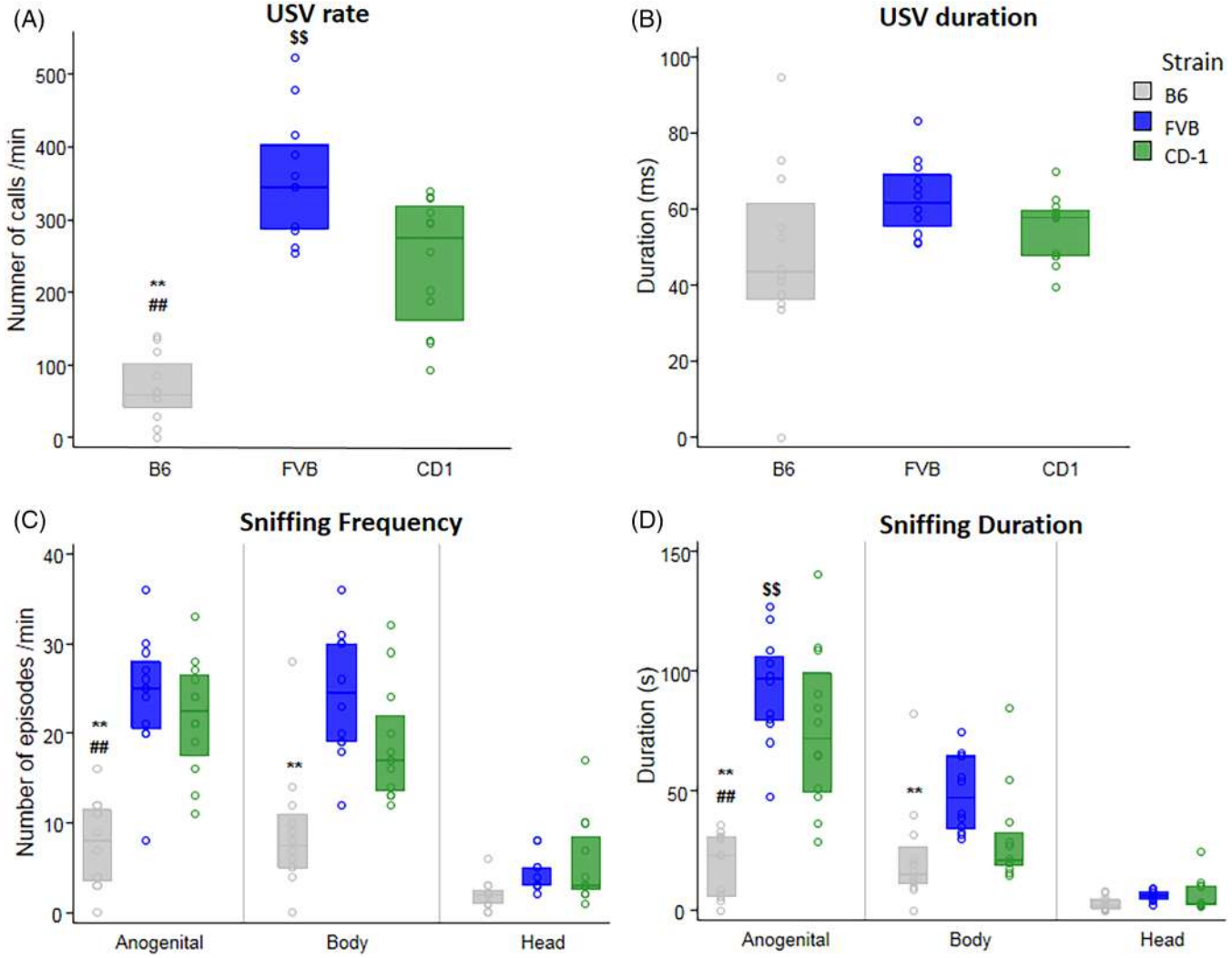

FIGURE 3 Female–female social interaction. A 3-min session measured parameters of interaction of a resident female ($n = 12$ for each strain) with an unfamiliar female of the same strain. USV rate (A) and duration (B), sniffing frequency (C), and sniffing duration (D). Significant differences between B6 and FVB ($p < 0.05^{*}$, $p < 0.01^{**}$); B6 and CD-1 ($p < 0.05^{\#}$, $p < 0.01^{\#\#}$); FVB and CD-1 ($p < 0.05^{\$}$, $p < 0.01^{\$\$}$). Data are expressed as median ± 1st and 3rd interquartile.

female interaction). Analysis revealed: 1) a main effect of strain on the proportion of eight call categories (complex, two-components, downward, chevron, short, composite, frequency steps, and flat); 2) a main effect of social context on the proportion of seven call categories (complex, upward, downward, chevron, composite, frequency steps, and flat; 3) a significant strain × social context interaction on the proportion of eight call categories (complex, two-components, downward, chevron, short, composite, frequency steps, and flat).

Detailed statistical analysis are reported in Table 3. Call differences between strains are also described in the Supporting Information, Table S1.

As for individual data variability across inbred and outbred strains, our analyses of absolute values of differences from mean group values identify significant differences only within the female–female social interaction test (Kruskal–Wallis test: $H = 6.164$, $p < 0.0459$), namely CD-1 female data show higher variability than B6 data (Mann–Whitney test: $U = 16$, $p = 0.0037$) whereas no evidence of significant difference in variability was detected within the other groups (Figure S1).

4 | DISCUSSION

Our findings corroborated results from previous studies showing that strain or genetic background, age, sex, as well as responsiveness to the environmental stimuli (i.e., body temperature) influence the social mouse behaviors, specifically how mice communicate and socially interact with conspecifics.[28,46,49–51] In the current study, differences among three strains emerged both in the USV rate and vocal repertoire (according to spectrographic features of each call), as well as in social performances.

We initially focused on vocal differences among strains at neonatal stages. This assessment is crucial considering that USV analysis is

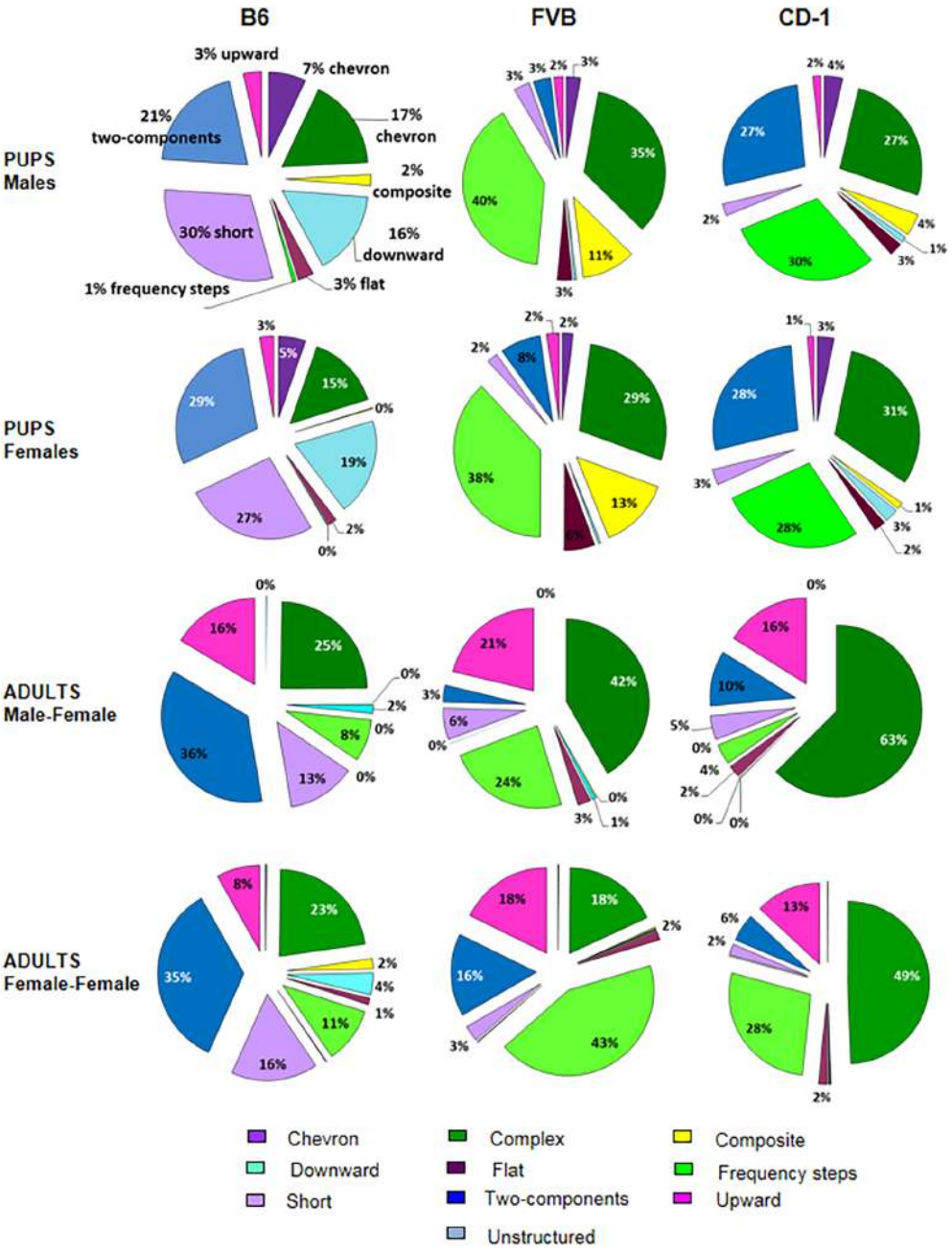

FIGURE 4 Distributions of call categories in pup B6, FVB and CD-1 (at pnd 8) and adult B6, FVB and CD-1 during each type of social encounter. Pie graphs display the percentages of the different call categories within each strain for male pups at pnd 8, female pups at pnd 8, adult male–female, and female–female social interactions. Percentages were calculated in each strain as number of calls in each category for each subject/total number of calls analyzed in each subject.

one of the few assays that can be performed during the first postnatal weeks of life to define developmental trajectories. It also has high value from a translational perspective, since developmental trajectories are investigated in several mouse models of neurodevelopmental disorders, often generated on different genetic background.[1,52,53] We chose to analyze five time points rather than one single postnatal day, to verify potential strain differences in USV profile across the entire neonatal period. Testing a pup for a brief period (3 min session at each time point) could not be considered a stressful event, if compared with prolonged periods of maternal isolation (from 15 to 45 min/day).[54–56] Thus, in the absence of stressful conditions, it can be assumed that repeated assessments of pups did not substantially affect their USV performance during development. Our analyses detected differences already at early stages for both USV rate and duration: CD-1 pups emitted a higher number, while B6 pups a lower number of calls compared with the other two strains, in line with previous studies.[10,43,57–60] When looking at the mean duration of calls, both FVB and CD-1 calls were longer than B6. Thus, B6 pups significantly produced a reduced number of calls and those emitted were also shorter than FVB and CD-1,[43,44] as also confirmed by the analysis of sonographic structure (see below). In comparison to B6, FVB pups produced longer USVs, in line with previous data collected on those strains.[36] The ontogenetic profile of USV rate also showed strain differences: B6 pups had a peak in the calling rate earlier, around postnatal day 4, while CD-1 and FVB pups around pnd 6–8.[2,32,57–59,61] As Figure 1C depicted, the large

degree of differences in USV rate emission at pnd 4, 6, and 8 is because of the outbred CD-1 pups. In response to social isolation and with the aim to activate maternal care, the increased USV emission of CD-1 pups could be considered as an index of more anxiety-like behavior in comparison to "low anxiety" profile of inbred B6 and FVB pups.[62] Several factors could be seen as "confounders" in ultrasonic emission, such as body temperature and weight.[63] In our analysis, strain differences in rate and duration of vocalizations were somehow paralleled by differences in temperature and weight. However, analyses of co-variance, with either temperature or body weight as covariates, indicated that strain differences were still well detectable, ruling out the possibility that the different patterns shown by B6 pups was substantially because of lower values of body weight and temperatures.

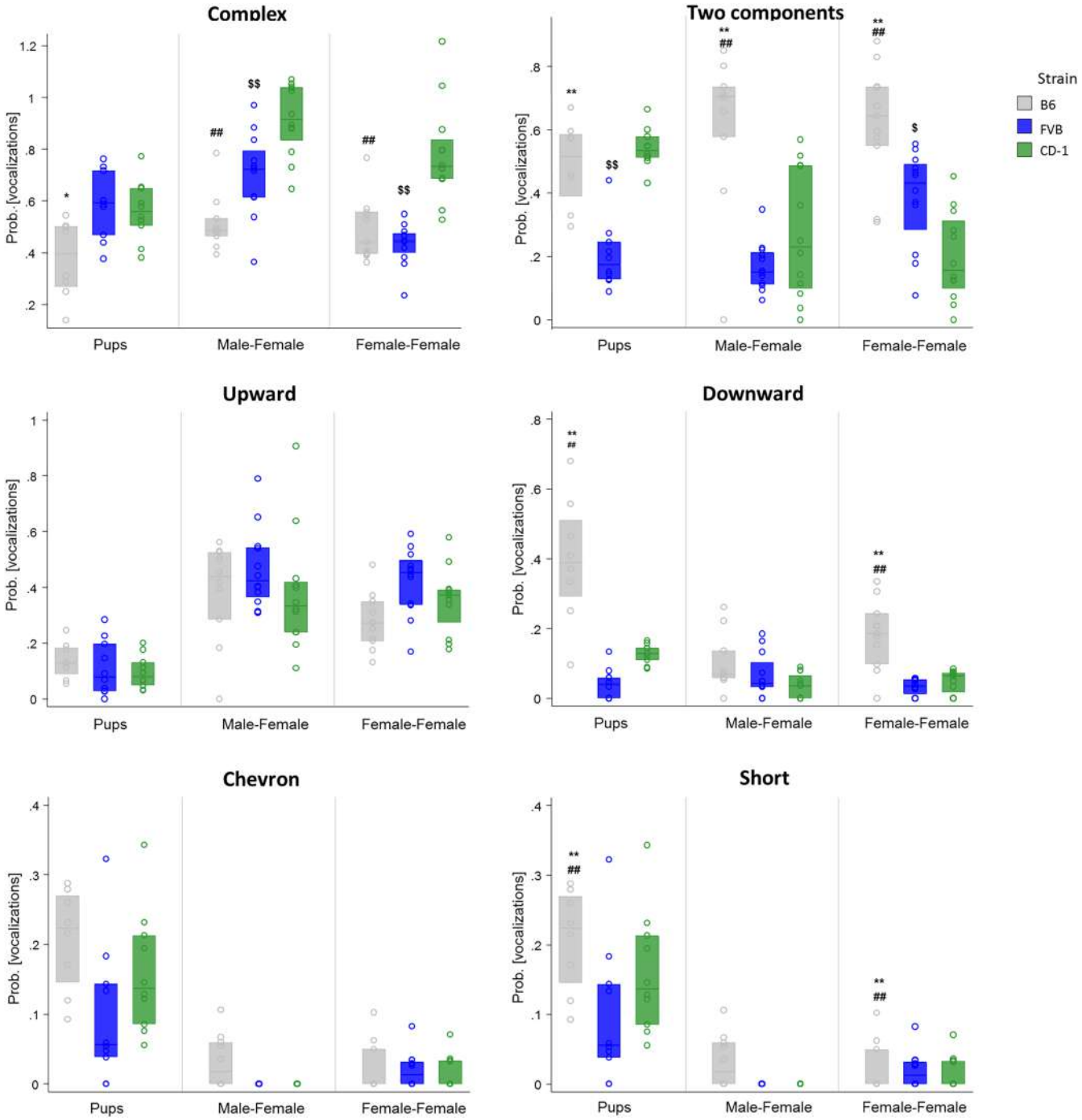

FIGURE 5 Production of calls across strains. Probability of producing calls from each of the nine categories of USVs at pnd 8 (pups), during adult male–female and female–female social interactions. Data are expressed by angular transformation (y-axes are different for each graph). Significant differences between B6 and FVB ($p < 0.05^{*}$, $p < 0.01^{**}$); B6 and CD-1 ($p < 0.01^{\#\#}$); FVB and CD-1 ($p < 0.05^{\$}$, $p < 0.01^{\$\$}$)

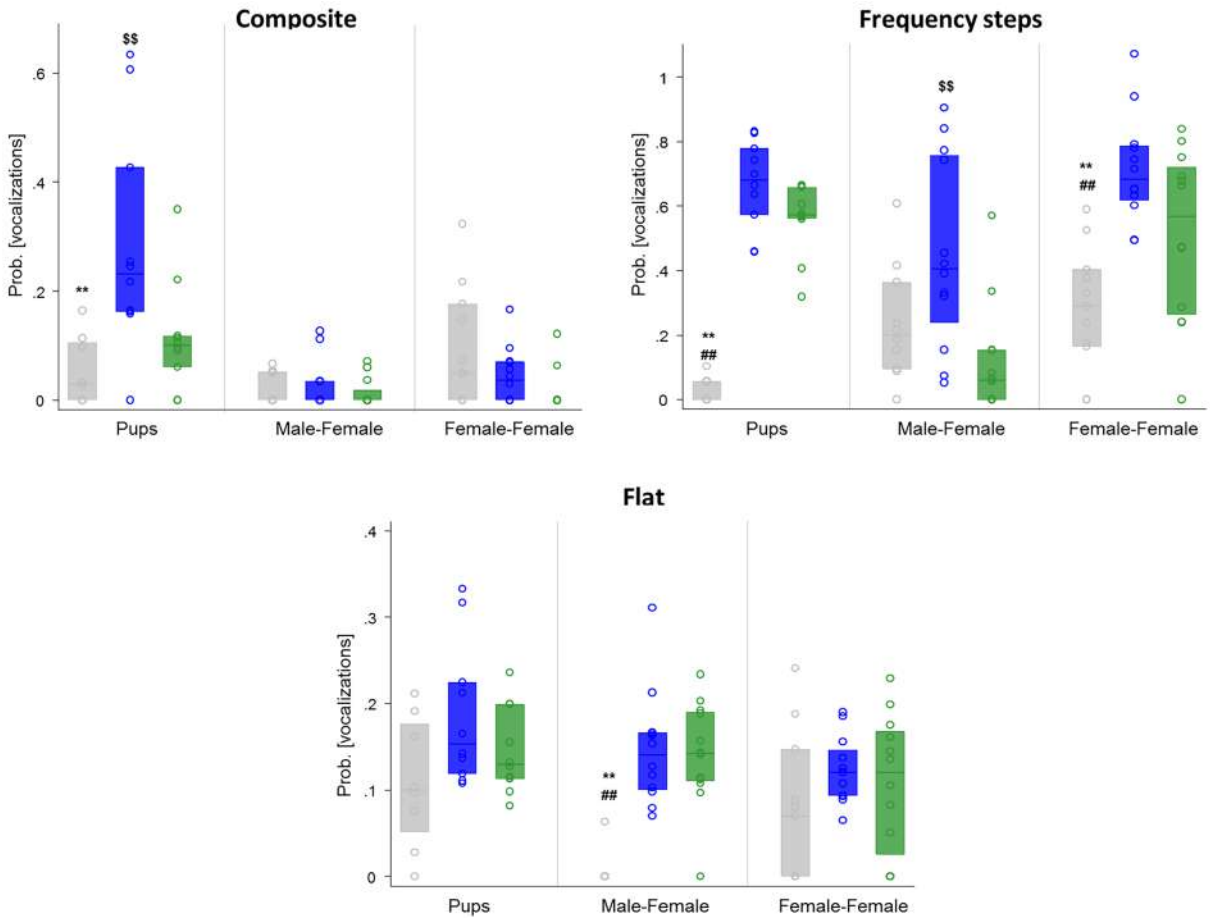

FIGURE 5 (Continued)

TABLE 3 Statistical analysis of call categories across strains and social contexts (neonatal stage, adult male–female and female–female interaction)

	Main effect of strain	Main effect of social context	Interaction of strain x social context
Complex	$F(2, 88) = 36.48$, $p < 0.01$	$F(2, 88) = 17.81$, $p < 0.01$	$F(4, 88) = 6.58$, $p < 0.01$
Two-components	$F(2, 88) = 35.35$, $p < 0.01$	$F(2, 88) = 45.70$, NS	$F(4, 88) = 10.40$, $p < 0.01$
Upward	$F(2, 88) = 45.11$, NS	$F(2, 88) = 40.67$, $p < 0.01$	$F(4, 88) = 11.74$, $p < 0.01$
Downward	$F(2, 88) = 52.21$, $p < 0.01$	$F(2, 88) = 22.70$, $p < 0.01$	$F(4, 88) = 13.22$, NS
Chevron	$F(2, 88) = 6.46$, $p < 0.01$	$F(2, 88) = 73.20$, $p < 0.01$	$F(4, 88) = 3.12$, $p < 0.05$
Short	$F(2, 88) = 39.40$, $p < 0.01$	$F(2, 88) = 86.55$, NS	$F(4, 88) = 4.67$, $p < 0.01$
Composite	$F(2, 88) = 6.16$, $p < 0.01$	$F(2, 88) = 17.65$, $p < 0.01$	$F(4, 88) = 7.17$, $p < 0.01$
Frequency steps	$F(2, 88) = 37.49$, $p < 0.01$	$F(2, 88) = 12.35$, $p < 0.01$	$F(4, 88) = 6.92$, $p < 0.01$
Flat	$F(2, 88) = 15.34$, $p < 0.01$	$F(2, 88) = 5.09$, $p < 0.01$	$F(4, 88) = 2.37$, $p = 0.05$

Abbreviation: NS, not statistically significant.

Our second investigation targeted the adult vocal profile, in association to social behavior. We found that FVB males vocalized more and for longer time than B6 in presence of a sexually receptive female, as pups did at neonatal age for activating maternal care. The increased USV emission of FVB males was associated with increased social investigation (both frequency and duration) of the female partner, focusing on the anogenital area. In a similar manner, FVB females expressed high levels of social behavioral responses, since they vocalized more than B6 and CD-1, displayed more frequent and longer anogenital and body area sniffing than B6, as well as longer anogenital sniffing than CD-1. As for B6 strain, both males and females, as compared with FVB and CD-1 mice, conserved a low vocal and social

profile, differently from previous data found in adult male–male interactions.[64] Noteworthy, B6 has been extensively applied as genetic background to generate mutant mouse lines and used as "control" strain in several experiments, for example, in comparison with inbred strains with low social responsiveness.[65,66] It has been also considered as a suitable candidate for behavioral studies since episodes of aggressive attacks rarely occur during social interaction.[28,47,64,67] Present data indicate that B6 is the strain that has lower USV emission rate as neonate and adults (both sexes), and emit simpler and shorter calls. Thus, it might not be the preferential strain to choose in experimental settings and conditions that are hypothesized to reduce USVs and social responses. Our data also evidenced that CD-1 males had a lower production of USVs and a reduced interest to investigate the partner than FVB, and CD-1 females showed an intermediate profile for levels of USV emission and social investigation of the female partner when compared with both B6 and FVB strains.

The analysis of variability in the number of calls emitted by the three strains showed comparable levels of variability around mean values for each strain. This result substantially confirms and extends recent data showing that outbred CD-1 calls are not more variable than inbred FVB ones.[35] Such a convergence on comparable degree of variability in inbred and outbred USVs, using different methodological approaches and targeting different USV parameters (fairly counterintuitive on the basis of heterogeneity of genetic background), is worth of further comparisons of behavioral traits, not necessarily limited to vocalization patterns.[68] Nevertheless, present CD-1 USV profile suggests that this outbred strain can be suitable for both neonatal and adult studies focused on modulation (by different conditions or agents) of the vocal repertoire.

Further, we investigated more in detail both neonatal and adult vocal repertoires, classifying them into nine categories based on spectrographic appearance. For neonatal repertoire, we focused on postnatal day 8 to be consistent with previous data.[32,43,44] Differently from the other two strains, B6 pups (both sexes) emitted a significant portion of calls with a simple sonographic structure (short and downward) and also as adults maintained the highest percentage of short calls among the three strains. In addition, B6 pups (both sexes) also showed a remarkable percentage of two components (as CD-1 pups), and maintain this feature also in the adult vocal repertoire, whereas the same call category is no longer so prevalent in adult FVB and CD-1 mice. These features are in full agreement with our previous data.[43,46,47] The vocal repertoire of FVB mice depicted a different sonographic pattern: they emitted more composite calls (on postnatal day 8) and frequency steps calls (across social contexts) than the other strains, as also previously observed during both infancy and adolescence.[36,43] Similarly to FVB, CD-1 strain displayed a complex and modulated vocal repertoire in both neonatal and adult social contexts: CD-1 mice produced more two-component and frequency-step calls during neonatal age, while more complex and fewer frequency-step calls during adult social encounters. Branchi and colleagues already reported that 8-day-old CD-1 mice emitted a higher percentage of frequency steps and complex calls, along with low numbers of flat and short calls, although the spectrum of USVs was classified into five categories only (flat, complex, frequency steps, short, and composite).[32] In the current study, upward call represented the only call category that did not vary among strains at adulthood: B6, FVB and CD-1 mice emitted this type of call in a similar manner in both male–female and female–female social interaction, suggesting that upward calls production is a stable element of adult social interaction. These results are supported by a similar acoustic signature with a high proportion of upward calls in B6 adult mice,[69] and in wild-derived male house mice following urine exposure,[25] during male–female direct social interaction,[21] and after sexual priming.[23] All together, these findings highlight that inbred B6 mice produce simpler syllables with fewer internal changes and shorter duration, inbred FVB mice produce syllables with repeated internal changes (i.e.: more frequency-step and complex calls), while outbred CD-1 mice seem to have an intermediate profile.

Following mice from early age to the adulthood, it can be detected that pups displayed a wider vocal profile, based on six or more types of calls (i.e.: complex, two-component, frequency steps, composite, downward and short), while adults preferred to emit primarily four types of call categories (i.e., upward, complex, two-component and frequency steps). Mouse pups thus have a less defined vocal signature during early development and tend to define it with age, supporting the idea of a progressive change towards adult acoustic features and syllable composition.[52,70] At adulthood, a different prevalence of call categories is observed in each strain. Such strain-dependent patterns of call categories could affect mate-choice and/or probability to interact with a conspecific. Additional studies are needed to better understand the role and characteristics of vocal communication during social interactions (e.g., differences between courtship and mating or USV production) across strains, since few studies have detailed this aspect so far.[71,72]

It is worth highlighting that USVs are a useful tool for evaluating emotion and motivation in rodents.[73,74] In juvenile and adult rats, call categories have been extensively used as measures of emotional/affective state. Flat 22-kHz USVs indicate a negative affective state, while frequency 50-kHz USVs indicate a positive affective state.[75,76] To our knowledge, the meaning of each call category is not well-established in mice, although a main hypothesis is that the spectrotemporal call complexity may reflect motivational and emotional states. For example, the presence of complex and harmonic calls has been considered as a valuable index of positive emotions in mice.[77] Crucially, emotional and motivational aspects play a role in modulating the number of USVs and the type of call in several mouse models of psychiatric disorders, as autism spectrum disorders, schizophrenia, and stress-related disorders.[78] Comparing the use of call types by different strains, B6 mice disproportionally emitted higher pitch and more downward modulated calls than BALB mice during vigorous social approaches.[11] In our study, we supposed that more complex and modulated vocal repertoires depicted in FVB and CD-1 mice, in comparison to B6 ones, were functionally related to the strain-dependent differences in behavioral responses to social stimuli and environmental factors, as well as it appeared in other emotional behaviors (i.e., anxiety, stress).[79,80]

A first limitation to the current study is the difficulty to distinguish which mouse of the dyad vocalized during the adult social interaction with current available protocols and most common analysis tools. Previous investigations dealt with the issue to determine whether the source of call production is exclusively the mouse test.[16,81] We suppose that in our experimental setting, given the protocol we applied to record USVs, the test mouse (male, in the male–female interaction; resident female in the female–female interaction) is the primary vocal emitter. However, we cannot rule out that also the female partner vocalized during the social encounter. Indeed, it has been recently demonstrated, through a multiple-microphone array system that allows to identify the vocalizer of a group of mice, that during male–female interactions also female mice vocalize, to interact with the male and transfer social and receptive information[16]; even in these more naturalistic settings, however, greater proportion of vocalizations are produced by males. During our qualitative analysis carried out by visual inspection of spectrograms, we did not detect overlapping of signals; although we cannot exclude that it sporadically occurred, we can rule out that it systematically affected our measurements. A novel insight comes from a recent deep learning approach evidencing the differences between male and female USVs through the investigation of spectrotemporal properties. The full spectrogram characteristics were informative about the emitter's sex, at least during analysis of female–male social interaction.[82] Future and in-depth investigations including the use of multiple-microphones, source localization methods and machine learning approaches are needed to determine the location of the sound and identify the mice vocalizing in the dyadic social interaction.

A second limitation consists in the limited number (three) of strains and types of social contexts/age (pups during maternal isolation, adult males and females during social encounters) used. Future studies are needed to expand such specific vocal evaluation to other mouse strains and additional settings, possibly synchronizing USV recordings and social behavior in adult mouse.[83]

As a third limitation, we did not evaluate potential strain-related differences in mother's behavior after reunion and their effects on subsequent USV neonatal emission. Data from literature indicated that increased maternal responsiveness may lead to a reduction in "isolation-induced USVs."[10] C57BL/6 mothers exhibited higher levels of maternal responsiveness in comparison to BALB/c ones, which in turn were associated to a lower number of neonatal USVs.[10] Thus, in our study it is possible the dams belonging to different strains recognized distinct signals (i.e., olfactory) from a pup placed in the nest following the experimental session and they activated differently maternal care and maternal responsiveness.

5 | CONCLUSION

Our data fit with other investigations emphasizing that USVs carry relevant social information about species, strain, sex and individuals, and potentially vary in response to mouse internal state, social experience and behavioral interactions with conspecifics.[19,52,81,84,85]

Present USV quantitative and qualitative results demonstrate that there is a large degree of variance among B6, FVB and CD-1 mouse strains across age and social contexts. We conclude that the number of USVs, as well as their acoustic features and sound shapes, may be influenced by strain, age and social context of assessment. In association with USV detection, analyzing behavioral social responses during two types of adult encounters represent an additional item that capture more defined mouse social profiles.

Our findings illustrate the importance of considering strain, as well as age and social/environmental conditions of mice, prior to set up USV experimental paradigms. Also background strain of genetically modified lines should be taken into account when dealing with previous findings or planning pharmacological/toxicological experiments. The intermediate profile of the outbred CD-1 strain, compared with inbred B6 and FVB strains, should be considered when dealing with the choice of inbred and outbred strains appropriate to different contexts or experimental settings.

AUTHOR CONTRIBUTIONS

Laura Ricceri and Maria Luisa Scattoni conceived the hypothesis and designed the study; Laura Ricceri, Maria Luisa Scattoni, and Maria Adelaide Marconi performed experiments; Angela Caruso, Maria Adelaide Marconi, and Laura Ricceri analyzed and plotted the data; Angela Caruso, Laura Ricceri, and Maria Luisa Scattoni wrote the manuscript. All authors read and edited the manuscript.

ACKNOWLEDGMENTS

The authors thank Tommaso Salvitti for editing graphs; Luigia Cancemi for helping with animal maintenance and care. The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Maria Luisa Scattoni https://orcid.org/0000-0002-6659-0280

REFERENCES

1. Caruso A, Ricceri L, Scattoni ML. Ultrasonic vocalizations as a fundamental tool for early and adult behavioral phenotyping of autism Spectrum disorder rodent models. *Neurosci Biobehav Rev*. 2020;116:31-43. doi:10.1016/j.neubiorev.2020.06.011
2. Elwood RW, Keeling F. Temporal organization of ultrasonic vocalizations in infant mice. *Dev Psychobiol*. 1982;15(3):221-227. doi:10.1002/dev.420150306
3. Noirot E. Ultrasounds and maternal behavior in small rodents. *Dev Psychobiol*. 1972;5(4):371-387. doi:10.1002/dev.420050410
4. Zippelius HM, Schleidt WM. Ultraschall-Laute bei jungen Mausen. *Naturwissenschaften*. 1956;21:1-2.
5. Blumberg MS, Alberts JR. Ultrasonic vocalizations by rat pups in the cold: an acoustic by-product of laryngeal braking? *Behav Neurosci*. 1990;104(5):808-817. doi:10.1037//0735-7044.104.5.808
6. Blumberg MS, Sokoloff G. Do infant rats cry? *Psychol Rev*. 2001;108(1):83-95. doi:10.1037/0033-295x.108.1.83

7. Cohen-Salmon C, Carlier M, Roubertoux P, Jouhaneau J, Semal C, Paillette M. Differences in patterns of pup care in mice. V--pup ultrasonic emissions and pup care behavior. *Physiol Behav*. 1985;35(2):167-174. doi:10.1016/0031-9384(85)90331-2
8. Ehret G. Categorical perception of mouse-pup ultrasounds in the temporal domain. *Anim Behav*. 1992;43(3):409-416.
9. Tasaka GI, Feigin L, Maor I, et al. The temporal association cortex plays a key role in auditory-driven maternal plasticity. *Neuron*. 2020;107(3):566-579 e7. doi:10.1016/j.neuron.2020.05.004
10. D'Amato FR, Scalera E, Sarli C, Moles A. Pups call, mothers rush: does maternal responsiveness affect the amount of ultrasonic vocalizations in mouse pups? *Behav Genet*. 2005;35(1):103-112.
11. Panksepp JB, Jochman KA, Kim JU, et al. Affiliative behavior, ultrasonic communication and social reward are influenced by genetic variation in adolescent mice. *PloS One*. 2007;2(4):e351. doi:10.1371/journal.pone.0000351
12. Ju A, Hammerschmidt K, Tantra M, Krueger D, Brose N, Ehrenreich H. Juvenile manifestation of ultrasound communication deficits in the neuroligin-4 null mutant mouse model of autism. *Behav Brain Res*. 2014;270:159-164. doi:10.1016/j.bbr.2014.05.019
13. Maggio JC, Maggio JH, Whitney G. Experience-based vocalization of male mice to female chemosignals. *Physiol Behav*. 1983;31(3):269-272. doi:10.1016/0031-9384(83)90186-5
14. Whitney G, Nyby J. Cues that elicit ultrasounds from adult male mice. *Am Zool*. 1979;19(2):457-463.
15. Heckman JJ, Proville R, Heckman GJ, Azarfar A, Celikel T, Englitz B. High-precision spatial localization of mouse vocalizations during social interaction. *Scientific Reports*. 2017;7(1):3017. doi:10.1038/s41598-017-02954-z
16. Neunuebel JP, Taylor AL, Arthur BJ, Egnor SE. Female mice ultrasonically interact with males during courtship displays. *eLife*. 2015;4:e06203. doi:10.7554/eLife.06203
17. Warren MR, Spurrier MS, Roth ED, Neunuebel JP. Sex differences in vocal communication of freely interacting adult mice depend upon behavioral context. *PloS One*. 2018;13(9):e0204527. doi:10.1371/journal.pone.0204527
18. Burke K, Screven LA, Dent ML. CBA/CaJ mouse ultrasonic vocalizations depend on prior social experience. *PloS One*. 2018;13(6):e0197774. doi:10.1371/journal.pone.0197774
19. Holy TE, Guo Z. Ultrasonic songs of male mice. *PLoS Biol*. 2005;3(12):e386. doi:10.1371/journal.pbio.0030386
20. Kanno K, Kikusui T. Effect of Sociosexual experience and aging on number of courtship ultrasonic vocalizations in male mice. *Zoolog Sci*. 2018;35(3):208-214. doi:10.2108/zs170175
21. Nicolakis D, Marconi MA, Zala SM, Penn DJ. Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success. *Front Zool*. 2020;17:10. doi:10.1186/s12983-020-00353-1
22. Nyby JG. Auditory communication among adults. *Handbook of Mouse Auditory Research*. CRC Press; 2001:16.
23. Zala SM, Nicolakis D, Marconi MA, et al. Primed to vocalize: wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female. *PloS One*. 2020;15(12):e0242959. doi:10.1371/journal.pone.0242959
24. Barnes TD, Rieger MA, Dougherty JD, Holy TE. Group and individual variability in mouse pup isolation calls recorded on the same day show stability. *Frontiers in Behavioral Neuroscience*. 2017;11:243. doi:10.3389/fnbeh.2017.00243
25. Marconi MA, Nicolakis D, Abbasi R, Penna DJ, Zala SM. Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures. *Anim Behav*. 2020;169:169-197.
26. Melotti L, Siestrup S, Peng P, et al. Individuality, as well as genetic background, affects syntactical features of courtship songs in male mice. *Anim Behav*. 2021;180:179-196.
27. Verjat A, Rodel HG, Feron C. Isolation calls in house mouse pups: individual consistency across time and situations. *Dev Psychobiol*. 2019;61(8):1135-1145. doi:10.1002/dev.21884
28. Maggio JC, Whitney G. Ultrasonic vocalizing by adult female mice (Mus musculus). *J Comp Psychol*. 1985;99(4):420-436.
29. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (Mus musculus musculus). *PloS One*. 2017;12(12):e0188647. doi:10.1371/journal.pone.0188647
30. Hoier S, Pfeifle C, von Merten S, Linnenbrink M. Communication at the garden fence--context dependent vocalization in female house mice. *PloS One*. 2016;11(3):e0152255. doi:10.1371/journal.pone.0152255
31. Moles A, Costantini F, Garbugino L, Zanettini C, D'Amato FR. Ultrasonic vocalizations emitted during dyadic interactions in female mice: a possible index of sociability? *Behavioural Brain Research*. 2007;182(2):223-230. doi:10.1016/j.bbr.2007.01.020
32. Branchi I, Santucci D, Vitale A, Alleva E. Ultrasonic vocalizations by infant laboratory mice: a preliminary spectrographic characterization under different conditions. *Dev Psychobiol*. 1998;33(3):249-256. doi:10.1002/(sici)1098-2302(199811)33:3<249::aid-dev5>3.0.co;2-r
33. Brudzynski SM, Kehoe P, Callahan M. Sonographic structure of isolation-induced ultrasonic calls of rat pups. *Dev Psychobiol*. 1999;34(3):195-204.
34. Scattoni ML, Michetti C, Ricceri L. Chapter 42 - rodent vocalization studies in animal models of the autism Spectrum disorder. In: Brudzynski SM, ed. *Handbook of Behavioral Neuroscience*. Elsevier; 2018:445-456.
35. Binder MS, Shi HD, Bordey A. CD-1 outbred mice produce less variable ultrasonic vocalizations than FVB inbred mice, while displaying a similar developmental trajectory. *Front Psych*. 2021;12:687060. doi:10.3389/fpsyt.2021.687060
36. Peleh T, Eltokhi A, Pitzer C. Longitudinal analysis of ultrasonic vocalizations in mice from infancy to adolescence: insights into the vocal repertoire of three wild-type strains in two different social contexts. *PloS One*. 2019;14(7):e0220238. doi:10.1371/journal.pone.0220238
37. Kikusui T, Nakanishi K, Nakagawa R, Nagasawa M, Mogi K, Okanoya K. Cross fostering experiments suggest that mice songs are innate. *PloS One*. 2011;6(3):e17721. doi:10.1371/journal.pone.0017721
38. Wohr M, Dahlhoff M, Wolf E, Holsboer F, Schwarting RK, Wotjak CT. Effects of genetic background, gender, and early environmental factors on isolation-induced ultrasonic calling in mouse pups: an embryo-transfer study. *Behav Genet*. 2008;38(6):579-595. doi:10.1007/s10519-008-9221-4
39. Arriaga G, Jarvis ED. Mouse vocal communication system: are ultrasounds learned or innate? *Brain Lang*. 2013;124(1):96-116. doi:10.1016/j.bandl.2012.10.002
40. Ashbrook DG, Roy S, Clifford BG, et al. Born to cry: a genetic dissection of infant vocalization. *Frontiers in Behavioral Neuroscience*. 2018;12:250. doi:10.3389/fnbeh.2018.00250
41. Bryant CD, Zhang NN, Sokoloff G, et al. Behavioral differences among C57BL/6 substrains: implications for transgenic and knockout studies. *J Neurogenet*. 2008;22(4):315-331. doi:10.1080/01677060802357388
42. Matsuo N, Takao K, Nakanishi K, Yamasaki N, Tanda K, Miyakawa T. Behavioral profiles of three C57BL/6 substrains. *Frontiers in Behavioral Neuroscience*. 2010;4:29. doi:10.3389/fnbeh.2010.00029
43. Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T + tf/J mouse model of autism. *PloS One*. 2008;3(8):e3067. doi:10.1371/journal.pone.0003067
44. Wiaderkiewicz J, Glowacka M, Grabowska M, Jaroslaw-Jerzy B. Ultrasonic vocalizations (USV) in the three standard laboratory mouse strains: developmental analysis. *Acta Neurobiol Exp*. 2013;73(4):557-563.

45. Rugh R. *The Mouse: its Reproduction and Development*. Oxford University Press; 1990.
46. Scattoni ML, Crawley J, Ricceri L. Ultrasonic vocalizations: a tool for behavioural phenotyping of mouse models of neurodevelopmental disorders. *Neurosci Biobehav Rev*. 2009;33(4):508-515. doi:10.1016/j.neubiorev.2008.08.003
47. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T + tf/J mice during three types of social encounters. *Genes Brain Behav*. 2011;10(1):44-56. doi:10.1111/j.1601-183X.2010.00623.x
48. Vigli D, Palombelli G, Fanelli S, et al. Maternal immune activation in mice only partially recapitulates the autism Spectrum disorders symptomatology. *Neuroscience*. 2020;445:109-119. doi:10.1016/j.neuroscience.2020.05.009
49. Gourbal BE, Barthelemy M, Petit G, Gabrion C. Spectrographic analysis of the ultrasonic vocalisations of adult male and female BALB/c mice. *Naturwissenschaften*. 2004;91(8):381-385. doi:10.1007/s00114-004-0543-7
50. Heckman J, McGuinness B, Celikel T, Englitz B. Determinants of the mouse ultrasonic vocal structure and repertoire. *Neurosci Biobehav Rev*. 2016;65:313-325. doi:10.1016/j.neubiorev.2016.03.029
51. Sasaki E, Tomita Y, Kanno K. Sex differences in vocalizations to familiar or unfamiliar females in mice. *R Soc Open Sci*. 2020;7(12):201529. doi:10.1098/rsos.201529
52. Grimsley JM, Monaghan JJ, Wenstrup JJ. Development of social vocalizations in mice. *PloS One*. 2011;6(3):e17460. doi:10.1371/journal.pone.0017460
53. Zeskind PS, McMurray MS, Garber KA, et al. Development of translational methods in spectral analysis of human infant crying and rat pup ultrasonic vocalizations for early neurobehavioral assessment. *Front Psych*. 2011;2:56. doi:10.3389/fpsyt.2011.00056
54. Burenkova OV, Averkina AA, Aleksandrova EA, Zarayskaya IY. Brief but enough: 45-min maternal separation elicits behavioral and physiological responses in neonatal mice and changes in dam maternal behavior. *Physiology & Behavior*. 2020;222:112877. doi:10.1016/j.physbeh.2020.112877
55. D'Amato FR, Cabib S. Chronic exposure to a novel odor increases pups' vocalizations, maternal care, and alters dopaminergic functioning in developing mice. *Behav Neural Biol*. 1987;48(2):197-205. doi:10.1016/s0163-1047(87)90738-2
56. D'Amato FR, Cabib S, Puglisi-Allegra S, et al. Effects of acute and repeated exposure to stress on the hypothalamo-pituitary-adrenocortical activity in mice during postnatal development. *Horm Behav*. 1992;26(4):474-485. doi:10.1016/0018-506x(92)90015-n
57. Bell RW, Nitschke W, Zachman TA. Ultra-sounds in three inbred strains of young mice. *Behav Biol*. 1972;7(6):805-814. doi:10.1016/s0091-6773(72)80172-x
58. Hahn ME, Hewitt JK, Schanz N, Weinreb L, Henry A. Genetic and developmental influences on infant mouse ultrasonic calling. I. a diallel analysis of the calls of 3-day olds. *Behav Genet*. 1997;27(2):133-143. doi:10.1023/a:1025637408900
59. Hennessy MB, Li J, Lowe EL, Levine S. Maternal behavior, pup vocalizations, and pup temperature changes following handling in mice of 2 inbred strains. *Dev Psychobiol*. 1980;13(6):573-584. doi:10.1002/dev.420130603
60. Thornton LM, Hahn ME, Schanz N. Genetic and developmental influences on infant mouse ultrasonic calling. III. Patterns of inheritance in the calls of mice 3-9 days of age. *Behav Genet*. 2005;35(1):73-83. doi:10.1007/s10519-004-0857-4
61. Robinson DJ, D'Udine B. Ultrasonic calls produced by three laboratory strains of Mus musculus. *J Biol*. 1982;197:383-389.
62. Bouwknecht JA, Paylor R. Behavioral and physiological mouse assays for anxiety: a survey in nine mouse strains. *Behavioural Brain Research*. 2002;136(2):489-501. doi:10.1016/s0166-4328(02)00200-0
63. Caruso A, Sabbioni M, Scattoni ML, Branchi I. Quantitative and qualitative features of neonatal vocalizations in mice. In: Brudzynski SM, ed. *Handbook of Ultrasonic Vocalization* Elsevier; 2018:139-145: chap 13.
64. Faure A, Pittaras E, Nosjean A, Chabout J, Cressant A, Granon S. Social behaviors and acoustic vocalizations in different strains of mice. *Behav Brain Res*. 2017;320:383-390. doi:10.1016/j.bbr.2016.11.003
65. Crawley JN. *What's Wrong with My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice*. Wiley-Liss Publication; 2000.
66. Moy SS, Nadler JJ. Advances in behavioral genetics: mouse models of autism. *Mol Psychiatry*. 2008;13(1):4-26. doi:10.1038/sj.mp.4002082
67. Canastar A, Maxson SC. Sexual aggression in mice: effects of male strain and of female estrous state. *Behav Genet*. 2003;33(5):521-528. doi:10.1023/a:1025722700138
68. Tuttle AH, Philip VM, Chesler EJ, Mogil JS. Comparing phenotypic variation between inbred and outbred mice. *Nat Methods*. 2018;15(12):994-996. doi:10.1038/s41592-018-0224-7
69. Matsumoto YK, Okanoya K. Mice modulate ultrasonic calling bouts according to sociosexual context. *R Soc Open Sci*. 2018;5(6):180378. doi:10.1098/rsos.180378
70. Liu RC, Miller KD, Merzenich MM, Schreiner CE. Acoustic variability and distinguishability among mouse ultrasound vocalizations. *J Acoust Soc Am*. 2003;114(6 Pt 1):3412-3422. doi:10.1121/1.1623787
71. Nomoto K, Hashiguchi A, Asaba A, et al. Female C57BL/6 and BALB/c mice differently use the acoustic features of male ultrasonic vocalizations for social preferences. *Experimental Animals*. 2020;69(3):319-325. doi:10.1538/expanim.19-0119
72. Sugimoto H, Okabe S, Kato M, et al. A role for strain differences in waveforms of ultrasonic vocalizations during male-female interaction. *PloS One*. 2011;6(7):e22093. doi:10.1371/journal.pone.0022093
73. Lahvis GP, Alleva E, Scattoni ML. Translating mouse vocalizations: prosody and frequency modulation. *Genes Brain Behav*. 2011;10(1):4-16. doi:10.1111/j.1601-183X.2010.00603.x
74. Wohr M, Schwarting RK. Affective communication in rodents: ultrasonic vocalizations as a tool for research on emotion and motivation. *Cell Tissue Res*. 2013;354(1):81-97. doi:10.1007/s00441-013-1607-9
75. Burgdorf JS, Brudzynski SM, Moskal JR. Using rat ultrasonic vocalization to study the neurobiology of emotion: from basic science to the development of novel therapeutics for affective disorders. *Curr Opin Neurobiol*. 2020;60:192-200. doi:10.1016/j.conb.2019.12.008
76. Simola N, Granon S. Ultrasonic vocalizations as a tool in studying emotional states in rodent models of social behavior and brain disease. *Neuropharmacology*. 2019;159:107420. doi:10.1016/j.neuropharm.2018.11.008
77. Gaub S, Fisher SE, Ehret G. Ultrasonic vocalizations of adult male Foxp2-mutant mice: behavioral contexts of arousal and emotion. *Genes Brain Behav*. 2016;15(2):243-259. doi:10.1111/gbb.12274
78. Granon S, Faure A, Chauveau F, Cressant A, Ey E. Why should my mouse call me? Acoustic communication in mouse models of social disorders: ultrasonic vocalizations as an index of emotional and motivational states. In: Brudzynski SM, ed. *Handbook of Behavioral Neuroscience*. Elsevier; 2018:423-431:chap 40.
79. Eltokhi A, Kurpiers B, Pitzer C. Behavioral tests assessing neuropsychiatric phenotypes in adolescent mice reveal strain- and sex-specific effects. *Scientific Reports*. 2020;10(1):11263. doi:10.1038/s41598-020-67758-0
80. Kopp C, Vogel E, Misslin R. Comparative study of emotional behaviour in three inbred strains of mice. *Behav Processes*. 1999;47(3):161-174. doi:10.1016/s0376-6357(99)00057-1
81. Hanson JL, Hurley LM. Female presence and estrous state influence mouse ultrasonic courtship vocalizations. *PloS One*. 2012;7(7):e40782. doi:10.1371/journal.pone.0040782
82. Ivanenko A, Watkins P, van Gerven MAJ, Hammerschmidt K, Englitz B. Classifying sex and strain from mouse ultrasonic vocalizations using deep learning. *PLoS Comput Biol*. 2020;16(6):e1007918. doi:10.1371/journal.pcbi.1007918

83. Sangiamo DT, Warren MR, Neunuebel JP. Ultrasonic signals associated with different types of social behavior of mice. *Nat Neurosci*. 2020;23(3):411-422. doi:10.1038/s41593-020-0584-z
84. Chabout J, Serreau P, Ey E, et al. Adult male mice emit context-specific ultrasonic vocalizations that are modulated by prior isolation or group rearing environment. *PloS One*. 2012;7(1):e29401. doi:10.1371/journal.pone.0029401
85. Hoffmann F, Musolf K, Penn DJ. Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. *Physiology & behavior*. 2012;105(3):766-771. doi:10.1016/j.physbeh.2011.10.011

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Caruso A, Marconi MA, Scattoni ML, Ricceri L. Ultrasonic vocalizations in laboratory mice: strain, age, and sex differences. *Genes, Brain and Behavior*. 2022; 21(5):e12815. doi:10.1111/gbb.12815

CHAPTER 6: Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap)

Abbasi, Reyhaneh; Balazs, Peter; Marconi, Maria Adelaide; Nicolakis, Doris; Zala, Sarah M.; Penn, Dustin J.

PLoS Computational Biology, impact factor: 4.8; **published**

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap)

Reyhaneh Abbasi[1,2,3]*, **Peter Balazs**[1], **Maria Adelaide Marconi**[2], **Doris Nicolakis**[2], **Sarah M. Zala**[2☯], **Dustin J. Penn**[2☯]

1 Acoustic Research Institute, Austrian Academy of Sciences, Vienna, Austria, **2** Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine, Vienna, Austria, **3** Vienna Doctoral School of Cognition, Behaviour and Neuroscience, University of Vienna, Vienna, Austria

☯ These authors contributed equally to this work.
* reyhaneh.abbasi@oeaw.ac.at

Citation: Abbasi R, Balazs P, Marconi MA, Nicolakis D, Zala SM, Penn DJ (2022) Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap). PLoS Comput Biol 18(5): e1010049. https://doi.org/10.1371/journal.pcbi.1010049

Editor: Dan Stowell, Tilburg University Faculty Humanities: Tilburg University Tilburg School of Humanities and Digital Sciences, NETHERLANDS

Received: June 13, 2021

Accepted: March 22, 2022

Published: May 12, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010049

Data Availability Statement: The sound files of wild mice that we used to evaluate our model is available online ((https://mousetube.pasteur.fr/) or

Abstract

House mice communicate through ultrasonic vocalizations (USVs), which are above the range of human hearing (>20 kHz), and several automated methods have been developed for USV detection and classification. Here we evaluate their advantages and disadvantages in a full, systematic comparison, while also presenting a new approach. This study aims to 1) determine the most efficient USV detection tool among the existing methods, and 2) develop a classification model that is more generalizable than existing methods. In both cases, we aim to minimize the user intervention required for processing new data. We compared the performance of four detection methods in an out-of-the-box approach, pretrained DeepSqueak detector, MUPET, USVSEG, and the Automatic Mouse Ultrasound Detector (A-MUD). We also compared these methods to human visual or 'manual' classification (ground truth) after assessing its reliability. A-MUD and USVSEG outperformed the other methods in terms of true positive rates using default and adjusted settings, respectively, and A-MUD outperformed USVSEG when false detection rates were also considered. For automating the classification of USVs, we developed *BootSnap* for supervised classification, which combines bootstrapping on Gammatone Spectrograms and Convolutional Neural Networks algorithms with Snapshot ensemble learning. It successfully classified calls into 12 types, including a new class of false positives that is useful for detection refinement. *BootSnap* outperformed the pretrained and retrained state-of-the-art tool, and thus it is more generalizable. *BootSnap* is freely available for scientific use.

Author summary

House mice and many other species use ultrasonic vocalizations to communicate in various contexts including social and sexual interactions. These vocalizations are increasingly investigated in research on animal communication and as a phenotype for studying the genetic basis of autism and speech disorders. Because manual methods for analyzing

(https://zenodo.org/record/5771669#.YiohQ9XML3g)). The sound files of laboratory mice, uploaded by Chabout et al. (2015), are already available online ((https://duke.box.com/shared/static/6j08fzyto8nuxxstk6bcpi9n52bk5bu4.wav) and (https://duke.box.com/shared/static/y5o7zw8jx9ugb2qocozyaup7xlby1sr8.wav)). Codes needed to reproduce our results are available online (https://github.com/ReyhanehAbbasi/BootSnap).

Funding: This work was supported by the FWF P 34922-N project ("NoMASP: Nonsmooth Nonconvex Optimization Methods for Acoustic Signal Processing") to PB and by a grant (FWF P 28141-B25) of the Austrian Science Foundation (http://www.fwf.ac.at) to DJP and SMZ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

vocalizations are extremely time consuming, automatic tools for detection and classification are needed. We evaluated the performance of the available tools for analyzing ultrasonic vocalizations, and we compared detection tools for the first time to manual methods ("ground truth") using recordings from wild-derived and laboratory mice. For the first time, class-wise inter-observer reliability of manual labels used for ground truth are analyzed and reported. Moreover, we developed a new classification method based on ensemble deep learning that provides more generalizability than the current state-of-the-art tool (both pretrained and retrained). Our new classification method is free for scientific use.

Introduction

The ultrasonic vocalizations (USVs) of house mice (*Mus musculus)* and rats (*Rattus norvegicus)* are surprisingly complex, and they are increasingly being investigated to better understand animal communication [1–3] and as a model for studying the genetic basis of autism and speech disorders [4,5]. Rodents emit USVs in discrete units called *syllables* or *calls* (these terms are metaphors and do not imply that rodents use words, or that their vocalizations function to attract other mice). USV syllables are separated by gaps of silence and they have been classified into several different categories by researchers visually inspecting spectrograms [1–3,6–10] i.e., the squared modulus of the short-time Fourier transforms (STFT) [11], or, less often, by statistical clustering analyses [12–17]. USVs are classified according to their shape and other spectro-temporal features, including the length of each syllable, their frequency, and degree of complexity. Classification provides the basis for subsequent analyses of USVs, such as repertoire size (e.g., Nicolakis et al. [7], Marconi et al. [6]) and sequences or *syntax* (e.g., Heckman et al. [3],Chabout et al. [18]).

Several classifications of USVs have been proposed, which vary from three to 12 different classes, and there is no consensus on how they should be classified. Researchers agree that there is a qualitative distinction between simple versus complex types of USVs (with the latter having frequency-jumps or harmonics), but not with other proposed classes, as their differences are fuzzy. Many proposed classes are quantitative variations within these two major categories (e.g., simple USVs show quantitative differences in length and shape, and complex syllables can have one or more frequency-jumps). A recent study concluded that USVs do not cluster into distinctive types, and instead form a continuum [19]. However, since USVs are mainly classified by human researchers, the crucial question is how do rodents perceive and respond to variations in USVs. Continuous differences in these calls might still be perceived as categorically discrete by rodents, just as we perceive continuous speech as discrete words and variations in wavelengths of light as different colors. Few studies have addressed questions about perception so far, and the evidence suggests that mice differentiate some though not other USV classes (see Outlook below). Moreover, house mice emit different types of USVs depending upon the social contexts and potential receivers [10,18,20–24], and they alter their syllable type usage over the time during courtship and mating [25–27]. Thus, identifying variations in USVs in different contexts is central to studying the functions of these vocalizations (Nicolakis et al. [7], Marconi et al. [6]).

The main technical challenge for USV processing and analyses includes developing better methods for detecting and classifying these vocalizations, since most analyses are still conducted manually by visual inspection of spectrograms, which is extremely time-consuming.

The first step in this signal processing task is USV detection, which is a challenging problem due to the low signal-to-noise ratio (SNR) in most recording conditions. Manually detecting

each USV can take an enormous amount of time, particularly with large datasets. Semi-automatic methods are useful, but they are still time-consuming (e.g., semi-automatic detection using Avisoft SASLab Pro and manual checks requires 1–1.5 hours to detect merely 150–300 USVs [28], and some datasets contain tens of thousands of USVs [6]). The time required to classify USVs takes even longer than detection, and classification is a necessary step to evaluate qualitative differences in vocalizations and to conduct analyses of USV sequences (syntax) (e.g., von Merten et al. [8]). Several software tools have recently become available for automating USV detection, including MUPET [13], MSA [14], DeepSqueak (DSQ) [16], USVSEG [29], Automatic Mouse Ultrasound Detector (A-MUD) [30], Ultravox (Noldus; Wageningen, NL) (commercial), and SONOTRACK (commercial). These tools enhance the efficiency of processing USV data, but they can generate errors for several reasons. Failing to detect actual USVs (the probability is given by the false-negative rate or FNR) can result in missing actual differences in the vocalizations, and erroneous detections (false detection rate or FDR) can lead to failure to detect actual differences and generate false differences. The challenge for any USV detection algorithm is maximizing the true positive rate (TPR) while minimizing the FNR and FDR. Moreover, automatic methods can have systematic biases depending on how they are developed. For example, automated methods developed using only one mouse strain, one sex, or only in one state or context can increase both types of error when applied to other mice or conditions (see S1 Table for the mice and recording conditions used for developing different USV detection tools if applied in other settings). Thus, automated methods can greatly enhance the efficiency of processing USV data, but it is critical that they can be generalized. Results should be treated with caution until the error rates in the detection and classification method are evaluated for particular datasets, or their generalizability is demonstrated.

To our knowledge, five studies have compared the performance of USV detection algorithms: (1) Binder et al. [28] compared MSA and Avisoft for detecting USVs emitted from different strains of mice (C57BL/6, Fmr1-FVB.129, NS-Pten-FVB, and 129). They concluded that Avisoft outperformed MSA for C57BL/6 and NS-Pten-FVB strains, but these two methods performed similarly for strain 129. Thus, there are strain-specific differences between these two detection tools. (2) Another study [31] compared the quantity of USVs detected by Avisoft to those detected by Ultravox (2.0) and reported significant differences in USV detection and weaker than expected overall correlations between the systems under congruent detection parameters. (3) Van Segbroeck et al. [13] compared MUPET and MSA for detecting USVs emitted by B6D2F1 males from MouseTube [3] and found that these methods generated similar call counts and spectro-temporal measures of individual syllables. (4) Coffey et al. [16] compared MUPET, Ultravox, and DSQ for detecting USVs by analyzing the TPR and precision (the ratio of detected true USVs to false positives). For this purpose, they manipulated a recording from MouseTube in two ways to gradually degrade its quality. In the first experiment, increasing levels of Gaussian white noise were added to recordings, and DSQ outperformed MUPET and Ultravox in terms of TPR and precision in all Gaussian noise levels. In the second experiment, real noise was added to recordings, and DSQ again outperformed MUPET in terms of precision and Ultravox in terms of precision and TPR. (5) Zala et al. [30] compared the performance of Avisoft and A-MUD (version 1.0) in identifying USVs of wild-derived *Mus musculus musculus*. They concluded that the latter method is superior in terms of TPR and FDR. Zala et al. [32] have since provided an updated version of A-MUD, which overcomes previous difficulties in identifying faint and short USVs.

Our first aim was to systematically compare the performance of four commonly used USV detection tools, MUPET, DSQ, A-MUD, and USVSEG, and to determine which is the most efficient and requires the least user intervention. We addressed three main questions:

1. How does the performance of different USV detection methods compare to each other? Previous studies indicate that A-MUD outperforms Avisoft, which outperforms MSA; MSA is comparable to MUPET and DSQ outperforms MUPET and Ultravox. To our knowledge, no study has systematically compared the performance of A-MUD and DSQ, or evaluated more than two of these methods together, though Coffey et al. [16] recently, compared DSQ, MUPET, and Ultravox.
2. How does the performance of USV detection methods compare to ground truth (i.e., manual detection by trained researchers)? Evaluation of detection methods rarely includes such a positive control, which is a crucial comparison to obtain absolute versus relative estimates of performance (e.g., see [30]). Binder et al. [28], Binder et al. [31], and Van Segbroeck et al. [13] compared Avisoft and MSA, Ultravox and Avisoft, and MUPET and MSA based on the number of USVs detected by each of the two methods, but no comparisons were made with ground truth. Coffey et al. [16] used only ca. 100 manually detected USVs as ground truth for comparing DSQ, MUPET, and Ultravox.
3. How well do USV detection tools perform when using novel datasets that differ from the original training set (often called, generalization performance, out-of-sample error, or out-of-the-box performance)? To our knowledge, only one study [28] has tested whether USV detection methods generalize to other mouse strains (comparing only Avisoft and MSA), and only one study has compared MSA and MUPET for different recording conditions (males vocalizing in response to female urine, an anesthetized female, and awake female) [13]. Van Segbroeck et al. [13] and Coffey et al. [16] only used recordings from a hybrid strain (B6D2F1), and Zala et al. [30] used wild-derived *Mus musculus*. Consequently, it is unclear how well current detection methods perform whenever applied to new recordings that differ from the data used to develop and evaluate the tool. The problem of generalization is well known in the machine learning community and there are several approaches to improve "transfer learning" [33].

Therefore, we compared the "out-of-the-box" performance of these USV detection tools with each other, and with ground truth, and we assessed their performance with novel datasets. For these comparisons, we used recordings of laboratory mice (*Mus laboratorius*) and wild-derived house mice (*Mus musculus musculus*), and using recordings under different social contexts and recording conditions. The data were obtained from sources not involved in the developmental phase for our tools (see Data and Methods). To evaluate the absolute performance of these models, we applied a new dataset of manually detected USVs as ground truth with a total of 3955 USVs. We minimized adjusting the detection parameters or re-training these tools because such additional user interventions would add more variables and make it impossible to compare their efficiency. One could include re-training before using or testing a detection tool with a novel dataset, but then the data would have to be re-labeled, which defeats the purpose of using an automated tool. To evaluate performance, we compared TPR (i.e., how often USVs are correctly detected) and FDR (how often background noises are mistakenly detected as USVs). Signal detection theory explains the inevitable trade-off between FPs and FNs [34], and therefore, the most effective tool will provide an optimal balance of these types of errors.

We also aimed to develop an improved method for detecting FPs, as a second refinement or data cleaning step to remove noise before classifying USVs or making other analyses. Whenever analyzing recordings of mice, there are always background noises i.e., non-USV sounds generated from recording instruments or movements of the mouse and bedding especially during social interactions. FNs are problematic as they result in a loss of data for

subsequent analyses; however, false positives from detection are more problematic for statistical analyses and training a classification tool. One can set the parameters of detection such that it errs on the negative rather than the positive set, as FPs can be deleted in the refinement step. To remove FPs, MUPET and DSQ include a preliminary detection refinement step using either an unsupervised approach, which groups data based on similarity measures rather than manually labeled USVs (both approaches), or a supervised approach, which requires manually labeled USVs for training a classifier (DSQ and [35]). Our preliminary evaluation found that DSQ outperformed MUPET in the detection refinement step (using the K-means clustering [36]), however, its performance differs depending on the data. Thus, we designed a method better suited to deal with the problems mentioned above and we compared our method with DSQ for detecting FPs, as this is a critical step for accurate USV classification.

Our second aims were to evaluate the state-of-the-art method for automated USVs classification, and to develop a better method, i.e., an out-of-the-box, high-performance, and supervised method that requires minimal human intervention. Automatic classification of USV syllable types can be achieved through unsupervised [12–14,16,17] and supervised [16] classifiers. The advantage of unsupervised classification (often called 'clustering') is that it is considered to be more objective, as it does not require a predefined number of classes or manually labeled observations. Hence, the number of classes is based on the information contained in the dataset rather than the researchers' assessment. These clusters do not always match the classification of USVs by researchers and it is unclear how they are perceived by mice (see Outlook below). In contrast, supervised classification (classification *sensu stricto*) methods require that researchers first classify or assigning labels to USVs for training a classifier (machine learning), which has higher accuracy compared to clustering [37,38]. One needs to use supervised classification for comparing the results between datasets and manual labels. To our knowledge, only a few studies have used supervised methods for classifying mouse USVs (see S1 Text).

Since the generalizability of USV classifiers has never been investigated (unlike methods for classifying bird vocalizations [39]), it is not known how well the current methods can classify USVs for novel datasets. Again, to evaluate a classification method, a systematic evaluation of a new dataset not used for training or testing is needed. We identified four key factors that can reduce the performance and generalizability of USV classifiers:

1. Noise is a potential problem for classification, as for detection, but this issue has not received sufficient consideration. Some methods used only recordings that had low background noise (high SNR data) for developing and testing their models (e.g., [40], [16], and [41]). This approach seems logical but it results in reduced performance when using more typical recordings of mice having a low-SNR [42]. This problem is exacerbated if the model is developed using predefined features extracted from spectrograms (e.g., see [40]), as the extraction of these features from low-SNR signals already introduces high variance.

2. Imprecise USV detection generates subsequent classification errors. As the main output after detection is usually the time and frequency range of USVs, the classification will only include the region of the spectrogram limited to the detected minimum and maximum USV frequency [16,40]. Our investigations, however, revealed that faint portions of USVs are often not included inside this window, leading to significant errors in feature estimation and classification.

3. Neural networks are being increasingly used for USV classification [16,41,43]. Machine learning is an iterative method and it can fail to find the most effective weights for classification, however, because the algorithm takes a path that reduces the error and this can lead to

focusing on specific weights, which may not be very useful. Becoming trapped in a local minimum is a common problem, and it can reduce the generalizability of a classifier [44]. This problem can be overcome by using ensemble machine learning methods [45], a procedure that uses multiple algorithms, and the final output is obtained from combining the outputs of these models. This approach makes it possible to obtain a model with better performance than any of its component models and it allows for more flexible structures, though developing ensembles require additional training time.

4. Limited training and evaluation inflate model performance. The performance of any model is overly-optimistic whenever the same type of data (e.g., same mouse strain or recording context) is used for both model development and evaluation [40,41,43]. Using such a limited training set conceals the model's shortcomings in dealing with different strains or recording conditions, but surprisingly, previous studies have never considered this issue.

Thus, to develop new and improved methods for USV classification, we had the following aims:

1. Apply a CNN Snapshot Ensemble classifier based on the stochastic gradient descent algorithm, which is accurate even with noisy (low-SNR) data.
2. Use the full time-frequency images based on the entire frequency range and reduce the dimensionality (and thereby the computational load and the possibility of overfitting) using Gammatone filters applied to the spectrograms.
3. Compare our new method with pretrained (as an out-of-the-box model) and retrained DeepSqueak (DSQ), which is currently the state-of-the-art classification tool, and evaluate these methods using USVs recorded under different conditions and from different mice strains than the conditions and strains used in the training step.

Data and methods

USV data

Subjects. The data used in this study was first divided into two meta-sets: we have used one development set (DEV) to develop, train and test the developed detection and classification methods. To test the generalizability of the methods we use an additional evaluation (EV) set. For a direct test, as well as estimating the meta-parameters of the classifier, using stratified 8-fold cross-validation, the DEV dataset was further divided into three subsets including DEV_train, DEV_validation, and DEV_test (Table 1). We report the performance of the proposed classifier in Sections "Selecting the architecture of the classifier", "Evaluating BootSnap for classifying USVs", and "Inference classification" over the DEV_validation and DEV_test datasets. The DEV dataset (Zala et al. [30]) combined two pre-existing datasets: the first dataset was from 11 wild-derived male and 3 female mice (*Mus musculus musculus*) recorded for 10 min in the presence of an unfamiliar female stimulus [24]. In the second data set, 30 wild-derived male mice (*M. musculus musculus*) were recorded for 10 min in the presence of an unfamiliar female on 2 consecutive days, first unprimed and then sexually primed. These were F1 and F2 descendants from wild-caught *M. musculus musculus*, respectively (which for brevity, we refer to as 'wild mice'), whereas laboratory mice are domesticated hybrids of three *Mus* subspecies, and mainly *Mus musculus domesticus*.

The EV dataset consists of two datasets, and a part was obtained from wild mice ('EV_wild') (as in DEV), but under different conditions [6]. The vocalizations were obtained from 22 sexually experienced adult wild-derived (F3) male *M. musculus musculus* [6]. Male vocalizations

Table 1. Number of instances for each class in the different datasets.

Data set	Number of members in each class													
DEV_train	c	**c2**	**c3**	**c4**	**c5**	**h**	**d**	**up**	**u**	**f**	**us**	**s**	**ui**	**FP**
	308	241	69	0	0	124	299	4343	298	1277	74	291	543	4849
DEV_validation	53	42	12	0	0	21	52	753	52	221	13	51	94	840
DEV_test	50	39	11	0	0	20	48	695	48	205	12	47	87	776
EV_wild	c	**c2**	**split**				**Rise**						**ui**	**FP**
	20	224	334				1025						110	234
EV_lab	61	404	739				819						200	389

https://doi.org/10.1371/journal.pcbi.1010049.t001

were recorded without and also during the presentation of a female urine stimulus over three recording weeks, one time per week and each time for 15 minutes. To evaluate classifier performance, we used three arbitrarily chosen recordings out of these 66 recordings, and manually classified them for this study. The other part of the EV data is taken from the MouseTube dataset used for developing DSQ ('EV_lab') (B6D2F1 mice recorded by Chabout et al. [18]) and two arbitrarily selected recordings were sampled out of these 168 recordings. Although we only used a few recordings to evaluate the methods, these recordings contained a large number of USVs (Table 1). In order to prevent any potential bias in the performance of our method, we selected 4 datasets that differed in their recording methods and other characteristics, such as recording contexts (males with fresh or frozen female urine, males or females with a stimulus female separated by a divider), subjects' previous experience (males without or with sociosexual experience), microphone used (condenser ultrasound microphone Avisoft-Bioacoustics CM16/CMPA and USG Electret Ultrasound Microphone Avisoft Bioacoustics / Knowles FG) and genetic background of mice (wild-derived mice of F1-F3 generation and B6D2F1/J laboratory mice). See S2 Text for more detailed information on all datasets.

Detection. For USV detection, we applied A-MUD (version 3.2) using its published default parameters for both the DEV and the EV datasets. Because FPs and syllables are detected during the detection process, we call the detected segments 'elements' rather than 'syllables'. The parameters that affect A-MUD performance are o1_on, o1_off and if oo is enabled, oo_on and oo_off, which are amplitude thresholds in decibel. For this study, we use two A-MUD outputs: the element time slot and the estimated track of the instantaneous frequency over time (frequency track; FT), called 'segment info' (Fig 1). We also compared A-MUD to the three other detection tools, MUPET, DSQ, and USVSEG. To ensure a comparison, where

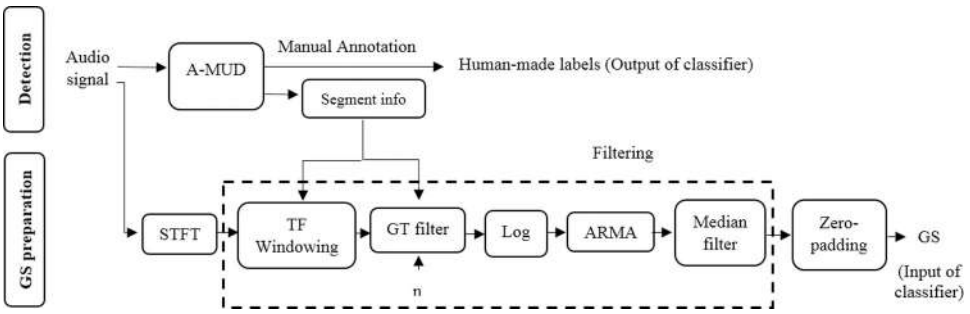

Fig 1. Block diagram showing the procedure for USV detection and input preparation for the classifier. ***n*** is the Gammatone (GT) filter order. STFT, A-MUD, ARMA, and GS are the abbreviation for short-time Fourier transform, automatic mouse ultrasound detector, autoregressive moving average, and Gammatone spectrograms, respectively. TF in 'TF windowing' is the abbreviation for time-frequency. In this step, we restrict the spectrogram to the time of interest, where the segment is detected, and to the frequency of interest, i.e., 20 kHz to 120 kHz.

https://doi.org/10.1371/journal.pcbi.1010049.g001

A-MUD is certainly not privileged, the parameters of A-MUD were fixed while those of the other approaches were optimized, through trial-and-error, i.e., we used the best parameters, which provide the highest true positive rates for each detection tool, and not the default settings. The parameters used for evaluating the different tools are presented in S2 Table.

Since the detection tools that we compared in this study were developed and evaluated using USVs of wild mice (A-MUD) and laboratory mice (DSQ, USVSEG, and MUPET), we also use USVs from both types of mice for our evaluation (two recordings for wild mice from the DEV and EV_wild + two recordings for the laboratory mice from EV_lab). The DEV_1 (1 sound file from DEV data), EV_wild_1 (sound file 1 from EV_wild data), EV_lab_1 (sound file 1 from EV_lab data), and EV_lab_2 (sound file 2 from EV_lab data) signals consist of 947, 771, 1013, and 1224 USVs, respectively.

Manual annotation of detections. After automatically detecting all elements, the DEV dataset was manually classified into 12 classes (Fig 2), depending on the USVs' spectro-temporal features [5–7,9,32,46] (S3 Table). These classes are based on frequency changes [32] (> 5

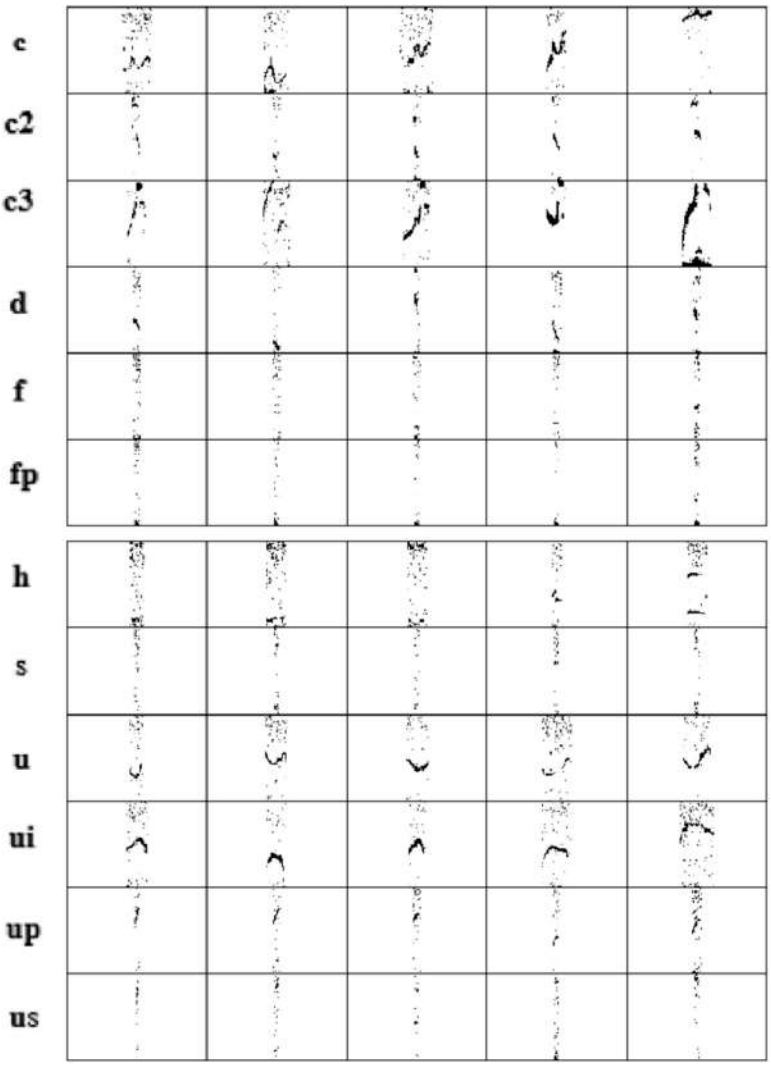

Fig 2. Gammatone Spectrograms (GSs) of five members of 12 studied classes. These GSs have the minimum Manhattan distance to other members of 12 USV classes in the development dataset.

https://doi.org/10.1371/journal.pcbi.1010049.g002

kHz increase 'up', > 5 kHz decrease 'd'), on the number of components (corresponding to breaks in the frequency track; 'c2' with 2 and 'c3' with 3 components), on changes of frequency direction (≥ 2 changes 'c') or shape (u-shape, 'u', u-inverted shape, 'ui'), on frequency modulation (< 5kHz, 'f'), on time (5–10 ms, 's', < 5ms, 'us'), and harmonic elements, 'h'. It is worth noting that there are 2 more USV classes, USVs with 4 'c4' and 5 'c5' components. Due to their infrequency, however, they are excluded from the training task (DEV dataset), but they are used for the evaluation step (EV dataset).

When using low-SNR recordings, or recordings with faint or short USVs, certain background noises are sometimes mistakenly detected as USVs. These errors are false positives (FPs), whereas USVs that are missed are false negatives (FNs). As mentioned above, minimizing one of these types of errors increases the other one, due to inevitable tradeoffs in signal detection [47]. FPs are preferable over FNs, as they can be excluded in a follow-up step, and thus we included FP as a target class. The DEV dataset contained 16958 elements including 6465 FPs in total (Table 1).

When comparing our model with DSQ, the EV data (EV_lab and EV_wild) were manually labeled into 6 classes: 'c2', 'split' (pool of 'c3', 'c4', 'c5', and 'h'), 'c', 'ui', 'FP', and 'Rise' (pool of 'up', 'd', 'f', 's', 'us', and 'u'). We created the classes 'split' and 'Rise' because DSQ reported them together with 'c2', 'c', 'ui', and 'FP' as the output classes. The EV dataset consisted of 4500 elements including FP, of which 1947 and 2615 instances belonged to wild mice and laboratory mice, respectively.

Input images for the classifier. Handcrafted, predetermined features (such as slope, modulation frequency, number of jumps, etc.) are affected by noise, so the development of a classifier based on these features increases the error of the classification, as discussed in the Introduction. Therefore, we developed an image-based supervised classification built on the STFT of detected elements, followed by a set of filters and a zero-padding method (Fig 1).

After applying the time segmentation obtained from A-MUD, a STFT (NFFT = 750) with a 0.8-overlapped Hamming window is applied to the signals, as shown in Fig 1. The desired information in the frequency interval of 20 kHz to 120 kHz and in the time interval of detected elements is extracted ("TF windowing", Fig 1).

A spectrogram (the squared modulus of the STFT) is often used for the analysis of USVs and machine learning approaches [16,41]. But the problem is that spectrograms lead to high computational demands and, because of redundancy, they pose high risks of model overfitting. Following Van Segbroeck et al. [13], a Gammatone (GT) filter bank [48] was therefore used to reduce the size of the STFT array along the frequency axis from 251 × 401 to 64 × 401 while simultaneously maintaining the key spectro-temporal features. It can be interpreted as a pooling operator using a re-weighting step, which is motivated by a comparison with filterbanks adopted to human auditory perception [49]. Therefore, we adapted the frequency distribution to make our method applicable to the auditory range of mice.

GT filter bank computations are provided in a MATLAB script by [50]. These computations were converted into the Python language for the present study. For each filter, a central frequency and bandwidth are required. The bandwidth and center frequency equations obtained in MUPET are also employed here (see S2 Text). In MUPET, the midpoint frequency parameter (Eq 2 in S2 Text) used to calculate the central frequencies was chosen as 75 kHz. The midpoint frequency can be interpreted as the frequency region where most information is processed [13]. Because the authors acknowledged that this value may not apply to all mice, we estimated the optimum value by calculating the median frequency (i.e., 63.5 kHz) from the FTs of all detected syllables, omitting FPs (S1 Fig). Then, in a pilot test, we updated this value to 68 kHz to minimize the information loss from USVs. The central frequency was calculated based only on the DEV data. A more detailed explanation of how to determine these two

parameters is given in the S2 Text (the Gammatone filterbank section). To further eliminate the background noise from the images, following MUPET, we calculated the maximum value between the Gammatone-filtered STFT pixels and the floor noise (10^{-3}). The logarithm of the output was smoothed using an auto-regression moving-average (ARMA) filter [51] with order 1 (S2 Text). Finally, a median filter [52] was applied to remove stationary noise. The product of the pre-processing is a smoothed, denoised spectrogram with reduced size of 64*401, called Gammatone spectrograms (GSs). Fig 2 shows the GSs of five samples of each 12 studied classes. These samples have the minimum Manhattan distance to other members of each class.

CNN classifier. For our study, we used convolutional neural networks (CNNs), a particular form of the deep neural network [53] first introduced by [54] and further developed by [55]. A brief description of how this model works, how we implemented it, and how the DSQ classifier is retrained is provided in the S2 Text.

We have evaluated our classifier for different values of its hyperparameters and architecture to achieve the best performance. These parameters were the number of convolution layers (i.e., 3, 4, and 5), the number of filters in each convolution layer (16, 32, 64, and 96), the kernel size in the first convolution layer (i.e., (3, 3), (5, 5), and (3, 18)), the drop out percentages (i.e., 0.5, 0.6, and 0.7), the size of dense layers (i.e., 32, 64, and 128), and the learning rate (cosine annealing learning rate scheduler [44], fixed learning rate = 10^{-3}, and decreasing learning rate = (10^{-3} to $5 * 10^{-6}$)).

In this study, we used categorical cross-entropy (CCE) [53,56]. For the reduction of the overfitting [57] L^2 regularization [58] is added to CCE. To optimize the loss function, we used the stochastic gradient descent with Nesterov momentum [59] and we initialized the weights of the convolution and FC layers using the He-initialization [60]. To reduce overfitting and to promote the generalizability of the model [61], we performed the augmentation of the training dataset using random shifts of width and height by 10%. We chose the following architecture for the classifier based on the comparison of the model performance on DEV_validation data (S4 Table).

The architecture of our network is shown in Fig 3. In this depiction, e.g., Conv2D (32, 3*18) denotes a 2-dimensional convolution layer with a kernel size of 3*18 and 32 filters. The FC (128) is a fully connected layer with 128 neurons. After two FC layers, a dropout layer with the probability of 0.5 is used. This step reduces the risk of overfitting [62].

Imbalanced data distribution. As shown in Table 1, the DEV_train dataset is significantly unbalanced, with 69 occurrences of the 'c3' and 4849 of the 'FP' class, a typical situation in real applications of machine learning. To investigate how this uneven distribution affects

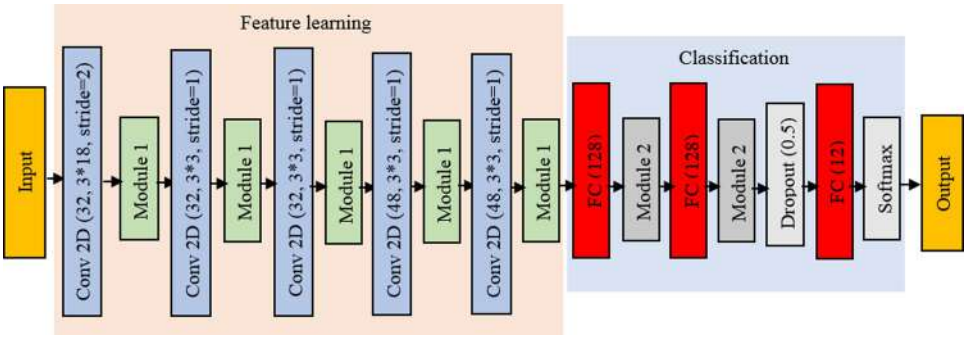

Fig 3. Classifier architecture. Module 1 consists of the following layers: Batch normalization + ELU + Maxpooling 2*2. Module 2 consists of the following layers: Batch normalization + ELU. Conv2D (32, 3*18) is a 2-dimensional convolution layer with a kernel size of 3*18 and the number of filters is 32. FC (128) is a fully connected layer with 128 neurons.

https://doi.org/10.1371/journal.pcbi.1010049.g003

the performance of the classifier, in addition to the original DEV_train data, we fit the model with resampled DEV_train using three different approaches.

1. In the first approach, the original input data are bootstrapped *10* times to increase the generalizability and reliability of the classifier [63,64]. Here, we used bootstrap to increase (decrease) the randomness (variance) during the model development. In each bootstrap iteration, samples are drawn from the original dataset with repetition, so some samples may appear more than once or some not at all. Then, we fitted a model for each bootstrapped dataset. The final model performance was evaluated by the average over the 10 models. Bootstrapping reduced the ratio of data imbalance from 76 to 4.
2. In the second scenario, all classes, except the classes 'c3' and 'us', which only have a maximum data number of 69 and 74, are randomly under-sampled to 124 samples.
3. In the last scenario, all classes, except 'FP' and 'up', are over- and under-sampled to the number of samples of the majority class, i.e., 4849. We used the Synthetic Minority Oversampling Technique Edited Nearest Neighbor (SMOTEENN) [65] and the number of neighbors was selected as 3.

To tackle the imbalanced distribution, during the model training we also weighed the loss function inversely proportionally to the number of class members [66] for the original, bootstrapped, and under-sampled data using the following equation:

$$WCCE = -\sum_{i=1}^{C} cw_i \, y_i \, log \, log \, (p_i), \;\; where \; cw_i = \frac{N}{c * n_i} \tag{1}$$

N and n_i are the total number of samples and class members. CCE [53,56] in Eq 4 in the S2 Text was updated to WCCE.

Model ensemble. The weights optimized on a particular dataset are not guaranteed to be optimal (or even useful) for another dataset. At the same time, different machine-learning algorithms can lead to different results even for the same dataset. In ensemble methods [45] the final output is taken from combining the outputs of different models and thus reducing the variance of the classifier output. Rather than training a model from scratch for different sets of hyperparameters, we produced 5 trained models during the training of a single model using Snapshot Ensemble with cosine annealing learning rate scheduler [44]. The use of the Snapshot Ensemble does not add complexity to the classifier, whereas it does help to take advantage of ensemble learning without needing to train additional models. The ensembles were trained consecutively, so the final weights of one model are the initial weights of the next. In this approach, the CNN weights are saved at the minimum learning rate of each cycle (S2 Fig), which occurs after every 40 epochs. To determine the best combination of these 5 models, we have cross-validated 4 approaches: 1) using the predictions of the 5th model, 2) using the average prediction from the last 3 models, 3) combining the predictions of the last 3 models by Extreme Gradient Boosting Machines (XGBMs) [67], and 4) combining the predictions of all 5 models using XGBMs. In explaining the third and fourth methods, instead of taking the average of the predictions (used for the second method), the predictions of the last three and five models of the DEV_validation data together with their ground truth are used for training the XGBMs. In this case, the final output of the classifier is the output of XGBMs.

Thus, to develop our classifier, these four ensemble methods were applied for each resampling approach namely under-sampling, over-sampling, and bootstrapping, and for the original data.

Inter-observer reliability (IOR). Our ground truth (or 'gold standard') was based on manual classification by researchers, and we used two independent observers to classify USVs

and then to evaluate our ground truth, we evaluated the reliability of our ground using class-wise inter-observer reliability (IOR). The first 100 USVs of 10 sound files were manually classified into 15 USV types by two of the authors, and both have much experience (Marconi et al. [6],Nicolakis et al. [7],Zala et al. [32]). We used five arbitrarily selected sound files from the DEV dataset and all five sound files used for the EV dataset (EV_wild and EV_lab). Both observers were blind to their respective labels and the original labels used for the development or evaluation of our classifier. The USV labels were extracted and exported into *Excel* files. The exported parameters included the start time, end time, and USV type of each vocalization. Then, the labels from both observers were aligned according to the start time of each segment. Thus, vocalizations with the same starting time were compared between the two observers. Segments that were labeled as false positives by the observers but detected by A-MUD as candidate USVs, were included; and segments that were labeled as unclassified ("uc") were excluded from the analyses. Segments classified as the same type by both observers were scored as 'agreement'. Segments that were either detected by only one observer or were classified into a different class were scored as 'disagreement'. Then, we calculated the percentage of correctly classified USVs by both observers, reported as IOR. We calculated the IOR for DEV and EV data for all segments (including FPs), and when including and excluding USVs detected by only one observer and not the other (i.e., labeled as 'missed' USVs). In addition to the original data, we calculated the IOR and F1-score when excluding 's' and 'us' classes, to evaluate how these two classes affected the IOR, and when pooling the original data into 12, 11, 6, 5, 3, and 2 classes, respectively, to compare the IOR and F1-score with the performance of our classifier.

Results

Comparing detection algorithms

Fig 4 shows the performance (TPR and FDR) of the four detection tools, MUPET, (pretrained) DSQ, USVSEG, and A-MUD (S1–S4 Data). A-MUD was tested using its default parameters, whereas the others were implemented using the combination of parameters that provided the

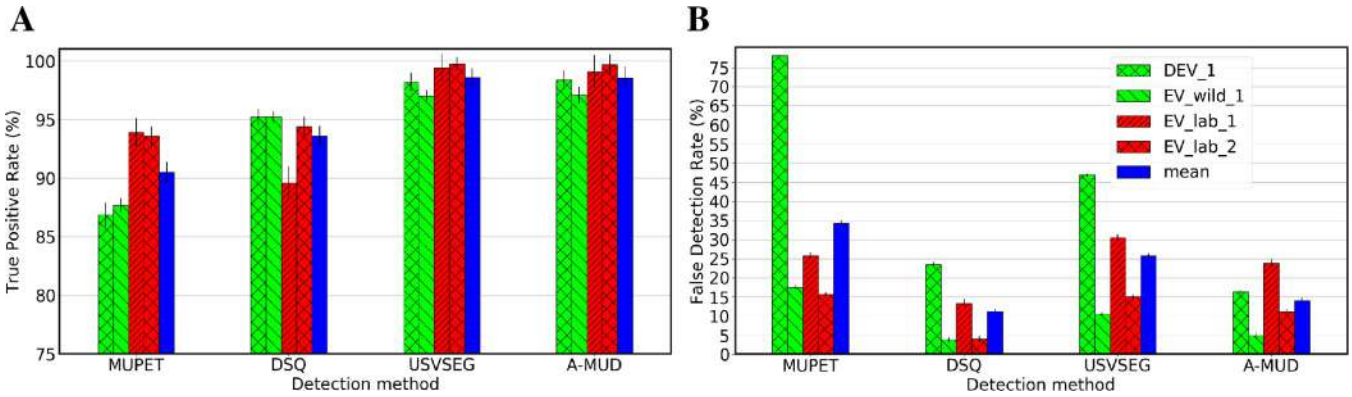

Fig 4. Best performance of four USV detection methods for four recordings. (A) The True Positive Rate shows the ratio of the number of USVs correctly detected to the total number of manually detected USVs * 100. (B) The False Detection Rate shows the ratio of the number of unwanted sounds (noise) incorrectly detected as USVs to the total number of detected elements * 100. Error bars represent the estimated variance calculated from the bootstrap resampling method. MUPET was implemented with the noise-reduction parameter set at 2, minimum syllable duration of 5 ms, and a minimum frequency of 30 kHz [13]. DSQ used its detection with the short rat call_network_v2 network with a high "recall" parameter [16]. USVSEG applied its detection with the threshold parameter set at 2.5, the minimum gap between syllables at 5 ms, and the minimum length of USVs at 4 ms [29]. A-MUD was run using its default parameters [30]. The legend shows the four recordings that were compared for each method (i.e., laboratory mice vs wild mice for both DEV (i.e., DEV_1 and EV_wild_1) and EV datasets (i.e., EV_lab_1 and EV_lab_2) and the mean of these four recordings. DEV_1 and EV_lab_1 are examples of low-SNR recordings and EV_lab_2 is an example of high-SNR recording.

https://doi.org/10.1371/journal.pcbi.1010049.g004

best results for the chosen dataset. We also compared the performance of these methods using other parameters (S3 Fig).

A-MUD (with the default parameters) and USVSEG (with the tuned parameters) correctly detected the largest number of USVs (TPRs were all >97%) whereas MUPET had the lowest mean TPR (90%) (Fig 4). A-MUD and USVSEG also provided the best performance when evaluating the detection of USVs from low-SNR recordings (DEV_1 and EV_lab_1, which include USVs from wild-derived and laboratory mice, respectively). We evaluated the performance of USVSEG using recordings of laboratory and wild mice and found that it has a higher TPR for laboratory mice using any of its settings (S3 Fig). This result is likely because this method is primarily parameterized and evaluated based on recordings of laboratory mice. In contrast, A-MUD (with the default parameters) has a high TPR for both types of data, despite that it was parameterized and evaluated using recordings of wild mice only. The presence of faint USVs (in EV_wild_1) had little effect on the TPR for most methods, except for MUPET. The TPR for this method was reduced from 93% to 86% when recordings contained faint USVs. By comparing FDRs, we found that DSQ had the lowest error rates, though it has fewer mean TPR than A-MUD (93.6% vs 98.6%). This shows that users need to be aware of the limitations of using these tools (like DSQ) without re-training and fine-tuning.

Visual inspection of the results indicates that the highest variance of TPR (~ 1.2%) and FDR (~1%) when comparing all the tools occurred in the data EV_lab_1. The TPR of USVSEG reached A-MUD (98.6%), whereas it underperformed A-MUD in terms of FDR (25.7% vs. 13.7%). Also, by examining the output of USVSEG, we found that most of its FPs are fragmented faint USVs, so they do not resemble FPs and, thus, must be manually removed from this group and assigned to the USVs.

Our results also show that A-MUD and USVSEG underestimated the duration of USVs in wild mice and overestimated them in laboratory mice (S3 Text). The slopes (and intercepts) between USV duration estimated by the two tools and observations are not statistically different (permutation test, p-values $> .05$). These results can explain some of the errors in the classification of USVs because overstimulation (underestimation) may cause the inclusion of noise (removal of useful information) in the USV segmentations. Further investigation of this error is beyond the scope of this paper.

Selecting the architecture of the classifier

To develop our classifier, the detected elements were first manually classified into 12 types of USVs (ground truth). In addition to the original data, three types of resampling approaches were examined (under-sampling, over-sampling, and bootstrapping) to overcome the uneven distribution between USV classes. For each type of resampling, four model ensemble methods were applied to the outputs, which include the predictions of the last Snapshot ensemble ('sn'), the average prediction of the last 3 Snapshot ensemble models ('sn_avg_3'), and a combination of the predictions of the last 3 ('sn_xgb3') and 5 Snapshot ensemble models ('sn_xgb5') by XGBMs. To investigate the effect of snapshot ensemble and bootstrapping approaches on model performance, we considered the classifier trained using a learning rate of 10^{-3} (called 'single model' in Fig 5) and original data as the baseline. Fig 5 shows the performance of the models with different combinations of resampling and ensemble methods compared to the baseline model.

The comparison of F1-score obtained from baseline model (68.9±2.3%) and model trained using Snapshot ensemble (based on original data) (70.6±1.4%) shows the superiority of Snapshot ensemble. In addition, bootstrapping data (without using Snapshot ensemble) increased the F1-score by about 6% compared to the baseline model. The bootstrap and under-sampling

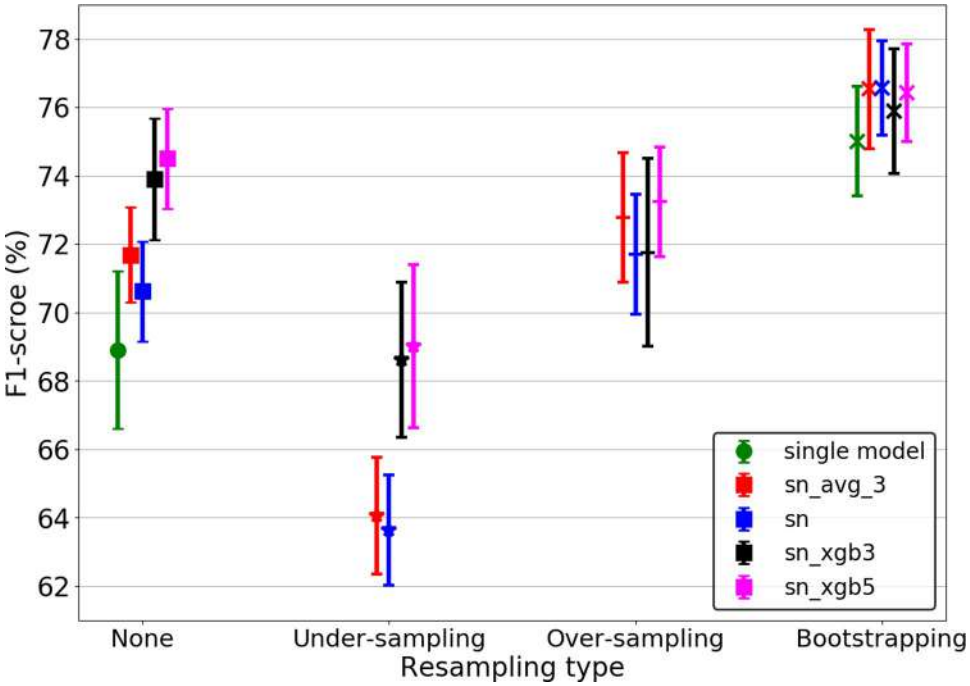

Fig 5. Performance of classifiers based on four resampling methods for four types of ensemble models. The single model performance is only displayed in two types of resampling (including 'none' and 'bootstrapping'), to better understand the effect of the 'bootstrapping' approach on the baseline model (which is based on original data and fixed learning rate). Using snapshot ensemble-based method, for each type of resampling (including 'none', 'under-sampling', 'over-sampling', and 'bootstrapping'), four ensemble models have been applied to the outputs last Snapshot ensemble ('sn'), the average prediction of the last 3 Snapshot ensemble models ('sn_avg_3'), and combining the predictions of the last 3 ('sn_xgb3') and 5 Snapshot ensemble models ('sn_xgb5') by XGBMs. The mean ± STD of macro F1-score of test datasets over 8-fold cross-validation are shown.

https://doi.org/10.1371/journal.pcbi.1010049.g005

methods always had the highest and lowest average F1-score, respectively, regardless of the ensemble method. Using the last model obtained from the Snapshot ensemble gave the highest average F1-score (76.6%) with bootstrapping. 'sn_xgb5' outperformed the other ensemble methods for the original data and two other resampling methods (under-sampling and over-sampling). The last model of the Snapshot ensemble also provided the lowest variation in bootstrapped data (STD = 1.4%). The differences between the ensemble methods are not large if used together with bootstrapping.

Neither the under-sampling (F1-scores = 69%) nor the over-sampling (F1-scores = 73.5%) methods improved the performance of the model compared to the best model from the original data (F1-score = 74.5%). While this result is not surprising for the under-sampled case, the performance of the oversampling case shows that the variance is here not a problem for small classes. The poor performance of the model fed by under-sampled data can be attributed to the random discard of samples and thus the deletion of useful information. The over-sampling method may have failed to improve the model performance because the images produced by the SMOTEENN are very similar to the original data (S4 Fig) leading to model overfitting. As a result, the combination of bootstrapped data and the last Snapshot model (hereafter called *BootSnap*) provided the best classifier.

Next, we examined the class-wise performance of the best model for each combination of resampling and ensembling method, including original + 'sn_xgb5', under-sampled + 'sn_xgb5', over-sampled + 'sn_xgb5', bootstrapped + 'sn' (*BootSnap*), and baseline model. As shown in Fig 6, *BootSnap* improved the F1-scores of classes 'c2', 'up', 'ui', 'c3', and 'us' by

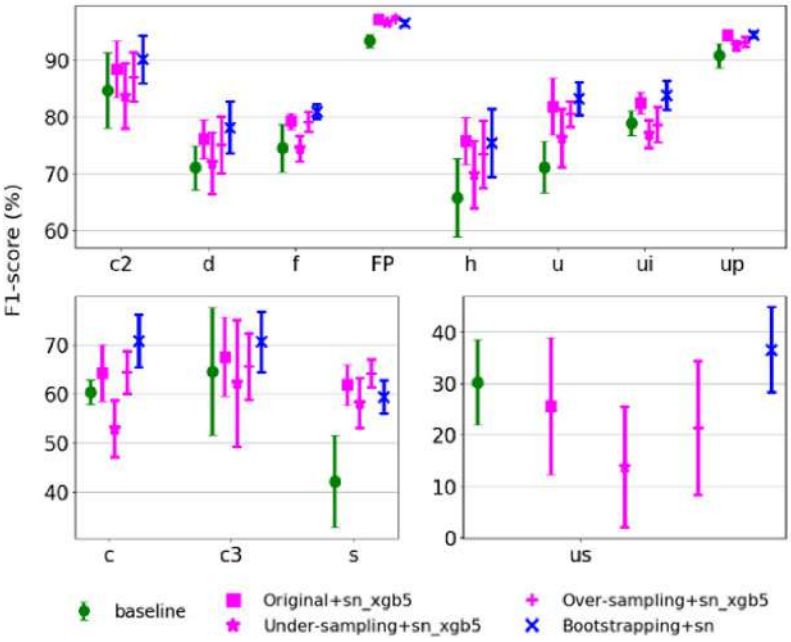

Fig 6. Performance of baseline model and the best model for each combination of resampling and ensemble methods. The performance of the baseline model (trained using a learning rate of 10^{-3} and without the use of resampling and ensemble methods) and the best model resulting from each ensemble method for different USV classes are shown. The mean ± STD of the class-wise macro F1-scores is based on the 8-fold cross-validation.

https://doi.org/10.1371/journal.pcbi.1010049.g006

around 5% and classes 'd', 'h', 'u', 'c', and 's' by around 10% or more. It also improved the F1-score of classes 'c' and 'c3' by around 5% and class 'us' by around 10%, compared to other combinations of resampling and ensemble methods. The number of classes 'c3' and 'us' in the original data is lower than in other classes, and bootstrapping seems to effectively increase the number of class members used during the model development. For classes, 'c2', 'd', 'f', and 'u', *BootSnap* increased the average macro F1-score by around 2%-3%. The classes 'FP', 'h', 'ui', and 'up' in the original + 'sn_xgb5' and *BootSnap* models have approximately equal average macro F1-score. Using the XGB output ensembling for bootstrapped data and SMOTEENN increased model complexity and did not improve the performance of the classifier.

Somewhat surprisingly, the average macro F1-score of the classes 'h' and 'ui' did not increase by bootstrapping, so it seems that the number of these data points is sufficient for our method. It appears that bootstrapping did not help only for the class 's', but the abundance of class members of 'up' and 'FP' in the original data defused the effect of bootstrapping. The average macro F1-score of *BootSnap* in the class 's' is about 2% less than in the model fed by the original data.

BootSnap also reduced the variation in the macro F1-scores for almost all USV classes, and the largest reduction in variation was for classes 'u', 'c3', and 'us'. However, the classes 'us' and 'c3' had the highest macro F1-score STD in all resampling methods; a result that might be due to the very low number of samples in these two classes (99 and 93 members respectively).

Evaluating *BootSnap* for classifying USVs

To evaluate the performance of *BootSnap* for different types of USVs, we generated a row-wise normalized confusion matrix (or error matrix) [68] (Fig 7). To prepare this matrix, we used

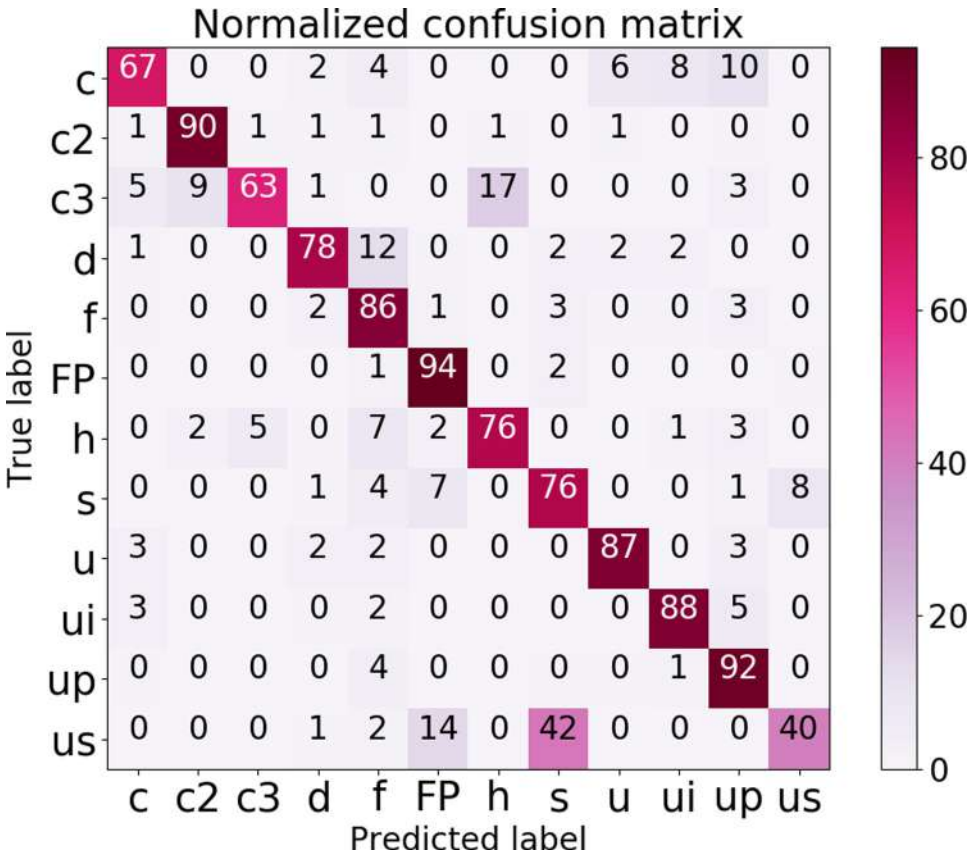

Fig 7. Confusion matrix of a 12-class classification using *BootSnap*. The main diagonal represents the recall of each USV class. The other values in each row are FNRs, which indicate the percentage of each class of USVs incorrectly labeled or classified.

https://doi.org/10.1371/journal.pcbi.1010049.g007

the manual annotations and predicted labels from *BootSnap* of the DEV_test dataset (of 8-fold).

This matrix shows that non-USVs ('FP') were classified with the highest recall (94%), which indicates that our model can successfully detect most falsely identified signals, and exclude them from further processing. It also shows that 40% to 92% of different types of USVs were accurately classified. The lowest recall was the 'us' class, and more than 40% of 'us' were mistakenly labeled as class 's' and 14% of the total members were assigned to the class 'FP'. The classification of 's' syllables (76%) was much more accurate than 'us', and the highest FNR value of this class ('s') belongs to the class 'us'. The misclassification of these two classes can be attributed to the use of the USVs length as the only feature used for manual classification, which is not reliable ('us' also shows much lower inter-observer repeatability in manual classification than other classes; S5 Fig). Class 'c3' had the second-lowest recall (63%), and most of its FNs were found with the classes 'h' (17%), 'c2' (9%), and 'c' (5%). These errors were due to the occurrence of harmonic patterns or faint jumps in the class 'c3'. The class 'c' had the third-lowest recall (67%), despite having a high number of members. The problem is that 'c' syllables were often mis-assigned due to their similarity in the spectrograms to 'ui', 'u', and 'up' types, which resulted in the highest FN rates in these three classes. Examination of the misclassified members of the class 'h' indicates that they were often assigned to the class 'f'. The highest portion of FNR (17%) of the class 'c3' is found with the class 'h'. The FNR of the class 'h' is 5%

with class 'c3'. In other words, the members of the class 'c3' are much more likely to be mistaken as the class 'h' than vice versa. It is because harmonic patterns are frequently seen with the second element (out of three elements) in the class 'c3', whereas the opposite rarely occurred in our recordings.

As shown in Fig 2, members of the class 'd' resemble the members of class 'f', which resulted in the class 'd' having the most FNs with the class 'f'. While there is no distinguished pattern of FNs distribution in other classes, it is important to note that FNs of the classes 'c2' and 'c3' mostly occur among themselves. Thus, the performance of the classifier is improved after pooling the 'c2' and 'c3' classes, as we show next.

Inference classification

Since it is unclear whether and how mice classify USVs, we report the performance of the best classifier (*BootSnap*) based on the different number of classes proposed in previous studies (Table 2). It is important to note that, unlike previous studies, we considered 'FP' as a target class. Since *BootSnap* was trained using 12 classes, we pooled different types of calls in various combinations, especially for the most similar types of syllables, to compare its performance with existing literature treating other numbers of classes. This comparison provides some insights into the classification of types of USVs by researchers.

The number of USV classes studied here ranged between 2 and 12 different types. As expected, classifying all 12 classes provided the lowest F1-score (76.6 ± 1.4%). In the next step, the classes 'us' and 's', which differ only in their duration, were pooled to form a new class, labeled 'short'. By combining these two classes, we expectedly found a significant increase in the F1-score (81.1 ± 1.6%). In addition, by combining these two classes, a significant number of 'us' and 's' types, which were mistakenly assigned as each other (Fig 7), were correctly classified as 'short'. In the next step, the classes 'up', 'd', 'f', 's', 'us', and 'u' were pooled to form the class called 'Rise', and the classes 'c3' and 'h' were included in the class 'split'. Aside from the class 'u', a common feature between classes pooled into 'Rise' was having no changes in their frequency direction. These classes were mostly false positives in the 12-member classification, and thus, after pooling, the F1-score significantly increased to 86.7±1.9%, compared to the 11-class classification.

Then, according to Wang et al. [69], the number of classes was reduced to five. We pooled the classes 'ui', 'c', and 'Rise'. These classes have no jumps in their spectrograms and thus the pooled new class was labeled 'no-jump'. Also, the classes 'h' and 'c3', which were pooled in the previous step into the class 'split', were separated again, but unlike the previous steps, the F1-score decreased (ca. 0.2%). This result might have been due to the separation of classes 'h' and 'c3' causing a large number of members of the latter class to be classified in the former

Table 2. BootSnap performance in classifying the DEV_test dataset in various combinations of classes.

Basis of classifications	# of classes	Different combinations of syllable types												Adapted from	F1-score (%)
original	**12**	FP	up	d	f	s	Us	u	ui	c	c2	c3	h	[32]*	**76.7±1.4**
Pool 's' and 'us'	**11**	FP	up	d	f	Short		u	ui	c	c2	c3	h	[5,46]	**81.1±1.6**
-	**6**	FP	Rise						ui	c	c2	split		[16]	**86.7±1.9**
Simple/complex	**5**	FP	no-jump								c2	c3	h	[69]	**86.5±2.2**
F- jumps	**3**	FP	no-jump								jumps and harmonics			[10]	**95.4±0.6**
FP/USV	**2**	FP	USV											-	**97.1±0.4**

*There are more references for 12 classes classification including [46], [5], [7], [6], and [9].

https://doi.org/10.1371/journal.pcbi.1010049.t002

Table 3. Comparison of pretrained DSQ (out of box model), retrained DSQ, and BootSnap performances. The performance metric is calculated based on supervised classification of USVs in EV_wild and EV_lab recordings. The values of macro F1-score (which is the average of F1-score over all classes) and class-wise F1-score (F1-score computed for each class) together with their resampling-based variance estimation are presented.

Scheme	macro F1-score (%)	Class-wise F1-score (%)					
		c	c2	split	FP	Rise	ui
pretrained_DSQ retrained_DSQ *BootSnap*	*EV_wild*						
	41±1 66±2 **67±1.6**	0±0 30±8 **35±6**	44±3 50±3 **58±3**	56±3 **83±1** 58±3	50±3 **98±1** 93±1	82±1 92±1 **92±0**	12±3 41±5 **66±4**
pretrained_DSQ retrained_DSQ *BootSnap*	*EV_lab*						
	49±1 40±1 **64±1**	24±3 8±2 **38±4**	43±3 48±2 **93±1**	74±1 53±2 **84±1**	66±2 54±2 **77±1**	69±2 **71±1** 61±2	16±4 6±2 **28±3**

https://doi.org/10.1371/journal.pcbi.1010049.t003

class (Fig 7). In the next step, all the members of the classes 'c2', 'c3', and 'h' were pooled into the class 'jumps and harmonics' and compared with the classes 'FP' and 'no-jump'. As mentioned before, all the FNs of the classes 'c2' and 'c3' were from each other (Fig 7), and as a result, pooling them in one class yielded an F1-score of about 95.4 ± 0.6%. Finally, we classified syllables and 'FP' into two separate classes, and this simple binary classification, which was mostly used in the USV detection step, was able to differentiate USVs from FPs with an F1-score of 97.1 ± 0.4%. These results again show how the performance of *BootSnap* depends upon the type of USV, and that pooling certain classes results in better accuracy. Pooling the USV classes in various combinations provides future researchers with a basis to compare their classifiers producing a different number of target classes with *BootSnap*.

Comparing *BootSnap* and DSQ: transferability to new datasets

We compared the performance of *BootSnap* to DSQ, which we consider to provide the state-of-the-art classification tool, and we used the EV_wild and EV_lab signals (Table 3 and S5 and S6 Data). As discussed in the Data section, the EV_wild data were obtained from wild-derived house mice (as in DEV), but under different conditions [6] and EV_lab data were from the MouseTube dataset (which is used for developing the original DSQ). To fairly evaluate the performance of the DSQ classifier, we have evaluated both the out-of-the-box (pretrained) and retrained models. In the out-of-the-box model, we have used classifier weights obtained from the original DSQ paper. In the retrained model, we used the classifier weights obtained from training the DSQ classifier using DEV data (S2 Text). Note that for *BootSnap* we have used the weights that were learned with the original training data (DEV_train). So, we did not retrain DSQ and BootSnap based on EV data, because this would then be a new learning approach and not an evaluation of the generalizability of the two approaches. *BootSnap* predictions were pooled into 6 classes, which included 'Rise', 'split', 'ui', 'c2', and 'c' (DSQ reported them as the output classes), and 'FP'. DSQ distinguishes FPs from USVs using a post hoc denoising network [16] before the classification step. For comparison, we considered 'FP' as one of DSQ's final outputs. Since *BootSnap* was developed based on 8 folds, we used the mode of 8 predictions to compare it with the DSQ output. It is also important to note that A-MUD was used to detect USVs in both algorithms to provide a fair basis for comparing the classification step in DSQ and *BootSnap* (this improved the average detection rate of DSQ by 5%).

As expected, all three methods–*BootSnap* and pretrained and retrained DSQ–performed better for the types of mice that were used to train them (wild mice for BootSnap, laboratory mice for the pretrained DSQ, and wild mice for the retrained DSQ, respectively; Table 3). DSQ had F1-scores of 41% (pretrained) and 66% (retrained) for wild mice and 49% (pretrained)

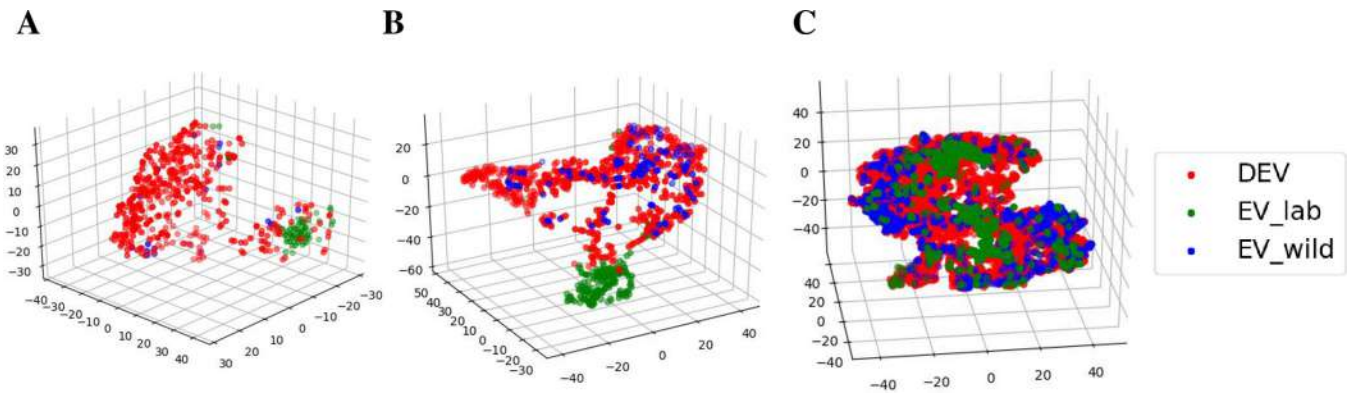

Fig 8. Scatterplots of USVs from three classes comparing different types of data and mice. 3-dimensional t-distributed stochastic neighbor embedding (t-SNE) representation of the classes (A) 'c', (B) 'ui', and (C) 'Rise' are shown. Colors indicate the dataset to which USVs belong.

https://doi.org/10.1371/journal.pcbi.1010049.g008

and 40% (retrained) for laboratory mice. *BootSnap* had an F1-score of 67% and 64% for wild and laboratory mice, respectively. Nevertheless, *BootSnap* outperformed pretrained and retrained DSQ for both types of mice overall. In terms of class-wise performance, *BootSnap* performed better than pretrained DSQ in nearly all the classes ('c', 'c2', 'split', 'FP', and 'ui', with higher F1-scores of 32%, 14%, 2%, 43%, and 54% for the EV_wild and higher F1-scores of 14%, 50%, 10%, 11%, and 12% for the EV_lab). The pretrained DSQ outperformed *BootSnap* for the EV_lab for one class, 'rise'. Retrained DSQ outperformed *BootSnap* in the classes 'split' and 'FP' in the EV_wild and the class 'Rise' in EV_lab. Regarding resampling-based variance estimation, the classes with a higher F1-score have less variance, which indicates that the result of that class (e.g., class 'split' in EV_lab) is more consistent.

Once again, an important point for developing and assessing the performance of a classifier is its generalizability, i.e., how well the model works when classifying data not used for the model development. In reviewing the above results, we observed that both DSQ and *BootSnap* had a relatively poor performance in the classification of the classes 'ui' and 'c'. Further examinations showed that the decline in their performance in these classes was due to the significant distance between new data and their training data. This distance is best seen in the three-dimensional t-distributed stochastic neighbor embedding (t-SNE) [70] representation (using the initial dimension of 40, the perplexity of 50, and the number of iterations of 2000) on vectorized GSs from both the DEV and EV datasets shown in Fig 8. The F1-scores of 'ui' and 'c' classes were very low for both *BootSnap* and DSQ for laboratory and wild mice, still, *BootSnap* outperformed DSQ. In the class 'Rise', there was large overlap between the USVs of wild and laboratory mice, which is in contrast with the classes 'ui' and 'c' (Fig 8C). Thus, the performance of both models for this class was much better than for other classes.

As a side remark, not directly linked to the topic of this paper, let us note that considering the data representation in Fig 8, the data from wild mice (DEV and EV_wild) and laboratory mice (EV_lab) could be effectively clustered using the samples from the 'c' and 'ui' classes.

Inter-observer reliability

When calculating the inter-observer reliability (IOR), excluding 'missed' segments, for the DEV dataset (n = 631 segments from 5 soundfiles) (S7 Data), we found ca. 80% agreement between two independent observers and ca. 84% agreement for the EV dataset (n = 578 segments from 5 soundfiles) (S8 Data), when including all classes (Tables 4 and S5). The removal of the 'missed' segments from all class combinations has a larger effect on IOR in the DEV

Table 4. Interobserver reliability and resampling-based variance estimation for the subsets of DEV and EV datasets. IOR values (in percentage) are given for different combinations of classes. Two IOR values are presented for each data and each combination of classes: IOR including 'missed' segments and IOR excluding 'missed' segments.

	Interobserver reliability in various combinations of classes (%)							
Data	**Original**	**Excluding 's' and 'us'**	**12 classes**	**11 classes**	**6 classes**	**5 classes**	**3 classes**	**2 classes**
DEV	79.5±1.6	83.6±1.6	79.5±1.6	80.6±1.6	83.8±1.1	89.2±1.1	89.2±0.8	92.4±0.8
	85.6±1.4	87.4±1.4	85.6±1.4	86.8±1.2	90.2±1.2	96±1.2	96±0.9	99.5±0.4
EV	84±1.6	85.6±1.6	88.7±1.4	88.9±1.3	90.1±1.3	93.2±1.1	94.6±1.1	97.9±0.7
	85.7±1.4	86.4±1.4	90.5±1.4	90.6±1.2	92±1.2	95±1.2	96.5±0.8	99.8±0.5

https://doi.org/10.1371/journal.pcbi.1010049.t004

data than in the EV data. This is probably because most of the USVs in the DEV dataset have low-SNR or they have a lower amplitude compared to USVs in the EV dataset since the EV dataset includes the EV_lab files which usually have a high-SNR. So, in the EV data, the probability of error in the detection tool and observer is lower due to the presence of louder USVs.

Excluding the 'us' and 's' USVs increased the IOR to 84% for the DEV data (9% of the segments excluded) and to 86% for the EV data (3.6% of the segments excluded), respectively. A detailed comparison of the manual classification by the two observers (S5 Fig) showed that the USV types 'us', 's', 'up', 'u', 'h', 'c', 'c3', 'c2', and 'ui' in the DEV dataset and 'us', 's', 'up', 'h', 'c4', 'c5', and 'ui' in the EV dataset accounted for the highest disagreement between observers. The disagreement for the type 'us' was likely due to detection error since 'us' USVs have <5 ms duration and be easy to be overlooked by another observer in noisy recordings. If there is a disagreement in the length of USVs (due to faint USVs or background noise) between observers, an "us" might be classified as 's' and 's' USV might be classified as 'd' or 'us'. We observed a low number of 's' and 'us' types when analyzing the EV dataset, especially within the recordings from laboratory mice (9% of 'us' and 's' in the DEV dataset compared to 3.6% in the EV dataset). Additionally, there can be disagreement between the USV types 'up' and 'ui'. This error is likely to occur due to the threshold of 5 kHz to measure the frequency modulation and used to distinguish between 'up' and 'ui'. USVs with upward frequency modulation of >5 kHz ('up') often ends with a slight downward frequency modulation, which can be close to 5 kHz. USVs often have a lower amplitude at the start or the end of the vocalization, and sometimes it can be difficult to measure the exact frequency modulation in a spectrogram. In summary, the main misclassifications are between 1) 'us' and 's', 2) 'c3' and 'h', 3) 'c3', 'c2', and 'c', 4) 'c', 'ui', 'u', and 'up', and 5) 'd' and 'f'. Usually, the fuzzy transition between the types is the main problem in manual classification. Thus, although USV syllables are discrete, they are not all very discrete, especially when the USVs are classified into a large number of classes (e.g., more than 5 according to Table 2). These findings confirm that the main difficulties of *BootSnap* and the manual classification are similar.

In our datasets, errors in manual classification were mainly due to (i) high background noise, (ii) duration or frequency thresholds used to define USV types, (iii) low or high amplitude of USVs (iv), and "noisy" vocalizations with many frequency-jumps emitted by laboratory mice. The disagreement in the manual classification of certain syllable types highlights the importance of finding a biologically relevant number of different USV classes, which can be reliably differentiated with low error rates by different observers.

Similar to the *BootSnap* F1-score, the IOR (Table 4) and F1-score (Table 5) of IOR data improved as we pooled the classes into fewer groups. For example, the IOR improved from 6 to 5 classes classification in the DEV (from 84% to 89%) and EV (from 90% to 93%) datasets. The improved IOR to 89% (DEV) and 94% (EV) after pooling all USVs with or without frequency jumps suggests that this might be a potential classification method, as this is more

Table 5. F1-score of the DEV_test and subsets of DEV (DEV_IOR) and EV datasets (EV_IOR) for IOR calculation. F1-score values (in percentage) are given for different combinations of classes. The numbers provided for DEV_test is the same as the numbers in Table 4. They are presented here again for easier comparison. Since we do not have 'missed' segments in the DEV_test data, these segments are removed when calculating the F1 score of DEV_IOR and EV_IOR datasets.

Setting	F1-score in various combinations of classes (%)					
	12	11	6	5	3	2
DEV_test	76.7±1.7	81.1±1.6	86.7±1.9	86.5±2.2	95.4±0.6	97.1±0.4
DEV_IOR	73.4±4.5	77.6±4.7	81.8±5.2	80.3±5.6	91±2	99.2±0.4
EV_IOR	82.8±3.5	83.9±2.7	89.7±1.9	84.2±3.6	97±0.6	99.6±0.4

https://doi.org/10.1371/journal.pcbi.1010049.t005

reliable between observers compared to a classification using ≥12 USV types. Additionally, manual classification showed an agreement of 92% (DEV) and 98% (EV) when distinguishing between USVs and 'FP' segments. The IOR increased to 99.5% (DEV) and 99.8% (EV) when excluding 'missed' segments.

Table 5 shows that in nearly all combinations of classes, the F1-score of DEV_test data (calculated between ground truth and *BootSnap* output) is similar to the F1-score of EV_IOR and higher than DEV_IOR datasets. The F1-score of EV_IOR and DEV_IOR datasets are calculated between two observers' labels. It can be concluded that the value of the F1-score generally increases with the pooling of the classes, and *BootSnap* classifies USVs with approximately equal accuracy as humans.

Comparing *BootSnap* and DSQ: sensitivity to new classes

One of the main performance factors of a classifier is how well it deals with classes for which it was not trained. The DEV data does not contain samples from two classes, 'c4' and 'c5'. Therefore, to address this issue, we analyzed the performance of pretrained and retrained DSQ and *BootSnap* focusing on these two classes, which were present in EV_wild data.

The results show that *BootSnap* assigned 68% and 32% of the members of these two classes to the classes 'c2' and 'c3', respectively. It is noteworthy that both 'c2' and 'c3' classes represent jump-included USVs, which is also a prominent feature of the classes 'c4' and 'c5'. Pretrained DSQ (retrained DSQ) assigned 3% (0%), 13% (6%), 46% (93%), 3% (0%), and 35% (1%) of the members of the classes 'c4' and 'c5' to the classes 'c', 'c2', 'c3', 'rise', and 'ui', respectively. Although the class 'ui' is relatively similar to the 'c4' and 'c5' classes based on visual inspection (S6 Fig), the difference is that there is no jump in the class 'ui' to which pretrained DSQ incorrectly assigned a significant number of classes 'c4' and 'c5'. Thus, we conclude that *BootSnap* uses a measure of similarity more fitted to USVs than pretrained DSQ, assigning new class samples to the most similar classes in the training data. The retrained DSQ, like BootSnap, assigned mostly all members of the classes 'c4' and 'c5' to jump-included classes ('c2' and 'c3').

Discussion and conclusions

The most important and novel contributions of our analyses include the following. First, we evaluated the performance and generalizability of four detection methods with each other, and we also assessed their absolute performance using ground truth. Only a few detection tools have been compared in previous studies, and they did not use a ground truth, or if so, they had a very small sample size. For example, the data for our ground truth consisted of 40 times more samples than the pretrained DSQ detector (i.e., 4000 vs 100), and therefore, our results should be much more robust. We used recordings from both wild house mice (*M. musculus musculus*) and laboratory mice, whereas most USV detection tools are designed (or machines are trained) using data from one or a few strains of laboratory mice. We found that A-MUD provided better overall performance compared to other detection methods, and without the

need for any manual parameter tuning or custom training of the networks. Second, we developed *BootSnap*, a new method for USV detection refinement (removing false positives or data cleaning) and classification, and we compared its performance and ability to generalize to novel datasets with DSQ classifier. We found that our new classification method outperformed DSQ–both the pretrained and retrained model–in nearly all aspects, including USVs of both the wild and laboratory mice. Below we address our main results in greater detail.

Comparing USV detection tools

Our first aim was to compare USV detection methods and evaluate their relative and absolute performance. We used wild mice, as well as laboratory mice, and we also compared recordings that had background noise (DEV_1 and EV_lab_1 signals) and faint (EV_wild_1) elements. We found that A-MUD and USVSEG detected the largest number of actual USVs (TPRs were all $> 97\%$ with A-MUD's *default* parameters and with the adaptive optimal parameters of USVSEG). DSQ and MUPET had the lowest mean TPRs (94% and 90%, respectively), and the pretrained (out-of-the-box) DSQ detector had the lowest FDR. Although DSQ has a lower FDR than other methods, it failed to detect ca. 6% of USVs on average, and to reduce FNR, one would need to train the detection network with labeled data, which would require manually resizing the window of each segment and its label (noise / USV). Although this can be done graphically in DSQ, it ultimately requires much manual intervention (user input).

USVSEG had a somewhat higher TPR for laboratory mice using any of its settings (99%) than wild mice (96%), and this is likely because USVSEG was primarily developed based on recordings of laboratory mice. A-MUD was parameterized using recordings of wild mice, though it still had high TPRs for both types of data, indicating that it is more generalizable than USVSEG. Unlike A-MUD, which was implemented using its default parameters, USVSEG has different performances for different parameter inputs. For example, in USVSEG, the use of the threshold parameter of 2.5, the minimum gap between syllables of 30 ms, and the minimum length of USVs of 4 ms leads to a significant increase in TPR for wild mouse data (above 90%) and a decrease in TPR for laboratory mice data (approximately 62%). While using a gap of 5 ms leads to improved TPR in both data sets. Another point is that the developers of USVSEG have suggested the user change the threshold parameter between 3.5 and 5.5. But we obtained the best performance of USVSEG for the wild mice data when the threshold was set to 2.5. We compared the performance of USVSEG and A-MUD for estimating the duration of USVs. Both methods underestimated the duration of USVs in wild mice and overestimated them in laboratory mice.

We compared how USVSEG and A-MUD detect USVs to better understand how these methods differ. USVSEG detects USVs using the following steps:

1. it calculates spectrograms using the multitaper method, which smooths the spectrogram and reduces background noises;
2. it flattens the spectrogram using cepstral filtering, which is performed by replacing the first three cepstral coefficients to zero and subtracting the median of the spectrogram (flattening eliminates impulse and constant background noises); and
3. it estimates the level of background noise to make a threshold for the resulting spectrogram.

In contrast, A-MUD (version 3.2) detects USVs using the following steps:

1. it applies an exponential mean to the spectrograms to reduce the noise contribution;

2. it estimates the envelope of the spectrograms using 6–8 cepstral DCT coefficients;
3. it computes the segmentation parameters, which are the amplitudes (m1-m3) and frequencies (f1-f3) of the three highest peaks in the spectrum for each time position; and
4. it searches for a segment based on 4 threshold values.

The reason that A-MUD (version 3.2) and USVSEG outperformed MUPET is presumably that A-MUD uses flattening rather than spectral subtraction for denoising. On the other hand, it seems that USVSEG in some cases leads to the failure of detection of ultrashort USVs, the false detection of two USVs as a single USV, the false segmentation of one USV as two or more USVs.

To summarize, A-MUD provides better overall performance compared to other methods for detecting USVs in audio recordings and without the need for any parameter tuning or custom training of the networks. For these reasons, we utilized A-MUD for our subsequent USV detection.

Comparing USV classification methods

Our second aim was to develop a new method for USV detection refinement and classification and compare its performance and generalizability with DSQ. To develop the classifier and to overcome the uneven distribution of classes, we examined three types of resampling approaches, under-sampling, over-sampling, and bootstrapping. For each type of resampling, four model ensemble methods were applied to the outputs: the predictions of the last Snapshot ensemble; the average prediction of the last 3 Snapshot ensemble models; and a combination of the predictions of the last 3 and 5 Snapshot ensemble models by XGBMs. We found that the differences between the ensemble methods are not large if used together with bootstrapping. This result can be interpreted in such a way that the ensemble of the models derived from bootstrapped data is already compensating for the uneven distribution statistically. We used bootstrapped data and the last model of snapshot ensembles as the best classifier ('*BootSnap*'). The classifier had the highest errors with classifying ultrashort ('us') USVs mainly due to their similarity with 's' USVs. These USVs do not differ qualitatively, they are not actually different syllable types, as they differ only in length. Another classification error was due to confusing 'c' and 'c3' syllables. The low recall in classifying 'c3' syllable types was likely due to their small number used for training, and also because some members have a harmonic element, much like 'h' types. The similarity in the spectrograms of 'c' to other classes as 'ui', 'u', and 'up' classes lead to errors in the classification of this class. On the other hand, the model classifies classes 'up', 'FP', and 'c2' with a recall higher than 90% and classes 'ui', 'u' and 'f' with a recall of more than 85%. These classes have a relatively larger number of members compared to other classes ('us' and 'c3') and their spectrograms are relatively different from each other. The overall F1-score of the model increased from 76.7% to 81.1% by pooling 's' and 'us' classes, which resulted in a more robust classification.

We compared the performance of *BootSnap* to the pretrained (out-of-the-box) and retrained DSQ classifier, which is currently the state-of-the-art classification tool. DSQ is a user-friendly and straightforward tool for analyzing mouse vocalizations and the user can train it for their data without the need for programming. In this analysis, however, we examined its generalizability and its out-of-the-box usability for novel data, as most users currently use this tool. It uses a 6-member syllable classification that includes 'Rise', 'split', 'ui', 'c2', 'FP', and 'c' types (i.e., a simpler classification approach based on 6 classes, Table 3). USVs from wild mice as well as laboratory mice were used to evaluate the generalizability of these two classifiers. As expected, in the *BootSnap* classifier (and in the retrained DSQ classifier, as well), the

closer the data is to the training domain, the better the overall performance. It has 87±1.9% F1-score for 6-class classification of USVs on DEV_test data (Table 2), but about 65% F1-score for EV datasets. We found that our new classification method outperformed both pretrained and retrained DSQ classifiers in nearly all aspects, including USVs of both the wild and laboratory mice (macro-F1 score of 66% vs 47% and 49%). Again, it is important to emphasize that the performance of retrained DSQ was worse than pretrained DSQ for EV_lab. The main reason for its reduced performance is likely due to the absence of laboratory mice in the DEV data, which would indicate that DSQ is less generalizable than BootSnap. This difference in performance is mainly because the DSQ classifier was developed using an architecture similar to our baseline model fed by high-SNR data, compromising its performance with new low-SNR recordings. In contrast, we used low-SNR data to develop our classifier and aimed to enhance its ability to generalize. We also used the Ensemble learning method, which is based on the Snapshot Ensemble and Bootstrapped input data for training the classifier. In Ensemble learning, base models are combined to prevent the final model from either overfitting or underfitting, making the model more stable and generalizable. So, the novelty of *BootSnap* comes from a clever combination of bootstrapping approach, snapshot ensemble, and Gammatone spectrograms.

BootSnap and the retrained DSQ classifier showed better performance than the pretrained (out-of-the-box) DSQ classifier in assigning new class samples to the most similar classes in training data. For example, our results show that the *BootSnap* retrained DSQ classifier assigned all instances with more than 3 jumps (similar to those not found in the training data) to similar classes with less than 3 jumps. However, the pretrained DSQ classifier allocated 30% of these new samples to the class without any jumps. The success of *BootSnap* as well as the retrained DSQ classifiers in assigning new class samples is due to the similarity of the data used for their development and EV_wild data. Our method also detects noise in new data much more accurately (F1-score of 93±1% vs. about 50±3% for EV_wild and 77±1% vs. 66 ±2% for EV_lab), and thus it is more useful for low-SNR data, which is a common challenge for USVs studies–especially studies aiming to record animals under social contexts. Also, *BootSnap* requires less user intervention to classify USVs, as for USVs classification using DSQ the user must first modify the frequency interval of USVs and then apply the classifier on them. But in *BootSnap*, after performing the detection by A-MUD, the whole interval of 20 kHz-120 kHz is used for classification. Another advantage in using *BootSnap* is that it is based on open-source software (Python) and, thus, it is free of charge, whereas DSQ is based on proprietary software (MATLAB), and requires the purchase of required licenses.

While completing the final draft of our present manuscript, a new tool, called 'VocalMat' [71], was published that detects and classifies USVs into 11 categories. The VocalMat classifier is trained on the USVs of mouse pups (5 to 15 days old) of both sexes of several inbred strains, including C57BL6/J, NZO/HlLtJ (New Zealand Obese), 129S1/SvImJ, NOD/ShiLtJ (Non-obese Diabetic NOD), and PWK/PhJ (descendants from a single pair of *Mus musculus musculus)*. It was developed using USVs in the frequency range of 45 kHz to 140 kHz. After contrast enhancement and applying several filters, the authors calculated the spectrogram (with the size of 227*227) of 12954 detected elements. Its classifier is the AlexNet model [72], which was pretrained on the ImageNet dataset. Like other studies, this classifier was not compared with other USV tools and the results on its generalizability were not provided. We evaluated the performance of VocalMat on its test data and found that the average class-wise accuracy is 79%, whereas *BootSnap* yielded an average class-wise accuracy of 83% for classifying DEV_test elements into 11 classes. The differences in the performances of these tools could be due to differences in the test data used for evaluation.

Evaluating ground truth: inter-observer reliability (IOR)

To our knowledge, this is the first time that the class-wise inter-observer reliability (IOR) of the ground truth, used to assess machine performance, has been evaluated. According to the IOR results, the agreement between two observers in the DEV and EV dataset was 80±1.6% and 89±1.4%, respectively. The mentioned values are related to the classification of segments into 12 classes. The USV classification was based both on A-MUD detections and on segments that were missed by A-MUD but manually detected by one or both observers. A closer look at the results reveals that mislabeling members of the classes 'us' as 's', 'ui' as 'up', and 'c' as 'ui' and to a lesser degree as 'up', and vice versa, is very likely. The reason for these errors in manual classification is their sensitivity to the arbitrary threshold (based on duration or modulation frequency) used in their definitions. In addition, the mislabeling can also occur in class 'h', as this class may include faint harmonic elements. Hence, part of the classification error of automated classification can be attributed to the error in the manual labeling of segments. However, any of these classes can be pooled to improve classification (from 80±1.6% using a 12-class classification to 84± 1.1% using 6-classes or to 92±0.8% using 2-class classification, see DEV dataset in Table 4), and such pooling also improved the F1-score of *BootSnap* (F1-score changed from 77% of 12-class classification to 87% of 6-class and 97% of 2-class classification, Table 2). Thus, pooling some types of USVs together improves the accuracy of *BootSnap*, which is expected since some types are very similar to each other. Consequently, *BootSnap* can be expected to perform better when classifying fewer types of syllables and that *BootSnap* can classify USVs with an accuracy similar to the results obtained from human inter-observer reliability. It is no surprise that the particular USVs that *BootSnap* does not classify well are the same ones that humans fail to show consistency. This result suggests that these types of USVs could just be human inventions, though it is certainly still possible that mice might treat them differently.

Outlook on USV classification

Our USV classification method is supervised, as with other models, and if users want to retrain the algorithm using their own recordings of mice, then manually labeled data must be provided. And despite the outperformance of *BootSnap* over DSQ, *BootSnap* still has difficulties with classifying new data containing complex USVs (with no jumps), u-inverted, and 1-jump USVs. Considering that our best model is based on the bootstrap technique, the computation time increases as the number of bootstrap iterations increases. By default, 10 repetitions are used for *BootSnap*, which means that *BootSnap* calculations will be 10 times slower than similar models. Because manual labeling of data is a difficult and time-consuming task, it is important to be able to apply a model trained on a single data source to other data as well. So, to further improve the generalizability of a classifier, in addition to implementing the bootstrap technique, we will increase the number of samples in the future by using more recordings. We expect that this approach will increase the predictive power of our classifier.

In summary, our ultimate goal was to develop an automated algorithm that provides an out-of-the-box method whose performance is as good or better than a human observer (manual classification). The human F1-score in EV data was higher than the F1-score of the out-of-the-box *BootSnap* model (89.7±1.9% vs 67%, respectively). Although *BootSnap* does not yet achieve our original aim, this first version provides an advance, as it outperforms other methods, including the state-of-the-art model (DSQ; 47% pretrained and 49% retrained). This leaves room for future research.

We emphasize that USVs have been classified by human researchers based on visual inspection of spectrograms or statistical clustering models, and very little is known about whether or

how mice can discriminate most of the various types of USVs that have been proposed. Mice can distinguish frequencies that differ by only 3% [73], and they can be trained to discriminate between simple versus complex USVs, and among certain variations in shape and frequency [74]. They can be trained to discriminate among USVs depending upon their spectro-temporal similarity, and they discriminate complex calls and up-shapes, but not u-shaped calls [75]. Mice fail to discriminate between synthetic sounds with different shapes, i.e., 'up'- vs. down-shapes [76]; however, the shapes of these synthesized sounds were very different from mouse USVs, and may have lacked characteristics that mice use for discrimination. More studies are needed to describe USVs produced in different contexts, and also determine whether mice can discriminate among different types of USVs. Such perception studies should include recordings with normal ranges of variation of syllable types within and between each category (i.e., mice should be better able to discriminate between- versus within-syllable type variation). Until such studies are conducted, the various types of USVs that have been proposed would be more accurately labeled as *USV variants* or *putative call types*.

Supporting information

S1 Table. Types of rodents and recording contexts used in different studies.
(XLSX)

S2 Table. The evaluated parameters for different USVs detection tools. Min-f and max-f in MUPET stand for min-frequency and max-frequency. Min-l in USVSEG, MUPET, and A-MUD stands for min-length.
(XLSX)

S3 Table. Definition of classes used in the labeling. Note that the number of members differs before and after down-sampling. F_e is the end frequency, F_s is the start frequency, F_{max} is the maximum frequency, and F_{min} is the minimum frequency. The number of members of each class corresponds to the DEV dataset.
(XLSX)

S4 Table. Performance of the classifier on DEV_validation data using various hyperparameters.
(XLSX)

S5 Table. The number of samples of each class of the observer 1 in DEV and EV subsets for IOR calculation. In the DEV sub-dataset (n = 5 soundfiles), there are very few samples, i.e., 2 from the classes 'c' and 'c4', 4 from the class 'c3', 5 from the class 'u' and 9 from the class 'h', thus the results of these classes are not very reliable. We found similar results in the EV sub-dataset (n = 5 sound files) where there are very few samples from the classes 'us', 'd', 'c', and 'c5'.
(XLSX)

S1 Text. USV classification literature review.
(DOCX)

S2 Text. Supplementary information on data and method.
(DOCX)

S3 Text. Comparing the estimated USV duration by USVSEG (using the optimal parameters) and A-MUD with the observed USV duration.
(DOCX)

S1 Fig. (A) Distribution of USVs Frequency Track (FT) values, extracted by A-MUD. The FT values are related to all detected syllables, omitting false positives. (B) The frequency response of 32 Gammatone filters (we have used 64 filters, but for simplicity 32 filters are plotted here) at the frequency range of 20 kHz to 120 kHz. (C) Two examples of the USVs spectrogram before (top row) and after applying the Gammatone filter and post-processing step (bottom row). This image shows that by applying the preprocessing steps on the spectrogram, the important information of the USVs is not lost, even though the size of the images is reduced from 251×401 to 64×401.
(TIF)

S2 Fig. Schedule scheme used for the learning rate. Using this scheme of the learning rate, the final weights of the model at every 40 epochs are the initial weights of the model in the next epoch. In this approach, the CNN weights are saved at the minimum learning rate of each cycle, i.e., at every 40 epochs.
(TIF)

S3 Fig. (A) true positive rate (TPR) and (B) false detection rate (FDR) of detection tools. In the main text, we compared the performance of 4 USV detection tools (USVSEG, A-MUD, DSQ, and MUPET) in a setting (i.e., input parameters) of which the selected parameters lead to their best average performance for the four-given data (DEV_1, EV_wild_1, EV_lab_1, and EV_lab_2). Here, we compared the performance of these methods using all the combinations used for their parameters (S2 Table). If we want to compare the best performance of each detection tool with the best performance of others, A-MUD and with a slight difference, USVSEG are in the first and second places, followed by DSQ and MUPET.
(TIF)

S4 Fig. Samples produced by Synthetic Minority Oversampling Technique Edited Nearest Neighbor (SMOTEENN). In the model design section, we used various approaches to deal with the problem of the imbalanced datasets, including using original, down-sampled, bootstrapped, and over-sampled data. The following figure presents the over-sampled data by SMOTEENN presented. The first column from the left is the original instance and the next columns are the resampled samples. The first, second, third, and last rows are from the classes 'c', 'c3', 'c2', and 'h', respectively. The images produced by the SMOTEENN are very similar to the original data, so, compared to the original data, this method did not help to improve the classifier performance.
(TIF)

S5 Fig. Agreement between two observers for two subsets of (A) model development (DEV) and (B) evaluation (EV) data. 'missed' segments are elements that are manually detected by only one observer. Both figures show high disagreement between the observers for both data in the 'us' and 'h' classes. In more detail, the amount of reliability in the DEV data in 'c3' and 'u' classes is very low. Differently, in the EV data, the reliability is less than other classes in 'c4' and 'c5' classes.
(TIF)

S6 Fig. Samples of USVs from the classes 'c4' and 'c5', USVs with 4 and 5 jumps, respectively. As mentioned in the results section (in the main text), the performance of a model is important when dealing with a new class. Because there was no sample of the 'c4' and 'c5' classes in the DEV data, we compared the output of the BootSnap and DSQ methods when the two classes were in the EV data. The following figure shows examples of members of these two classes in EV_wild data. *BootSnap* assigned 68% and 32% of the total members of these two

classes to the 2 and 3-jump included USVs, respectively. DSQ assigned the members of the classes 'c4' and 'c5' mostly to the 2 and 3-jump included USVs and 'ui'. Although the class 'ui' might be relatively similar to the 'c4' and 'c5' classes based on visual inspection, there is no jump in this class.
(TIF)

S1 Data. Data of 4 studied detection methods (USVSEG, MUPET, A-MUD, and DSQ) on DEV_1 recording.
(XLSX)

S2 Data. Data of 4 studied detection methods (USVSEG, MUPET, A-MUD, and DSQ) on EV_wild_1 recording.
(XLSX)

S3 Data. Data of 4 studied detection methods (USVSEG, MUPET, A-MUD, and DSQ) on EV_lab_1 recording.
(XLSX)

S4 Data. Data of 4 studied detection methods (USVSEG, MUPET, A-MUD, and DSQ) on EV_lab_2 recording.
(XLSX)

S5 Data. Data of EV_wild segments classification by BootSnap and pretrained and retrained DSQ compared to manual labels.
(XLSX)

S6 Data. Data of EV_lab segments classification by BootSnap and pretrained and retrained DSQ compared to manual labels.
(XLSX)

S7 Data. Data of interobserver reliability for the subsets of DEV datasets.
(XLSX)

S8 Data. Data of interobserver reliability for the subsets of EV datasets.
(XLSX)

Acknowledgments

We would like to thank Anton Noll for making A-MUD outputs available.

Author Contributions

Conceptualization: Reyhaneh Abbasi, Peter Balazs, Dustin J. Penn.

Data curation: Reyhaneh Abbasi, Maria Adelaide Marconi, Doris Nicolakis, Sarah M. Zala, Dustin J. Penn.

Formal analysis: Reyhaneh Abbasi.

Funding acquisition: Peter Balazs, Sarah M. Zala, Dustin J. Penn.

Investigation: Maria Adelaide Marconi, Doris Nicolakis, Sarah M. Zala.

Methodology: Reyhaneh Abbasi, Peter Balazs.

Project administration: Peter Balazs, Sarah M. Zala, Dustin J. Penn.

Resources: Reyhaneh Abbasi, Peter Balazs, Maria Adelaide Marconi, Doris Nicolakis, Sarah M. Zala, Dustin J. Penn.

Software: Reyhaneh Abbasi.

Supervision: Peter Balazs, Sarah M. Zala, Dustin J. Penn.

Validation: Reyhaneh Abbasi, Peter Balazs, Maria Adelaide Marconi, Doris Nicolakis.

Visualization: Reyhaneh Abbasi.

Writing – original draft: Reyhaneh Abbasi, Peter Balazs, Maria Adelaide Marconi, Doris Nicolakis, Sarah M. Zala, Dustin J. Penn.

Writing – review & editing: Reyhaneh Abbasi, Peter Balazs, Maria Adelaide Marconi, Doris Nicolakis, Sarah M. Zala, Dustin J. Penn.

References

1. Ehret G. Characteristics of vocalization in adult mice. Handbook of behavioral neuroscience. 25: Elsevier; 2018. p. 187–95. https://doi.org/10.1016/B978-0-12-809600-0.00018-4.
2. Brudzynski SM. Handbook of Ultrasonic Vocalization: A Window Into the Emotional Brain: Academic Press; 2018.
3. Heckman J, McGuinness B, Celikel T, Englitz B. Determinants of the mouse ultrasonic vocal structure and repertoire. Neuroscience & Biobehavioral Reviews. 2016; 65:313–25. https://doi.org/10.1016/j.neubiorev.2016.03.029 PMID: 27060755
4. Fischer J, Hammerschmidt K. Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication. Genes, Brain and Behavior. 2011; 10 (1):17–27. https://doi.org/10.1111/j.1601-183X.2010.00610.x PMID: 20579107
5. Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T+ tf/J mouse model of autism. PloS one. 2008; 3(8):e3067. https://doi.org/10.1371/journal.pone.0003067 PMID: 18728777
6. Marconi MA, Nicolakis D, Abbasi R, Penn DJ, Zala SM. Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures. Animal Behaviour. 2020. https://doi.org/10.1016/j.anbehav.2020.09.006.
7. Nicolakis D, Marconi MA, Zala SM, Penn DJ. Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success. Frontiers in zoology. 2020; 17:1–19. https://doi.org/10.1186/s12983-019-0343-8 PMID: 31911809
8. von Merten S, Hoier S, Pfeifle C, Tautz D. A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (Mus musculus domesticus). PloS one. 2014; 9 (5):e97244. https://doi.org/10.1371/journal.pone.0097244 PMID: 24816836
9. Musolf K, Meindl S, Larsen AL, Kalcounis-Rueppell MC, Penn DJ. Ultrasonic vocalizations of male mice differ among species and females show assortative preferences for male calls. PLoS one. 2015; 10(8). https://doi.org/10.1371/journal.pone.0134123 PMID: 26309246
10. Hoffmann F, Musolf K, Penn DJ. Ultrasonic courtship vocalizations in wild house mice: spectrographic analyses. Journal of ethology. 2012; 30(1):173–80. https://doi.org/10.1016/j.physbeh.2011.10.011 PMID: 22037196
11. Oppenheim AV, Schafer R, Buck J. Discrete-time signal processing. 2 ed. Upper Saddle River, NJ, USA: Prentice Hall: Pearson Education India; 1999.
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2 ed: Springer; 2009. p. 485–585.
13. Van Segbroeck M, Knoll AT, Levitt P, Narayanan S. MUPET—mouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. Neuron. 2017; 94 (3):465–85. e5. https://doi.org/10.1016/j.neuron.2017.04.005 PMID: 28472651
14. Chabout J, Jones-Macopson J, Jarvis ED. Eliciting and analyzing male mouse ultrasonic vocalization (USV) songs. Journal of visualized experiments: JoVE. 2017(123). https://doi.org/10.3791/54137 PMID: 28518074
15. Burkett ZD, Day NF, Peñagarikano O, Geschwind DH, White SA. VoICE: A semi-automated pipeline for standardizing vocal analysis across models. Scientific reports. 2015; 5(1):1–15. https://doi.org/10.1038/srep10237 PMID: 26018425

16. Coffey KR, Marx RG, Neumaier JF. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology. 2019; 44(5):859–68. https://doi.org/10.1038/s41386-018-0303-6 PMID: 30610191

17. Dou X, Shirahata S, Sugimoto H. Functional clustering of mouse ultrasonic vocalization data. PloS ONE. 2018; 13(5):e0196834. https://doi.org/10.1371/journal.pone.0196834 PMID: 29742174

18. Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Frontiers in behavioral neuroscience. 2015; 9:76. https://doi.org/10.3389/fnbeh.2015.00076 PMID: 25883559

19. Goffinet J, Brudner S, Mooney R, Pearson J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. Elife. 2021; 10:e67855. https://doi.org/10.7554/eLife.67855 PMID: 33988503

20. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+ tf/J mice during three types of social encounters. Genes, Brain and Behavior. 2011; 10(1):44–56.

21. Hoier S, Pfeifle C, von Merten S, Linnenbrink M. Communication at the garden fence–context dependent vocalization in female house mice. PloS one. 2016; 11(3):e0152255. https://doi.org/10.1371/journal.pone.0152255 PMID: 27022749

22. Seagraves KM, Arthur BJ, Egnor SR. Evidence for an audience effect in mice: male social partners alter the male vocal response to female cues. Journal of Experimental Biology. 2016; 219(10):1437–48. https://doi.org/10.1242/jeb.129361 PMID: 27207951

23. Matsumoto YK, Okanoya K. Mice modulate ultrasonic calling bouts according to sociosexual context. Royal Society open science. 2018; 5(6):180378. https://doi.org/10.1098/rsos.180378 PMID: 30110406

24. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (Mus musculus musculus). PloS ONE. 2017; 12(12):e0188647. https://doi.org/10.1371/journal.pone.0188647 PMID: 29236704

25. Matsumoto YK, Okanoya K. Phase-specific vocalizations of male mice at the initial encounter during the courtship sequence. PloS one. 2016; 11(2):e0147102. https://doi.org/10.1371/journal.pone.0147102 PMID: 26841117

26. Finton CJ, Keesom SM, Hood KE, Hurley LM. What's in a squeak? Female vocal signals predict the sexual behaviour of male house mice during courtship. Animal Behaviour. 2017; 126:163–75.

27. Warren MR, Spurrier MS, Roth ED, Neunuebel JP. Sex differences in vocal communication of freely interacting adult mice depend upon behavioral context. PLoS one. 2018; 13(9):e0204527. https://doi.org/10.1371/journal.pone.0204527 PMID: 30240434

28. Binder M, Nolan SO, Lugo JN. A comparison of the Avisoft (v. 5.2) and MATLAB Mouse Song Analyzer (v. 1.3) vocalization analysis systems in C57BL/6, Fmr1-FVB. 129, NS-Pten-FVB, and 129 mice. Journal of Neuroscience Methods. 2020; 346:108913. https://doi.org/10.1016/j.jneumeth.2020.108913 PMID: 32805316

29. Tachibana RO, Kanno K, Okabe S, Kobayasi KI, Okanoya K. USVSEG: A robust method for segmentation of ultrasonic vocalizations in rodents. PLoS One. 2020; 15(2):e0228907. https://doi.org/10.1371/journal.pone.0228907 PMID: 32040540

30. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Automatic mouse ultrasound detector (A-MUD): A new tool for processing rodent vocalizations. PloS ONE. 2017; 12(7):e0181200. https://doi.org/10.1371/journal.pone.0181200 PMID: 28727808

31. Binder M, Hernandez-Zegada CJ, Potter CT, Nolan SO, Lugo JN. A comparison of the Avisoft (5.2) and Ultravox (2.0) recording systems: Implications for early-life communication and vocalization research. Journal of neuroscience methods. 2018; 309:6–12. https://doi.org/10.1016/j.jneumeth.2018.08.015 PMID: 30118724

32. Zala SM, Nicolakis D, Marconi MA, Noll A, Ruf T, Balazs P, et al. Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female. PLoS ONE. 2020; 15(12):e0242959. https://doi.org/10.1371/journal.pone.0242959 PMID: 33296411

33. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009; 22(10):1345–59. https://doi.org/10.1109/TKDE.2009.191.

34. Wiley RH. The evolution of communication: information and manipulation. Animal behaviour. 1983; 2 (494):156–89.

35. Smith AA, Kristensen D. Deep learning to extract laboratory mouse ultrasonic vocalizations from scalograms. IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017: IEEE. https://doi.org/10.1109/BIBM.2017.8217964.

36. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence. 2002; 24(7):881–92. https://doi.org/10.1109/TPAMI.2002.1017616.

37. Goudbeek M, Cutler A, Smits R. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. Speech communication. 2008; 50(2):109–25. https://doi.org/10.1016/j.specom.2007.07.003.

38. Guerra L, McGarry LM, Robles V, Bielza C, Larranaga P, Yuste R. Comparison between supervised and unsupervised classifications of neuronal cell types: a case study. Developmental neurobiology. 2011; 71(1):71–82. https://doi.org/10.1002/dneu.20809 PMID: 21154911

39. Brandes TS. Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International. 2008; 18(S1):S163–S73. https://doi.org/10.1017/S0959270908000415.

40. Vogel AP, Tsanas A, Scattoni ML. Quantifying ultrasonic mouse vocalizations using acoustic analysis in a supervised statistical machine learning framework. Scientific reports. 2019; 9(1):8100. https://doi.org/10.1038/s41598-019-44221-3 PMID: 31147563

41. Premoli M, Baggi D, Bianchetti M, Gnutti A, Bondaschi M, Mastinu A, et al. Automatic classification of mice vocalizations using Machine Learning techniques and Convolutional Neural Networks. Plos one. 2021; 16(1):e0244636. https://doi.org/10.1371/journal.pone.0244636 PMID: 33465075

42. Wu X, Zhu X. Mining with noise knowledge: error-aware data mining. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2008; 38(4):917–32. https://doi.org/10.1109/CIS.2007.7.

43. Abbasi R, Balazs P, Noll A, Nicolakis D, Marconi MA, Zala SM, et al. Applying convolutional neural networks to the analysis of mouse ultrasonic vocalizations2019: Universitätsbibliothek der RWTH Aachen. https://doi.org/10.18154/RWTH-CONV-239263.

44. Huang G, Li Y, Pleiss G, Liu Z, Hopcroft JE, Weinberger KQ. Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:00109. 2017.

45. Zhou Z-H. Ensemble methods: foundations and algorithms: CRC press; 2012.

46. Hanson JL, Hurley LM. Female presence and estrous state influence mouse ultrasonic courtship vocalizations. PloS one. 2012; 7(7):e40782. https://doi.org/10.1371/journal.pone.0040782 PMID: 22815817

47. Macmillan NA, Creelman CD. Detection theory: A user's guide. 2 ed: Psychology press; 2004. https://doi.org/10.4324/9781410611147.

48. De Boer E, De Jongh H. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. The Journal of the Acoustical Society of America. 1978; 63(1):115–35. https://doi.org/10.1121/1.381704 PMID: 632404

49. Balazs P, Holighaus N, Necciari T, Stoeva D. Frame theory for signal processing in psychoacoustics. Excursions in Harmonic Analysis, Volume 5: Springer; 2017. p. 225–68. https://doi.org/10.1007/978-3-319-54711-4_10.

50. Slaney M. Auditory toolbox. Interval Research Corporation, Tech Rep. 1998; 10(1998):1194. https://doi.org/10.1007/978-3-642-37762-4_2.

51. Chen C-P, Bilmes JA, Kirchhoff K. Low-resource noise-robust feature post-processing on Aurora 2.0. Seventh International Conference on Spoken Language Processing; 2002.

52. Huang T, Yang G, Tang G. A fast two-dimensional median filtering algorithm. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979; 27(1):13–8. https://doi.org/10.1109/TASSP.1979.1163188.

53. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning: MIT press Cambridge; 2016.

54. Fukushima K. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980; 36:193–202. https://doi.org/10.1007/BF00344251 PMID: 7370364

55. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–324. https://doi.org/10.1016/10.1109/5.726791.

56. Murphy KP. Machine learning: a probabilistic perspective: MIT press; 2012.

57. Chen Y, Jiang H, Li C, Jia X, Ghamisi P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing. 2016; 54(10):6232–51. https://doi.org/10.1109/TGRS.2016.2584107.

58. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634.

59. Nesterov Y. A method for unconstrained convex minimization problem with the rate of convergence O (1/k^ 2). Doklady an ussr; 1983.

60. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision; 2015. https://doi.org/10.1109/ICCV.2015.123.

61. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, et al. Improving the generalizability of convolutional neural network-based segmentation on CMR images. Frontiers in cardiovascular medicine. 2020; 7:105. https://doi.org/10.3389/fcvm.2020.00105 PMID: 32714943

62. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014; 15(1):1929–58.

63. Yan Y, Chen M, Shyu M-L, Chen S-C. Deep learning for imbalanced multimedia data classification. IEEE international symposium on multimedia (ISM); 2015: IEEE.

64. Anguita D, Boni A, Ridella S. Evaluating the generalization ability of support vector machines through the bootstrap. Neural Processing Letters. 2000; 11(1):51–8. https://doi.org/10.1023/A:1009636300083.

65. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter. 2004; 6(1):20–9. https://doi.org/10.1145/1007730.1007735.

66. King G, Zeng L. Logistic regression in rare events data. Political analysis. 2001; 9(2):137–63.

67. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. https://doi.org/10.1145/2939672.2939785.

68. Sammut C, Webb GI. Encyclopedia of machine learning: Springer Science & Business Media; 2011. https://doi.org/10.1007/978-0-387-30164-8.

69. Wang H, Liang S, Burgdorf J, Wess J, Yeomans J. Ultrasonic vocalizations induced by sex and amphetamine in M2, M4, M5 muscarinic and D2 dopamine receptor knockout mice. PloS One. 2008; 3(4): e1893. https://doi.org/10.1371/journal.pone.0001893 PMID: 18382674

70. Lvd Maaten, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9 (Nov):2579–605.

71. Fonseca AH, Santana GM, Ortiz GMB, Bampi S, Dietrich MO. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. Elife. 2021; 10:e59161. https://doi.org/10.7554/eLife.59161 PMID: 33787490

72. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.

73. de Hoz L, Nelken I. Frequency tuning in the behaving mouse: different bandwidths for discrimination and generalization. PloS one. 2014; 9(3):e91676. https://doi.org/10.1371/journal.pone.0091676 PMID: 24632841

74. Neilans EG, Holfoth DP, Radziwon KE, Portfors CV, Dent ML. Discrimination of ultrasonic vocalizations by CBA/CaJ mice (Mus musculus) is related to spectrotemporal dissimilarity of vocalizations. PloS one. 2014; 9(1):e85405. https://doi.org/10.1371/journal.pone.0085405 PMID: 24416405

75. Screven LA, Dent ML. Perception of ultrasonic vocalizations by socially housed and isolated mice. Eneuro. 2019; 6(5). https://doi.org/10.1523/ENEURO.0049-19.2019 PMID: 31570420

76. Screven LA, Dent ML. Discrimination of frequency modulated sweeps by mice. The Journal of the Acoustical Society of America. 2016; 140(3):1481–7. https://doi.org/10.1121/1.4962223 PMID: 27914389

6. DISCUSSION AND CONCLUSIONS

The chapters in my dissertation mainly addressed hypotheses about the modulation of olfactory and acoustic signals and their signaling functions in wild house mice (*Mus musculus musculus*) and laboratory mice (*Mus laboratorius*), and here I summarize the main findings.

The first chapter shows that the excretion of volatile chemical compounds (VOCs) and pheromone-binding MUP proteins depend upon sex and social status. In seminatural conditions, the mice upregulated the concentration of urinary proteins in both sexes when compared to controls housed in standard conditions. The upregulation of urinary proteins potentially provides a reliable signal of social status and dominant males and females both had a higher reproductive success compared to subordinate individuals. The upregulation of MUPs in males, though not females, was correlated with their reproductive success. The higher expression of urinary proteins in dominant females in seminatural conditions might have been triggered by the presence of males, increasing the females' expression of MUPs close to the estrus (1) and the frequency of oestrus cycles over time (2-4). We also found temporal fluctuations in MUP expression in females, possibly due to seasonality. Our results showed that females produced the highest concentration of urinary creatinine in spring when most were pregnant (5). As in previous studies on other mice, males released expected higher concentrations of certain pheromones compared to females (6, 7). More detailed analyses of these volatile compounds showed that two were regulated according to male social status: 1) 6-hydroxy-6-methyl-3-heptanone (HMH), a volatile compound that influences the sexual development and behavior of females, and was positively correlated with the dominance status and reproductive success of males; 2) and trimethylamine (TMA), which is a bacterial metabolite, that was upregulated in subordinate males and negatively correlated with their reproductive success. TMA is known

to be used for species and sex recognition and mice show attraction at low concentrations but aversion at high concentrations (8). Thus, high TMA excretion could help explain how females recognize and assess subordinate males. In summary, the concentration of different MUPs and VOCs in males might signal to conspecifics of both sexes their social status, the territory boundaries and potentially avoiding direct aggressive behaviors with neighboring males, whereas proteins regulation in females' urine might signal receptivity and attract males (1, 2, 9). This is the first study to correlate the expression and regulation of pheromones with male social status and reproductive success, and the first to our knowledge in any mammal.

In the second chapter, we showed that exposing wild male mice to a female (socio-sexual priming) influenced their subsequent courtship USVs, including USV number, spectral features, and syllable type diversity compared to controls, and that the effects of the socio-sexual priming lasted for almost one month. We confirmed that the implemented version of the USV detection system A-MUD 3.2 provides a highly efficient method for detecting USVs compared to our previous version (10). Our results on sexual priming confirmed that the number of both, sonic (i.e., squeaks and vocalizations under 20 kHz) and ultrasonic vocalizations increased after priming and were also highly correlated. Primed males emitted three times more USVs than the unprimed mice. USVs in primed males were longer in duration and their spectral diversity increased 1 day and 20 days after sexual priming. We found a high interindividual variation in the vocal repertoire and spectrotemporal features as previously found in wild mice (11). Unprimed males had the most distinct and spectrally simple vocal repertoire with more USVs of short duration and non-distinct shape compared to the other primed groups. USVs from unprimed male showed a higher interindividual variation in the mean frequency, a shorter duration and a lower slope (i.e., frequency evolution over time of the vocalization). The low

variability found in the mean frequency of USVs emitted by primed males might suggest a possible female's preference on "attractive" acoustic stimuli with specific spectrotemporal features that might potentially induce female's receptivity and increase mating opportunities (12-14). Previous studies showed that sexual priming affects the regulation of male's sexual behavior enhancing the scent-marking behavior (15), reducing the latency to copulation and increasing the sperm density (16, 17). Our study provides the first evidence that socio-sexual priming increases males' vocalizations by enhancing the number and spectral complexity of USVs with long-lasting effects. However, the signaling functions and functional consequences of USV modulation after priming are still not known, yet these results are consistent with the hypothesis that USVs reliably signal sexual arousal and interest in mating (reviewed in 18, 19).

In the third chapter, our study provides three main results: 1) similar to the priming effect, all males increased the number, length, and spectral complexity of the USVs once the partition was removed and the social partners were allowed to directly interact; 2) unrelated pairs also increased the length and spectral complexity of USVs during direct interactions, differently to related pairs where males increased the length of USVs when physically separated by a perforated divider from the females; and 3) unrelated males sired a higher number of litters and females had a shorter latency to give birth compared to related breeding pairs, suggesting that USV emission can act as an important factor of inbreeding avoidance (20). Similar to laboratory mice, we found that wild males modulate the USV emission depending upon social contexts (before and during direct contact) (as reviewed in 21, 22) and different types of stimuli (14, 20). Our results suggest that unrelated pairs discriminate between kin and non-kin and that males modulate the USV emission accordingly when paired with an unrelated female. These

findings suggest that USV signals might contain kinship vocal signatures and potentially information on the identity of the partners (23).

In the fourth chapter, the results show that adult males can alter the USV emission upon female odor stimulation. Detailed analyses of courtship USVs show individually distinct features in the vocalizations that are consistent over time and might be critical for individual recognition. The alteration in USV features consisted in a higher USV number and duration, more spectral complexity, and a modification of the spectrographic features. Males showed a distinct "vocalizing style" given by strong individual differences in the number of USVs emitted that was consistent over time. Thus, males could be grouped into different groups according to their vocalizing style: low, intermediate, and high vocalizers, and consistent and non-consistent vocalizers. Differently from previous findings in wild mice, non-vocalizing was a rare consistent trait (23, 24). Different statistical approaches and the application of the random forest on the three recordings confirmed that some USV categories and spectrotemporal features were good candidates as individual vocal signatures. Although we did not reach a full consensus using different methods to detect the IVS, it was clear that the repertoire composition played an important role in individual discrimination. Machine learning methods confirmed that the frequency parameters, some syllable types, the USV duration and the interval between syllables were good candidates as IVS. Thus, mouse USVs contain individual vocal signatures that might be used for individual recognition, but these IVS need to be tested over a longer time for stability and in other social contexts for robustness, as shown in other species (25, 26). However, there are still important questions to address: 1) whether USV emission can be used for individual recognition and health status (27); 2) whether three consecutive weeks is a reliable time to detect individual vocal signatures (28, 29), as mice live in social groups and

undergo different social experiences over three weeks, such as inter- and intrasexual interactions (i.e., estrous cycles, territory formation by males, breeding)(30, 31) and parent-offspring interactions (i.e., rearing and weaning of new generation) (reviewed in 32, 33); 3) whether differences in the morphology of the larynx (as reviewed in 34) and in the physiology through endocrine control might favor individual recognition (33, 35) ; 4) and whether USV emission might be a heritable trait (36). This is the first study to show that female odor enhances the number and complexity of USVs over time with similar characteristics of courtship USVs in presence of a female. Features in the USVs are individually distinct and consistent over time, suggesting that courtship USVs might be a reliable signal for individual recognition. However, playback studies are needed to test whether females can identify from specific IVS candidates the identity of the vocalizing males, and thus confirm our findings.

In the fifth chapter, we showed that the three mouse strains differed from each other since the first neonatal stage. B6 pups and adults emitted a lower number of USVs with a shorter duration, and USVs were spectrally simple compared to the other strains. The outbred CD-1 pups vocalized more than the inbred FVB and B6 strains. As a developmental milestone, we reported that B6 pups showed a peak of the calling rate earlier than the other two strains (peak at post-natal day 4 compared to peak at post-natal day 6-8). Pups used a larger vocal repertoire that decreased during adulthood in all three strains (i.e., from 6 or more types as neonates to 4 categories when adults). During social interaction tests, inbred FVB strain performed more social investigations and the highest number of USVs as both neonates and adults. The repertoire composition in FVB mice included more spectrally complex USVs or with harmonic elements. The outbred CD1 strain showed an intermediate socio-communicative profile for both the number of USVs emitted and the vocal repertoire with a similar acoustic

profile during the neonatal phase to the B6, but with an increase in frequency step and complex USVs more similar to the FVB as adults. From visual inspection, each strain showed a vocal repertoire that varied according to the sex of the vocalizing individuals and their social partners, similar to findings in wild mice (11). Since the B6 adult males and females performed the lowest number of social interactions and USV emission, researchers should think carefully about whether using B6 as a background strain for mutant lines, or as reference strain in tests measuring social responsiveness (37, 38). The intermediate behavioral and communicative profile of the outbred CD1 suggests that this strain can be included in behavioral and acoustic tests with inbred strains, but raise questions about the real differences between inbred and outbred strains (39). For the first time, we provide evidence for a strain difference at different developmental stages and we confirmed a correlation between the occurrence of USV emission and the investigatory behavior, influenced by the sex and strain during social interactions. Our results suggest that each strain might have its characteristics that should be taken into account when choosing a strain for a specific research question, or creating a genetically modified line from a background strain (reviewed in 40). More studies are needed to confirm whether the USV modulation during development might get stabilized and allow strain recognition between different strains (41, 42).

In the sixth chapter, we showed the excellent performance of A-MUD 3.2 in detecting USVs in presence of background noise, and the flexibility and accuracy of Bootsnap in classifying USVs from different sources (laboratory strains, wild mice and different contexts). Both programs are free for scientific use and Bootsnap is based on Python, which is a free platform. When comparing different USV detection methods, we found that A-MUD 3.2 performed better than the other models trained by using only USVs of laboratory mice. Our

classification method “Bootsnap” outperformed the best available classifier (i.e., Deepsqueak), which was pretrained on laboratory mice recordings and also retrained with wild mice recordings. Our study showed some limitation in correctly classifying the USVs that are common to other software: 1) the presence of background noise during the recordings; 2) the similarity between the spectral shapes of some USV types, that caused problems also to trained researchers; 3) and the low amplitude at the start or end of some vocalizations that were difficult to include in the correct USV category. For the first time, we compared the performance of Bootsnap with manually classified soundfiles recorded from both, wild and laboratory mice. We reported a good inter-rater reliability that was higher for more spectrally distinct USV types, but lower when comparing USVs showing classification problems shared with the automatic classifiers. Bootsnap reached the highest accuracy when it used only two classes (i.e., false positives including segments with high background noise and USVs), but its performance degraded when the number of categories increased to more than six. A good trade-off between the number of USV categories and clustering accuracy was obtained for six different USV classes including the false positives as a separate category (i.e., background noise). Bootsnap applied the bootstrapping method to increase the number of elements in all categories to correctly classify less represented USV categories. This study presents an implemented version of the USV detection system that we previously developed and that outperforms the other methods even in low signal-to-noise ratio recordings. We coupled the USV detection system with a new state-of-the-art USV classifier whose performance is higher than other methods and has been compared with manual classification. These methods prove to be a key support for the analyses of USVs as non-invasive tools for behavioral tests and animal welfare in rodents (43-45).

In summary, our overall findings provide novel evidence that chemical and acoustic communication convey information on the sex, social and physiological status, age, kinship, and identity of the signaling mouse in standard and seminatural conditions. During courtship, USV modulation is strongly influenced by social contexts (i.e., presence of the female or her scent), and this change can be long-lasting after the socio-sexual priming and can carry information on the sexual experience of males. USVs can signal the genetic relatedness of the emitting individuals and might affect their reproductive success. USVs can also signal individual identity through individual vocal signatures. Some IVS candidates are consistent over three weeks and show interindividual differences. However, more playback studies are needed to confirm whether IVS candidates could be used by mice for individual recognition over a longer period of time.

In conclusion, we showed that studies in wild mice that carry a high genetic diversity and underwent natural and sexual selection, are of critical importance to investigate some functions related to the individual's fitness through chemical and acoustic signals. Determining the adaptive functions of USVs can provide insights into conducting and interpreting the results of behavioral assays that are mostly performed in the laboratory with *Mus laboratorius*, which is now represented by thousands of artificially selected strains, whose behavior is highly heterogeneous (46). Therefore, we should consider the translational value of the studies on the wild house mouse in standard and seminatural conditions.

References

1. Černá M, Kuntová B, Talacko P, Stopková R, Stopka P. Differential regulation of vaginal lipocalins (OBP, MUP) during the estrous cycle of the house mouse. Scientific Reports. 2017;7(1):11674. doi: 10.1038/s41598-017-12021-2.
2. Whitten WK. Modification of the oestrus cycle of the mouse by external stimuli associated with the male Journal of Endocrinology. 1958;17(3):307-13. doi: 10.1677/joe.0.0170307.
3. Jemiolo B, Alberts J, Sochinski-Wiggins S, Harvey S, Novotny M. Behavioural and endocrine responses of female mice to synthetic analogues of volatile compounds in male urine. Animal Behaviour. 1985;33(4):1114-8. doi: https://doi.org/10.1016/S0003-3472(85)80170-6.
4. Marchlewska-Koj A, Cavaggioni A, Mucignat-Caretta C, Olejniczak P. Stimulation of estrus in female mice by male urinary proteins. Journal of Chemical Ecology. 2000;26(10):2355-66. doi: 10.1023/A:1005578911652.
5. König B, Lindholm AK. The complex social environment of female house mice (Mus domesticus). In: Piálek J, Macholán M, Munclinger P, Baird SJE, editors. Evolution of the House Mouse. Cambridge Studies in Morphology and Molecules: New Paradigms in Evolutionary Bio. Cambridge: Cambridge University Press; 2012. p. 114-34.
6. Stopková R, Stopka P, Janotová K, Jedelský PL. Species-specific expression of major urinary proteins in the house mice (Mus musculus musculus and Mus musculus domesticus). Journal of Chemical Ecology. 2007;33(4):861-9. doi: 10.1007/s10886-007-9262-9.
7. Sampsell BM, Held WA. Variation in the major urinary protein multigene family in wild-derived mice. Genetics. 1985;109(3):549-68. Epub 1985/03/01. doi: 10.1093/genetics/109.3.549. PubMed PMID: 3979814; PubMed Central PMCID: PMCPMC1216287.
8. Li Q, Korzan WJ, Ferrero DM, Chang RB, Roy DS, Buchi M, et al. Synchronous evolution of an odor biosynthesis pathway and behavioral response. Curr Biol. 2013;23(1):11-20. Epub 2012/11/28. doi: 10.1016/j.cub.2012.10.047. PubMed PMID: 23177478; PubMed Central PMCID: PMCPMC3543494.
9. Ronald KL, Zhang X, Morrison MV, Miller R, Hurley LM. Male mice adjust courtship behavior in response to female multimodal signals. PLOS ONE. 2020;15(4):e0229302. doi: 10.1371/journal.pone.0229302.
10. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Automatic mouse ultrasound detector (A-MUD): a new tool for processing rodent vocalizations. PLoS One. 2017;12(7):e0181200. Epub 2017/07/21. doi: 10.1371/journal.pone.0181200. PubMed PMID: 28727808; PubMed Central PMCID: PMCPMC5519055.
11. Zala SM, Reitschmidt D, Noll A, Balazs P, Penn DJ. Sex-dependent modulation of ultrasonic vocalizations in house mice (Mus musculus musculus). PLoS One. 2017;12(12):e0188647. Epub 2017/12/14. doi: 10.1371/journal.pone.0188647. PubMed PMID: 29236704; PubMed Central PMCID: PMCPMC5728457.
12. White NR, Prasad M, Barfield RJ, Nyby JG. 40- and 70-kHz vocalizations of mice (Mus musculus) during copulation. Physiology & Behavior. 1998;63(4):467-73. doi: https://doi.org/10.1016/S0031-9384(97)00484-8.
13. Matsumoto YK, Okanoya K. Phase-specific vocalizations of male mice at the initial encounter during the courtship sequence. PLOS ONE. 2016;11(2):e0147102. doi: 10.1371/journal.pone.0147102.

14. Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Frontiers in Behavioral Neuroscience. 2015;9. doi: 10.3389/fnbeh.2015.00076.

15. Zala SM, Potts WK, Penn DJ. Scent-marking displays provide honest signals of health and infection. Behavioral Ecology. 2004;15(2):338-44. doi: 10.1093/beheco/arh022.

16. Koyama S, Kamimura S. Influence of social dominance and female odor on the sperm activity of male mice. Physiology & Behavior. 2000;71(3):415-22. doi: https://doi.org/10.1016/S0031-9384(00)00361-9.

17. James PJ, Nyby JG. Testosterone rapidly affects the expression of copulatory behavior in house mice (Mus musculus). Physiology & Behavior. 2002;75(3):287-94. doi: https://doi.org/10.1016/S0031-9384(01)00666-7.

18. Pfaus JG, Kippin TE, Centeno S. Conditioning and sexual behavior: a review. Hormones and Behavior. 2001;40(2):291-321. doi: https://doi.org/10.1006/hbeh.2001.1686.

19. Fernández-Vargas M. Vocal signals of sexual motivation in male and female rodents. Current Sexual Health Reports. 2018;10(4):315-28. doi: 10.1007/s11930-018-0179-9.

20. Musolf K, Hoffmann F, Penn DJ. Ultrasonic courtship vocalizations in wild house mice, Mus musculus musculus. Animal Behaviour. 2010;79(3):757-64. doi: https://doi.org/10.1016/j.anbehav.2009.12.034.

21. Ehret G. Chapter 18 - Characteristics of vocalization in adult mice. In: Brudzynski SM, editor. Handbook of Behavioral Neuroscience. 25: Elsevier; 2018. p. 187-95.

22. Premoli M, Memo M, Bonini SA. Ultrasonic vocalizations in mice: relevance for ethologic and neurodevelopmental disorders studies. Neural Regen Res. 2021;16(6):1158-67. Epub 2020/12/04. doi: 10.4103/1673-5374.300340. PubMed PMID: 33269765; PubMed Central PMCID: PMCPMC8224126.

23. Hoffmann F, Musolf K, Penn DJ. Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. Physiol Behav. 2012;105(3):766-71. Epub 2011/11/01. doi: 10.1016/j.physbeh.2011.10.011. PubMed PMID: 22037196.

24. von Merten S, Hoier S, Pfeifle C, Tautz D. A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (Mus musculus domesticus). PLOS ONE. 2014;9(5):e97244. doi: 10.1371/journal.pone.0097244.

25. Linhart P, Osiejuk TS, Budka M, Šálek M, Špinka M, Policht R, et al. Measuring individual identity information in animal signals: overview and performance of available identity metrics. Methods in Ecology and Evolution. 2019;10(9):1558-70.

26. Linhart P, Mahamoud-Issa M, Stowell D, Blumstein DT. The potential for acoustic individual identification in mammals. Mammalian Biology. 2022. doi: 10.1007/s42991-021-00222-2.

27. Lopes PC, König B. Choosing a healthy mate: sexually attractive traits as reliable indicators of current disease status in house mice. Animal Behaviour. 2016;111:119-26. doi: https://doi.org/10.1016/j.anbehav.2015.10.011.

28. Holy TE, Guo Z. Ultrasonic songs of male mice. PLoS Biol. 2005;3(12):e386. Epub 2005/10/27. doi: 10.1371/journal.pbio.0030386. PubMed PMID: 16248680; PubMed Central PMCID: PMCPMC1275525.

29. Melotti L, Siestrup S, Peng M, Vitali V, Dowling D, von Kortzfleisch VT, et al. Individuality, as well as genetic background, affects syntactical features of courtship songs in male mice. Animal Behaviour. 2021;180:179-96. doi: https://doi.org/10.1016/j.anbehav.2021.08.003.

30. Hoier S, Pfeifle C, von Merten S, Linnenbrink M. Communication at the garden fence - context dependent vocalization in female house mice. PloS one. 2016;11(3):e0152255-e. doi: 10.1371/journal.pone.0152255. PubMed PMID: 27022749.

31. Varholick JA, Bailoo JD, Palme R, Würbel H. Phenotypic variability between social dominance ranks in laboratory mice. Scientific Reports. 2018;8(1):6593. doi: 10.1038/s41598-018-24624-4.

32. Musolf K, Penn DJ. Ultrasonic vocalizations in house mice: a cryptic mode of acoustic communication. In: Macholan M, Baird SJE, Munclinger P, Pialek J, editors. Evolution of the house mouse: Cambridge University Press; 2012. p. 253-77.

33. Asaba A, Hattori T, Mogi K, Kikusui T. Sexual attractiveness of male chemicals and vocalizations in mice. Frontiers in Neuroscience. 2014;8. doi: 10.3389/fnins.2014.00231.

34. Fernández-Vargas M, Riede T, Pasch B. Mechanisms and constraints underlying acoustic variation in rodents. Animal Behaviour. 2022;184:135-47. doi: https://doi.org/10.1016/j.anbehav.2021.07.011.

35. Mills EGA, Dhillo WS, Comninos AN. Kisspeptin and the control of emotions, mood and reproductive behaviour. Journal of Endocrinology. 2018;239(1):R1-R12. doi: 10.1530/joe-18-0269.

36. Spence HR, Aslam AM, Hofer MA, Brunelli SA, Shair HN. Vocal coselection in rat pup ultrasonic vocalizations. Ecology and Evolution. 2016;6(7):1922-9. doi: https://doi.org/10.1002/ece3.1907.

37. Crawley JN. Behavioral phenotyping of transgenic and knockout mice: experimental design and evaluation of general health, sensory functions, motor abilities, and specific behavioral tests. Brain Res. 1999;835(1):18-26. Epub 1999/08/17. doi: 10.1016/s0006-8993(98)01258-x. PubMed PMID: 10448192.

38. Moy SS, Nadler JJ. Advances in behavioral genetics: mouse models of autism. Mol Psychiatry. 2008;13(1):4-26. Epub 2007/09/13. doi: 10.1038/sj.mp.4002082. PubMed PMID: 17848915.

39. Tuttle AH, Philip VM, Chesler EJ, Mogil JS. Comparing phenotypic variation between inbred and outbred mice. Nat Methods. 2018;15(12):994-6. Epub 2018/12/07. doi: 10.1038/s41592-018-0224-7. PubMed PMID: 30504873; PubMed Central PMCID: PMCPMC6518396.

40. Heckman J, McGuinness B, Celikel T, Englitz B. Determinants of the mouse ultrasonic vocal structure and repertoire. Neurosci Biobehav Rev. 2016;65:313-25. Epub 2016/04/10. doi: 10.1016/j.neubiorev.2016.03.029. PubMed PMID: 27060755.

41. Sugimoto H, Okabe S, Kato M, Koshida N, Shiroishi T, Mogi K, et al. A role for strain differences in waveforms of ultrasonic vocalizations during male–female interaction. PLOS ONE. 2011;6(7):e22093. doi: 10.1371/journal.pone.0022093.

42. Nomoto K, Hashiguchi A, Asaba A, Osakada T, Kato M, Koshida N, et al. Female C57BL/6 and BALB/c mice differently use the acoustic features of male ultrasonic vocalizations for social preferences. Exp Anim. 2020;69(3):319-25. Epub 2020/02/27. doi: 10.1538/expanim.19-0119. PubMed PMID: 32101835; PubMed Central PMCID: PMCPMC7445051.

43. Grimsley JMS, Sheth S, Vallabh N, Grimsley CA, Bhattal J, Latsko M, et al. Contextual modulation of vocal behavior in mouse: newly identified 12 kHz “mid-frequency” vocalization emitted during restraint. Frontiers in Behavioral Neuroscience. 2016;10. doi: 10.3389/fnbeh.2016.00038.

44. Smith BJ, Bruner KEP, Kendall LV. Female- and intruder-induced ultrasonic vocalizations in C57BL/6J mice as proxy indicators for animal wellbeing. Comp Med. 2019;69(5):374-83. Epub

2019/10/04. doi: 10.30802/aalas-cm-18-000147. PubMed PMID: 31578163; PubMed Central PMCID: PMCPMC6807721.

45. Bigelow LJ, Cohen AJ, Pimm R, Knight JB, Bernard PB. Ultrasonic vocalization analysis as a novel metric to assess cage enrichment in rats. J Am Assoc Lab Anim Sci. 2022;61(2):140-8. Epub 2022/01/28. doi: 10.30802/aalas-jaalas-21-000024. PubMed PMID: 35082005; PubMed Central PMCID: PMCPMC8956216.

46. Eppig JT, Motenko H, Richardson JE, Richards-Smith B, Smith CL. The International Mouse Strain Resource (IMSR): cataloging worldwide mouse and ES cell line resources. Mamm Genome. 2015;26(9-10):448-55. Epub 2015/09/17. doi: 10.1007/s00335-015-9600-0. PubMed PMID: 26373861; PubMed Central PMCID: PMCPMC4602064.