Department for Biomedical Sciences

University of Veterinary Medicine Vienna

Institute of Population Genetics

(Head: Univ. -Prof. Dr. rer. nat. Christian Schlötterer)

Examination of the regulatory activity of a selected haplotype-block via an enhancer-reporter assay

Master thesis submitted for the fulfilment of the requirements for the degree of

Master of Science (MSc)

submitted by

Lucas Schröfl

Vienna, 14.02.2020

Supervisor: Senti Kirsten-André, Dr. rer. nat.

Examiner: Wallner Barbara, Dr^{in}. vet. med.

Acknowledgment

"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

Albert Einstein

Table of Content

Introduction

1. Evolutionary forces and the process of adaptation

1.1 Evolutionary mechanisms

The process of evolution is shaped by many independent processes. Five major mechanisms dictate the fate of evolution predominantly. These are mutation, recombination, gene flow, genetic drift and selection (Barton *et al.* 2007).

Selection is the increased propagation of entities (e.g. populations, individuals, genes or alleles) as a consequence of some inherited traits carried by these entities. As a result, the number of these beneficial traits will then tend to increase within a discrete group (e.g. populations within an ecosystem, individuals within a population) (Barton *et al.* 2007). The entirety of all traits and features of an entity that are relevant for its propagation are summarized in its fitness value. Fitness is defined as the number of offspring (or copies of an allele) produced in the next generation and is made up of several separate fitness components (e.g. survival, lifespan, reproduction rate) that add up to give the overall fitness (Barton *et al.* 2007). The proportion of genotypes present in a population depends on the ratio of every individual genotypes fitness to the mean genotype fitness in the population, in discrete time. This ratio is termed relative fitness. The absolute fitness of a population portrays if the population as a whole is shrinking or expanding. The separation of absolute fitness and relative fitness corresponds to the separation of ecology and population genetics (Barton *et al.* 2007). Otherwise speaking, selection arises from and is the inevitable consequence of the systematic accumulation of inherited variation in reproductive success (Barton *et al.* 2007).

The other four of these five major evolutionary mechanisms, namely mutation, recombination, gene flow, and genetic drift, act at random concerning function and adaptation. They act to generate, maintain and shape genetic variation within and between populations (Barton *et al.* 2007). Those random processes will be defined in the aspect of generation and maintenance of genetic variation in a highly simplistic manner in the

following. Mutation creates genomic polymorphisms. Recombination shuffles these polymorphisms into a random combination of variants, thereby indirectly increasing the rate of adaptation. Gene flow allows the spread of genetic variants between discrete populations. Genetic drift facilitates the random propagation and loss of polymorphisms, thereby decreasing genetic diversity within populations and increasing genetic divergence between populations, in a population size-dependent manner (Barton *et al.* 2007, Kawecki *et al.* 2012, Long *et al.* 2015, Jonas *et al.* 2016, Elena & Lenski 2003).

These four fundamental processes govern the course of evolution, although they prevalently degrade the products of evolution, and selection is the only deterministic mechanism capable of generating functional, adapted and complex organisms (Barton *et al.* 2007).

1.2 Experimental Evolution

The idiosyncrasy of evolutionary principles and forces can be studied in laboratory populations as a consequence of conditions imposed by the experimenter, over several generations. The response to these conditions can be measured based on changes in phenotype, or in genome sequence. This experimental framework is called experimental evolution (EE) and allows surveillance of evolutionary processes in real-time, as well as the possibility to replicate experiments under identical conditions. This approach facilitates the identification of deterministic and stochastic evolutionary processes (Kawecki *et al.* 2012, Schlötterer *et al.* 2015). The type of selection regime (e.g. environmental, genetic, social, demographic) and the way the regime can be applied to the population, is only limited by the imagination of the scientist. Presumptions of relevant traits (in a given selection regime) often motivate EE studies, however, selection will act on any and all traits relevant to fitness, making straightforward descriptive studies possible (Kawecki *et al.* 2012).

The first experiment within the framework of experimental evolution was performed already in 1881 by Louis Pasteur, long before the study of evolutionary principles came into the grasp of science. He employed experimental evolution to produce a live attenuated vaccine for *Pasteurella multocida*, the causative agent of chicken cholera, by repeated passaging of the

bacteria onto an unnatural host, until the passaged strain was deprived in pathogenicity and virulence. Though the theoretical basis of Darwinian evolution involved in the vaccine production was obscure, other live attenuated vaccines against diseases such as typhoid, plague, and cholera followed soon (Kawecki *et al.* 2012, Harper *et al.* 2006, Plotkin & Plotkin 2011).

1.3 Evolve and Resequence

In the last half of a century, adaptive responses in EE studies were measured primarily by the identification of phenotypic changes, later combined with the description of genomic changes displayed by a limited number of genetic markers (allozymes, single nucleotide polymorphisms, microsatellites, insertion sequence elements) (Orozco-ter Wengel *et al.* 2012, Signor & Nuzhdin 2018, Teotonio *et al.* 2009, Haley & Birley 1983, De Jong & Bochdanovits 2003, Greer 2017, Schneider & Lenksi 2004), rather than on whole genome data, due to complexity, high expenses and low-throughput inherent to Sanger sequencing (Long *et al.* 2015, Schlötterer *et al.* 2015, Orozco-ter Wengel *et al.* 2012, Reuter *et al.* 2015). Because these genetic markers only covered a vanishingly small fraction of the genome, they illustrated a rather incomplete picture of the genetic signature of adaptation (Nuzhdin *et al.* 1993, Teotonio *et al.* 2009, Haley & Birley 1983, Dunham *et al.* 2002, Papadopoulos *et al.* 1999, Schneider & Lenski 2004). Nevertheless, EE studies never failed to identify at least one marker that showed a pattern of non-neutral evolution, suggesting that large fractions of the genome respond to selection (Nuzhdin *et al.* 1993, Teotonio *et al.* 2009, Haley & Birley 1983, Dunham *et al.* 2002, Papadopoulos *et al.* 1999, Schneider & Lenski 2004).

The recent advent of Next-Generation Sequencing (NGS) and the continuous decrease in DNA whole genome sequencing costs enabled the pursuance of the ultimate goal of EE: a complete description of the genetic signature which establishes adaptive phenotypes (Schlötterer *et al.* 2015). This combination of EE and high-throughput whole genome sequencing is referred to as Evolve and Resequence (ER, coined by Turner *et al.* 2011), and unifies the branches of molecular and population genetics (Kofler & Schlötterer 2014). Whole genome sequencing of pools of individuals (pool-seq) is the method of choice in E&R studies because it offers accurate whole-genome allele frequency estimates at lower costs than whole

genome sequencing of individuals (Schlötterer *et al.* 2014). E&R studies permit to description of frequencies of genome-wide polymorphisms that segregate within a population as a function of treatment and time (number of generations) (Long *et al.* 2015, Schlötterer *et al.* 2015). Further statistical analysis and the opportunity to contrast of these frequency trajectories across biological replicates allows the identification of genomic loci that are targeted by evolutionary processes (e.g. Kawecki *et al.* 2012, Elena & Lenksi 2003, Schlötterer *et al.* 2015, Barghi *et al.* 2017, Kofler & Schlötterer 2014, Orozco-ter Wengel *et al.* 2012).

This appealing experimental approach is applied to a wide variety of biological systems, ranging from the synthetic evolution of libraries of oligonucleotide *in vitro*, to isogenic asexual bacterial, or semi-sexual yeast populations, to obligate sexual multicellular eukaryotes (Long *et al.* 2015). Numerous different experimental settings are possible in E&R studies, but two approaches are most common. The first one is the investigation of evolution in asexual microorganisms with an isogenic starting population (i.e. base population), where accumulating mutations are the source of genetic variation which fuel adaptation. This evolutionary system is characterized by hard and soft selective sweeps, that accompany beneficial alleles on their way to fixation, and clonal interference caused by the competition of diverse beneficial alleles that are present in different clones (Long *et al.* 2015, Hermisson & Pennings 2005, Elena & Lenski 2003). In this evolutionary system, adaptation is constrained by the rate at which mutations occur and the distribution of their effect size (Hermisson & Pennings 2005, Kawecki *et al.* 2012).

In the second experimental setting, the base populations are initiated from genetically polymorphic obligate sexual metazoans (i.e. harboring substantial standing genetic variation). In this setting, the fate of adaptation highly depends on the quality and amount of genetic variation present in the base population, and additionally on the history of selective pressures which the population had encountered in the past (Hermisson & Pennings 2005). The targets of selection are identified by the contrast of observed allele frequency changes with population genetics theory. A typical genetic signature of adaptation from standing genetic variation is allele frequency change at many loci, suggesting polygenic adaptation. This results in a large number of putative selected candidate single nucleotide polymorphisms

(SNPs) and large fractions of the genome that seem to be selected (Hermisson & Pennings 2005, Barghi *et al* 2017, Orozco-ter Wengel *et al.* 2012, Teotónio *et al.* 2009, Kofler & Schlötterer 2014).

One preferred model organism for E&R studies in metazoans is *Drosophila melanogaster*, an organism that offers the advantages of a small and well-annotated genome, short generation time, resource- and space-savings, high levels of polymorphisms and a superlative amount of technical and genetic resources (Mohr *et al.* 2014, Matthews *et al.* 2005, Barghi *et al.* 2017). In *D. melanogaster* a variety of E&R studies have been conducted, applying diverse selection regimes (Turner *et al.* 2011, Orozco-terWengel *et al.* 2012, Kolss *et al.* 2009, Santos *et al.* 1997, Zhou *et al.* 2011, Turner & Miller 2012, Remolina *et al.* 2012), and large fractions of the genome that respond to selection and harbor large numbers of presumably selected SNPs, have been identified (Tobler *et al.* 2014).

Many of these putative selected SNPs are prevalently false positives (Barghi *et al* 2017, Tobler *et al.* 2014), and they only appear to be selected because they are closely linked to a truly beneficial allele located in *cis*. As a consequence, those false positive candidate SNPs exhibit a similar allele frequency change as the truly selected loci. This process is known as hitchhiking and results in linkage disequilibrium (LD) (Smith & Haigh 1974). The excess of LD in *D. melanogaster* is thought to result from large segregating inversions within natural *D. melanogaster* populations (Barghi *et al.* 2017, Tobler *et al.* 2014). Hence the genetic signature of adaptation of *D. melanogaster* can be interpreted only with caveats.

In a recent study, Barghi *et al.* (2017) showed that the sister species of *Drosophila melanogaster, Drosophila simulans,* offers an alternative model organism for E&R studies, which portrayed much narrower genomic regions under selection (22.6 megabase pairs (Mbp) in *D. simulans* vs. 84.4 Mbp in *D. melanogaster*) in a stressful, hot-temperature, environmental regime. Individual selected SNPs were identified in more narrow genomic regions and in lower numbers in *D. simulans* than in *D. melanogaster* (918 candidate selected SNPs in *simulans* vs. 11.115 candidate selected SNPs in *melanogaster*) (Barghi *et al.* 2017). These observations suggest *D. simulans* to be an adequate model organism for E&R studies (at least in this particular selection regime) (Barghi *et al.* 2017).

2. E&R study of *D. simulans* populations in a hot-thermal environment

In 2013, in the lab of Christian Schlötterer at the Institute of Population Genetics of the University of Veterinary Medicine, Vienna, an E&R study was initiated from South African outbred *Drosophila simulans* flies. Female wild flies were captured in South Africa and used to establish 1278 isofemale lines. These 1278 isofemale lines were used to set up three "non-overlapping" supergroups, each starting from 426 isofemale lines. Five replicates of each supergroup were initiated, resulting in fifteen replicate populations in total. These replicates (effective population size (N_e) ~ 300), starting from standing genetic variation, were exposed to a stressful, hot-temperature environmental selection regime (cycling in 12h rhythm, 12 hours at 18°C followed by 12 hours at 28°C, mimicking night and day). Pool-sequencing was performed every 10 generations. After 50 non-overlapping generations in the selective regime, the genomic pool-seq information of the "evolved" populations (timepoint F50) was compared to the genomic pool-seq information of the ancestral base populations (timepoint F0) (Langmüller, unpublished). SNPs that showed an allele frequency change greater than expected by neutral genetic drift were identified and clustered by analysis orientated to the approach of Neda Barghi *et al*. (2019). The analysis (based on high-quality, triple mapped, Y-translocation masked, polymorphic sites in the *D. simulans* South African base population) was conducted by Anna Maria Langmüller and lead to the reconstruction of a haplotype-block, characterized by a cluster of 12 significant, top-candidate SNPs (identified using the R package *haploReconstruct*, see Franssen et al. 2016). Furthermore, the pool-seq information of intermediate timepoints (pool-seq every ten generations: F10, F20, F30, F40) allowed a description of allele frequency dynamics of individual candidate SNPs over the entire experiment and across all replicates (Langmüller, unpublished).
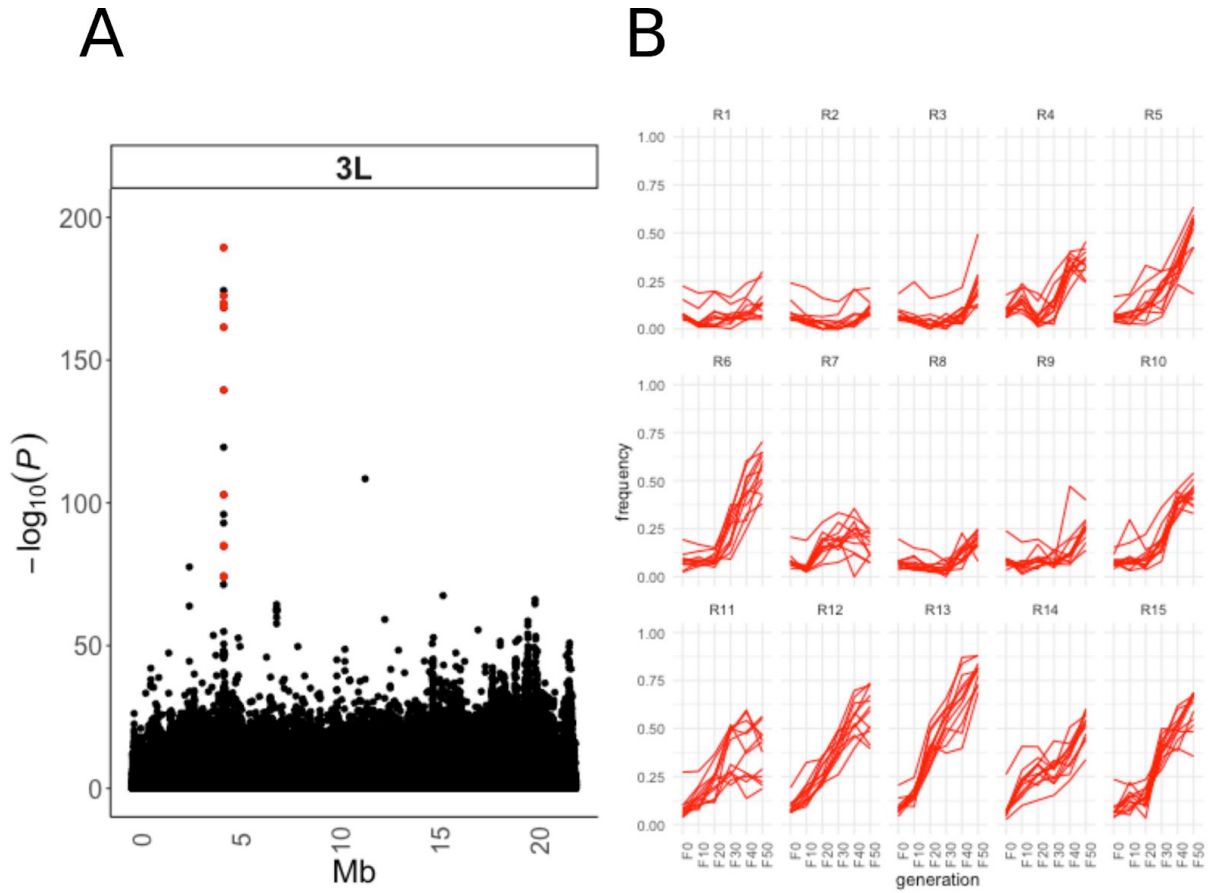
Figure 1 SNP cluster on 3L and trajectories of individual SNPs: (A) Manhattan plot showing the negative log10 transformed *p*-values from Cochran-Mantel-Haenszel test contrasting the ancestral (F0) with evolved (F50) populations, across positions on the chromosome 3L. The top-candidate, significant SNPs that were reconstituted as haplotype-block (*haploReconstruct*) are indicated in red. (B) Individual allele frequency trajectories of all top-candidate SNPs displayed over all 50 generations and across all replicates. Plots created by Anna Maria Langmüller.

These 12 top-candidate SNPs show an increase in allele frequency in 10 of the 15 replicates and are estimated to have a strong positive mean selection coefficient of 0.076 (estimated according to the approach of Taus *et al*. 2017). They cluster on 3L from position 4,568,235 to 4,571,357 (annotated genome *D.sim M252*, Palmieri *et al.* 2014). The identification of this remarkable narrow selected haplotype-block (3,122 bp) is extraordinary, as most E&R studies do not provide resolution of genomic regions that are subjected to evolutionary processes on the single-gene level (Turner *et al*. 2011, Remolina *et al*. 2012, Tobler *et al*. 2014, Zhou *et al*. 2011). This exceptional resolution proposes this genomic region to be a suited candidate for further follow-up experiments to elucidate the functionality of this sequence. Apart from the

12 candidate SNPs, this region harbors 328 additional non-significant SNPs (Langmüller, unpublished). The top-candidate polymorphisms are scattered over intronic and exonic regions of the protein-coding gene *GD13851* (one in exon no. 3, three in exon no. 4, eight in intronic regions). *GD13851* spans from Scf_3L:4,487,658 to 4,498,487 (*Drosophila simulans* R2.02, FB2019_05, Thurmond *et al.* 2019: Murphy 2014 *personal communication to FlyBase*, *Drosophila* 12 Genome Consortium 2008 *personal communication to FlyBase*, *Drosophila* 12 Genome Consortium *et al.* 2007; https://flybase.org/reports/FBgn0185548, 23.11.2019).
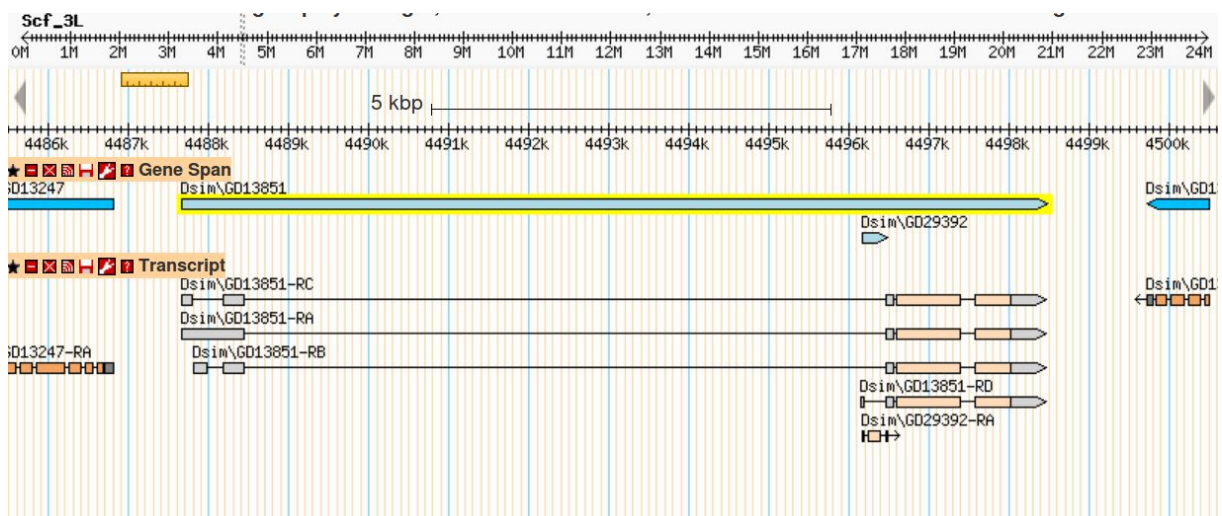
Figure 2 Approximately 15 kilobases (kb) of the genomic region illustrated, in which *GD13851* is located: The gene is depicted in light blue with yellow accentuation. All transcripts are portrayed, exons are marked by grey and pinkish boxes, introns indicated by a black line only. The Screenshot was obtained from FB2019_05, Thurmond *et al.* 2019, gBrowse: https://flybase.org/cgi-bin/gbrowse2/dsim/?Search=1;name=FBgn0185548 (Accessed: 23.11.2019)

3. Transcriptional regulation as selection target

A part of the intronic region of *GD13851* has been identified as a putative enhancer region by the open-access software JASPAR (http://jaspar.genereg.net/), a software that predicts TF binding motifs in genomic sequences (Sandelin *et al.* 2004). This putative enhancer region contains six out of the 12 top-candidate SNPs (see Figure 3).
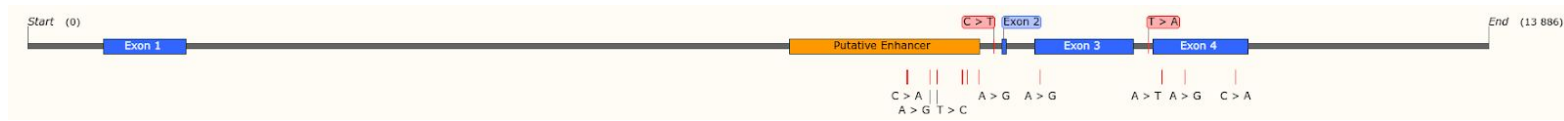
Figure 3 The locus of selection: The exons of *GD13851* are depicted in blue, introns are depicted as a grey line in between the exons. The region that was predicted by the software JASPAR (http://jaspar.genereg.net/) as an enhancer region is indicated in orange. The position of the SNPs that characterize both haplotype-blocks (ancestral and evolved) are shown in red, always in this order: "(ancestral > evolved)".

3.1 Gene-expression as main player in animal diversity

The genetic architecture that permits the great variety of phenotypes present on earth today, for example, the astonishing phenotypic richness of the metazoan kingdom, has always been of extraordinary interest in biology (Carroll 2008). Many techniques emerged in the second half of the 20th century that allow the investigation of the molecular and genetic fundamentals for this phenotypic diversity (e.g. morphology, anatomy, physiology, metabolism, behavior) and more and more light is shed on the genetic architecture of phenotypic idiosyncrasy and biological complexity.

Comparative genome analysis provided some unexpected findings, such as the realization that increased numbers of genes or increase in genome size do not account for increased complexity in organismal morphology, anatomy, physiology, and behavior (the human genome is only six times larger than some bacterial genomes) (Levine & Tijan 2003, Finlay & Esteban 2009, Taft *et al.* 2007). Furthermore, many developmental genes that encode for transcription factors (TFs) and cell-signaling molecules and which govern embryonic development, and thereby implicitly, morphology and anatomy, are highly conserved across phyla of the animal kingdom (Prud'homme *et al.* 2007, Carroll 2008). The formation of whole body parts and functionally equivalent organs is controlled by remarkably similar sets of orthologous genes in widely diverged animals (Prud'homme *et al.* 2007). These sets of fundamental genes show high structural conservation and functional equivalence (when homologs are placed in different taxa, e.g. mouse Pax-6 protein induces ommatidium

formation in *Drosophila*; *Cnidarian* Achaete-Scute homolog (*CnASH*) to induce formation of sensory organs in *Drosophila*) across bilaterians, despite hundreds of millions of years of independent evolution (Carroll 2008, Prud'homme *et al.* 2007). This extraordinary conservation seems to be the rule rather than the exception (Carroll 2008).

Another study that attracted a lot of attention was conducted by Mary-Claire King and A.C. Wilson in 1975, in the course of which they compared protein sequences of humans and chimpanzees by electrophoresis, and immunology, and a limited number of DNA fragments by DNA annealing methods. This study revealed a very modest degree of sequence divergence, in the same magnitude as between sister species in *Drosophila,* what is extraordinary regarding the profound divergence in anatomy, behavior, ecological context, and morphology between humans and chimpanzees (King & Wilson 1975).

How can the (relatively) small number of similar genes be exploited to generate the multitude of existing and increasingly complex phenotypes? One feasible option is the establishment of complex spatiotemporal gene expression patterns during an organism's life cycle (Levine and Tijan 2003). A vast body of studies has proven that spatiotemporal gene expression divergence contributes to higher-level phenotypic divergence, and modulated spatiotemporal gene expression is sufficient to recreate phenotypic differences (Wray 2007, Wittkopp & Kalay 2012).

3.2 Regulation of Gene Expression

The functionality of every polypeptide is dependent on multiple factors: 1) on the molecular function that the protein is executing, 2) on the biological process in which the protein is involved, 3) on the cellular compartment in which the protein is located, 4) and on the spatial (e.g. cell-type, organ) and 5) temporal units in which the protein is synthesized in the first place (Wray 2007, Lee *et al.* 2007). While only 2 % of the human genome codes for proteins, it is estimated that one-third of the genome comprises regulatory regions that control replication, condensation, pairing, segregation, and gene expression. Furthermore, 5-10 % of the total protein-coding capacity of metazoans is dedicated to proteins involved in transcription regulation (Levine & Tijan 2003). Unlike genome size or gene number, a strong

correlation is observed between biological complexity and the ratio of non-protein-coding DNA sequences to absolute genome size (Taft *et al.* 2007).

Gene expression can be regulated at least at any of the six following potential control points: 1) contextual activation of the gene (i.e. permissive chromatin structure), 2) initiation of transcription and elongation, 3) processing of the transcript, 4) transport from the nucleus to the cytoplasm, 5) translation of mRNA, and 6) degradation and turnover of the mRNA (Krebs *et al.* 2014). To initiate transcription and elongation, RNA polymerase II assembles with the basal transcriptional machinery at the promoter, a sequence in immediate vicinity to the transcription start site. The rate of transcription is often weak in the absence of additional factors (Krebs *et al*. 2014, Shlyueva *et al.* 2014). These additional factors that modulate gene transcription may be classified into two distinct categories: Factors that segregate in linkage to the gene (i.e. *cis*-regulatory elements) or factors that segregate independently to the gene (i.e. *trans*-regulatory elements) (Lemos *et al.* 2008, Wittkopp & Kalay 2012).

Trans-regulatory elements are proteins that interact with specific DNA sequences and thereby govern transcription (Alberts *et al.* 2014, p. 310-314). Transcription factors (TFs) are one type of *trans*-regulatory protein. TFs regulate the transcription of a target gene positively or negatively as a consequence of sequence-specific interactions with transcription factor binding motifs (Latchman 1997).

Cis-regulatory elements (CREs) are non-protein coding DNA regions that contain particular nucleotide sequences that are recognized by *trans*-regulatory proteins. The recognition is caused by intimate, thermodynamically favorable, non-covalent interactions (hydrophobic interactions, electrostatic forces, hydrogen bonding) between the protein surface and the nucleobases exposed in the minor or major groove of the DNA double-helix (Seeman *et al.* 1976, Tolstorukov *et al*. 2004, Von Hippel 2007). These protein-DNA interactions are among the strongest and most specific molecular interactions known in biology (Alberts *et al.* 2014, p. 266). Any one gene is usually regulated by multiple CREs, and typically each CRE contains multiple binding sites for *trans*-regulatory proteins, acting both positively and negatively on transcriptional activity (Prud'homme *et al.* 2007, Levine & Tijan 2003, Latchman 1997).

3.3 Enhancers

Enhancers are one type of CRE and typically consist of concentrated clusters of multiple transcription factor binding motifs and are typically 100 bp to 1000 bp in length (Long *et al.* 2016, Wittkopp & Kalay 2012, Prud'homme *et al.* 2007, Shlyuevva *et al.* 2014, Kim & Shiekhattar 2015). Even though the ability of enhancers to initiate and enhance transcription, if bound by TFs, is commonly emphasized, enhancers can also contain binding motifs for proteins acting to repress transcription (Arnosti & Kulkarni 2005, Crocker & Ilsley 2017).

Each transcription factor binding motif (referred to as motifs in the following) typically acts in an autonomous and modulatory manner and adds a context-dependent nuance to the overall activity of the enhancer (Long *et al.* 2016, Crocker & Ilsley 2017, Shlyueva *et al.* 2014). Motifs are degenerate sequence patterns (i.e. more than one nucleobase allowed at particular positions) with a length of 6 - 10 base pairs (bp) and summarize the binding preference of a particular TF (Shlyueva *et al.* 2014). The structural organization of most enhancers is highly flexible, in particular regarding: 1) the motif types (different motif types get recognized by different TFs), 2) individual motif affinities, 3) number of motifs, 4) their order, 5) spacings between motifs, 6) orientation of the motifs to each other, 6) and the local DNA shape superimposed to the motifs (Wittkopp & Kalay 2012, Long *et al.* 2016). The feature of degeneration allows for fine-tuning of transcription factor binding specificity via non-optimal matches at degenerate positions; employment of multiple suboptimal motifs may promote specificity of regulation without losing signal strength; optimization of suboptimal motif sequences have been shown to cause stronger and ectopic enhancer activity (Long *et al.* 2016). Functional conservation of enhancers between both closely and distantly related species is observed to be much more common than conservation at the sequence level. Enhancer regions show substantial motif turnover and losses of orthologous motifs are often compensated by motif gains at non-orthologous positions (Arnold *et al.* 2014, Taher *et al.* 2011). This predominant flexible structural organization of enhancers, in addition to the high degeneration of motifs, results in enhancer sequence conservation levels in between the conservation level of protein-coding sequences and non-functional sequences (Wittkopp & Kalay 2012).

However, this flexible organization of enhancers is no strict paradigm as other enhancers with deep structural and sequential conservation have been discovered. In those enhancers, any mutation that disrupts type, number, and positioning of motifs interferes with enhancer functionality, due to the highly cooperative manner in which TF complexes bind at these CREs (e.g. IFNβ gene enhancer, see Thanos & Maniatis 1995). A proposed rationale behind this organization is that these enhancers serve as on/off switches that rely on multiple strictly required inputs (TFs), and accordingly, it has been observed that these types of enhancers often cluster close to genes encoding for developmental transcription factors, so genes with enormous pleiotropic effects. Although, this enhancer organization rather seems to be the exception (Long *et al.* 2016, Arnosti & Kulkarni 2005). Mostly, enhancers are organized in between the spectrum of imperative conservation and complete flexibility, with defined spacing, order, and orientation required for some motifs, but not for others within a single enhancer (Long *et al.* 2016).

Enhancers are scattered over the genome and 1) act to drive transcription independent of their relative distance, location and orientation to their cognate promoter, 2) increase transcription of the linked gene from its correct transcription initiation site specified by the core promoter, 3) are capable to function with (many) heterologous promoters, 4) exhibit DNase hypersensitivity I, furthermore, enhancers contribute 5) modular, 6) additively, 7) autonomous, and 8) partly redundant to the overall expression pattern of their target gene (Long *et al.* 2016, Wittkopp & Kalay 2012, Prud'homme *et al.* 2007, Shlyuevva *et al.* 2014, Kim & Shiekhattar 2015, Small *et al.* 1993, Gray *et al.* 1994).

One unique feature of enhancers is their ability to drive transcription from their cognate promoter over long genomic distances. This allows a single gene to be controlled by a large number of independent enhancers, and generates enormous combinatorial complexity of gene-expression repertoires (Long *et al.* 2016, Shlyueva *et al.* 2014). Indeed, a defining feature of metazoan gene expression is the use of numerous enhancers (and also other CREs, e.g. promoters, insulators) to control the expression of a single gene (Levine & Tijan 2003). Moreover, in metazoans, enhancers are the primary regulators of gene expression in discrete spatiotemporal domains (Kim & Shiekhattar 2015, Prud'homme *et al.* 2007).

3.4 Enhancers in the aspect of evolution

The flexible, modular architecture and the additive and partly redundant functionality, both within and between enhancers, have profound implications for the evolutionary fate of those genomic elements.

Compared to *trans*-regulatory elements, *cis*-regulatory elements such as enhancers are attributed to be smaller mutational targets, thus having a lower probability of being mutated (Signor & Nuzhdin 2018, Lemos *et al*. 2008). In conformity with this, diversity in *trans*-regulatory factors contributes more to intraspecies gene expression variation (Signor & Nuzhdin 2018, Lemos *et al.* 2008). On the other hand, studies demonstrated a greater contribution of *cis*-regulatory divergence to interspecies gene expression differences (Arnold *et al*. 2014, Signor & Nuzhdin 2018, Wittkopp *et al.* 2004). The hypothesized rationale for these contrary findings is that mutations in protein-coding sequences are more likely to arise, but are also more reasonable to have pleiotropic effects, and therefore are expected to be (most likely) deleterious and as a consequence selected against (Signor & Nuzhdin 2018, McManus *et al.* 2010).

The modular, autonomous organization within and between enhancers implies that gene expression can be modulated in individual, restricted domains (e.g. tissues, cell-types, developmental stages), without any effect on gene expression in other domains. This modularity provides reduced pleiotropy to mutations in *cis*-regulatory sequences, and allows subcompartments of organisms to evolve somewhat independently; reduced pleiotropy allows selection to operate more efficiently by minimizing functional trade-offs (Wray 2007); modularity is thought to be a major driver of evolutionary novelty (Prud'homme *et al.* 2007, Wittkopp & Kalay 2012, Alberts *et al*. 2014, Jeong *et al*. 2008).

Differential gene expression resulting from *trans*-regulatory polymorphisms is also considered as (most probable) being recessive (or dominant) rather than additive (Signor & Nuzhdin 2018, McManus *et al*. 2010). *Cis*-regulatory divergence is expected to exhibit additive (i.e. co-dominant) inheritance because transcripts from the maternal and paternal

alleles are assumed to contribute independently to total levels of gene expression in the diploid genome (McManus *et al*. 2010). This implicitly affects the efficiency with that selection can act upon *cis*-regulatory elements: selection can act immediately on co-dominant traits, because they transfer a fitness effect as soon as they appear in the population, from a low allele frequency on. Recessive traits, on the other hand, supply a fitness effect only in homozygotes, and as a consequence, random genetic drift is required to raise the allele frequency until homozygotes start to appear within the population (Wray 2007). This increases the probability of emerging recessive beneficial alleles for getting lost and decreases their probability of fixation.

Additional to these (in aspect of evolvability) beneficial characteristics of enhancers, empirical findings that promoters are much more conserved across species (Wittkopp & Kalay 2012), and the deep functional conservation of *trans*-regulatory proteins, further suggests *cis*-regulatory elements such as enhancers to be the major mean for gene expression evolution (Prud'homme *et al*. 2007). We suggest that the region predicted by JASPAR (see Figure 3) is indeed an enhancer and that the candidate SNPs present in this sequence cause a change in spatiotemporal transcription of *GD13851*. We further propose that this modulation of transcriptional activity is the beneficial adaptation that leads to increased fitness in a stressful, hot-thermal environment.

4. O-Glycosylation of proteins in the CNS

GD13851 has no experimental supported evidence for gene ontology in *Drosophila simulans*. Nonetheless, the interpro project (2004-) has predicted *GD13851* to posses galactosyltransferase activity (classified as glycosyltransferase family 31), to play a role in the biological process of protein glycosylation and to be located in the cell membrane (Flybase Curators *et al*. 2004 -, from FB2019_05, Thurmond *et al*. 2019: https://flybase.org/reports/FBgn0185548, 23.11.2019). Four transcripts and four polypeptides are annotated (from FB2019_05, Thurmond *et al*. 2019: https://flybase.org/reports/FBgn0185548, 23.11.2019).

GD13851's syntenic ortholog from *D. melanogaster, CG11357,* has shared sequence ID of 92 % (7488/8170, query coverage: 93 %) across gene length (10830 bp in *simulans*, 11407 bp in *melanogaster*, sequence information from FB2019_05, Thurmond *et al.* 2019: https://flybase.org/download/sequence/FBgn0185548/FBgn, https://flybase.org/reports/FBgn0035558, 23.11.2019, *CG11357* annotated by: Ashburner 1999, FamiliarityBreedsContempt 1999), and up to ~94 % analogy (query coverage: 100 %) on amino acid sequence level (MEGABLAST, and BLASTP, default settings. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2019]; accessed: https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Furthermore, the orthologue *CG11357* is predicted to have a role in O-linked protein glycosylation by the InterPro Project (2004-), by the GO Reference Genome Project (Gaudet *et al.* 2011), and by Schwientek *et al.* (2002) (FB2019_05, https://flybase.org/reports/FBgn0035558, 23.11.2019). *CG11357* is also a gene without experimentally validated enzymatic function but with conserved structural domains. It is suspected to function as O-linked N-acetylgalactosaminyltransferase, O-linked acetylglucosaminyltransferase, and galactosyltransferase (FB2019_05, https://flybase.org/reports/FBgn0035558, Schwientek *et al.* 2002, Gaude *et al.* 2011, Yamamoto-Hino *et al.* 2015).

O-glycosylation describes the conjugation of polypeptides with monosaccharides or oligosaccharides at the oxygen atom of the hydroxyl group of serine, threonine, hydroxyproline, hydroxylysine or tyrosine. O-Glycosylation governs physical parameters like molecule stability and tertiary protein conformation and numerous biochemical and physiological processes: 1) cell-cell interactions, 2) protein-protein interactions, 3) protein folding, 4) protein localization, 5) secretion, 6) immunization, 7) stress response (Li *et al.* 2019, Staudacher 2015, Zhang *et al.* 2008). Secretory proteins and transmembrane proteins are typically glycosylated in multicellular eukaryotes, these modifications provide biophysical and biochemical properties, adequate for interactions in the extracellular matrix, which is rich in glycans and glycoconjugates. The glycosylation of transmembrane proteins and secretory proteins usually occurs in the endoplasmic reticulum-Golgi pathway (Colley *et al.* 2017).

O-linked N-acetylglucosaminyltransferase (O-GlcNAcylation) is abundant in the brain, aberrant O-GlcNAcylation has been associated with neurodegenerative diseases (e.g. Alzheimer's disease, Parkinson's disease), and extensive crosstalk between O-GlcNAcylation and phosphorylation has been observed, where imbalances lead to drastic outcomes (Hart *et al.* 2011). Furthermore, O-GlcNAcylation of the cytoplasmic Milton protein, which associates with kinesin-1 and enables axonal transport of mitochondria (Glater *et al.* 2006), correlates with mitochondrial motility in neurons and allows mitochondria to respond to intracellular glucose availability gradients (Pekkurnaz *et al.* 2014). Mitochondrial motility has direct implications on glucose metabolism, ATP supply, cell-signaling and Ca_{2+} buffering in cellular subcompartments in neurons (Pekkurnaz *et al.* 2014). A huge variety of neuronal proteins are O-GlcNAcylated and this modification has been witnessed to be determinant for neuronal plasticity (Hart *et al.* 2011). The Notch transmembrane receptor protein is central in neuronal development (reviewed in Justice & Jan 2002), and portrays multiple potential O-GlcNAc conjugation sites, on both its intracellular and extracellular domain. O-Glycosylation of Notch has the tendency to promote cleavage and activation of Notch (Stanley *et al.* 2010). The gene *brainiac*, another member of the glycosyl transferase family 31, exhibits a mutant phenotype that resembles the *notch* mutant phenotype during oogenesis (Schwientek *et al.* 2002).

CG11357 (*GD13851*'s ortholog in *D. melanogaster*) loss of function mutation caused by P-element insertion has been shown to result in increased habituation (Eddison *et al.* 2012). Compatible with this finding, activity of the Ca_{2+} channel, which mediates the essential mechanisms of long-term potentiation and depression in neurons, has been observed to be modulated by O-GlcNAcylation, with a reciprocal effect of O-GlcNAcylation level on Ca_{2+} transients (Rengifo *et al.* 2007). Whole-body gene silencing of *CG11357* by RNAi leads to death during development, before flies reach the third instar larval stage (Yamamoto-Hino *et al.* 2015).

Contemplating all the biological processes in which *GD13851*'s ortholog and protein family members are engaged, we evaluate the probability that *GD13851* is involved in neurobiological processes as high. We hypothesize that a supposed change in *GD13851* transcription pattern, resulting from the selected haplotype-block, occurs in the *D. simulans*

brain, and further that neurobiological mechanisms are modulated as a consequence of the changed *GD13851* expression in the fly brain.

5. Testing regulatory activity via enhancer-reporter assays

Enhancer-reporter assays are designed to test the ability of DNA sequences to regulate transcription of a reporter gene (e.g. GAL4, *lacZ*, GFP, QF, lexA) remote of their native contexts, from a minimal promoter, *in vivo* (Jory *et al.* 2012). In this experimental approach, a (enhancer) DNA sequence, typically 2-3 kb in length (Pfeiffer *et al.* 2008, Kvon *et al.* 2014, Jenett *et al.* 2012), whose regulatory activity is to be elucidated, is inserted into an appropriate vector, upstream of a reporter gene (this construct is then termed enhancer-reporter vector or donor plasmid for genomic integration). Consequently, this enhancer-reporter construct is integrated into the genome of a model organism and the detection of the reporter gene allows the *in vivo* characterization of the regulatory activity of the (putative) enhancer sequence (Kvon 2015, Janssen *et al.* 2006, Naylor 1999).

Typically the reporter gene activity recapitulates a fraction of the expression pattern of the gene from which the enhancer originates (Kvon 2015). This mirrors the modular, autonomous, and the additive nature of enhancers, where multiple enhancers collectively regulate the net temporal and spatial expression pattern of a single target gene (Jeong *et al.* 2008, Shlyueva *et al.* 2014, Pfeiffer *et al.* 2008, Gray *et al.* 1994). Another pattern that has been observed before is that enhancer sequences, if dislocated from their native genomic environment, drive transgene expression in cells and tissues in which the endogenous gene, from which the enhancer originates, is not expressed. This may be due to the absence of inhibitory CREs or lack of competition with other enhancers (Pfeiffer *et al.* 2008).

We aim to execute an enhancer-reporter assay to inspect the regulatory activity of either haplotype-block of the putative enhancer sequence (evolved haplotype-block or ancestral haplotype-block, i.e. with the "evolved" version of the top-candidate SNPs or with the "ancestral" version of the top-candidate SNPs, see Figure 3). Therefore we perform microinjections to integrate our enhancer-reporter constructs into the genome of *Drosophila*

simulans flies. Our objective is to 1) investigate the overall spatiotemporal regulatory activity of the reconstituted haplotype-block (see Figure 3) that contains the putative enhancer, and to 2) elucidate any non-conformity in regulatory activity between the selected and ancestral haplotype-version of this sequence. Especially the second point is of particular interest as any regulatory divergence between the two haplotype-blocks is hypothesized to be the causative genomic adaptation leading to increased fitness in a stressful, hot-temperature environment.

6. *Drosophila simulans* strain #2176

We use the *D. simulans* strain #2176 (*Dsim*#2176), which has been generated by Stern at el. (Genes Genomes Genetics, 2017), as recipient strain for the integration of our enhancer-reporter constructs. This strain possesses an endogenous attP site and has been generated by the random integration of the plasmid *pBac{3XP3::EYFP-attP}* into the genome of *D. simulans yellow (y^-) white (w^-, [w1])* (San Diego Species Stock Center stock number 14021-0251.013) by the transposase *piggyBac* (Stern *et al*. 2017). After the successful integration of the plasmid, EYFP expression has been knocked-down by CRISPR/Cas9-mediated targeted mutagenesis, for EYFP to not conflict with any further use. The plasmid *pBac{3XP3::EYFP-attP}* integrated into the left arm of chromosome two (2L) at the position 6392951, outside of any protein-coding regions (Stern *et al.* 2007).

Dsim#2176 harbors the allele *w1* that has been first described by Thomas Hunt Morgan as spontaneous mutation (Science, 1910). He observed the recessive mutation *w1* to causes a change in fly eye color from red (wildtype phenotype w^+) to white (mutant phenotype w^-), and furthermore, he noticed that this mutation was predominantly transmitted to male progeny. The allele *w1* is characterized as a loss of function (Sabl & Birchler 1993), hypomorphic (Lloyd *et al*. 2002), and amorphic (Judd 1995) allele, and results from a segment insertion at the X-linked *white* gene (Zachar & Bingham 1982). The *white* gene encodes for an ABC-type transmembrane transporter (Lloyd *et al*. 2002) that is strictly required for synthesis of the pigment ommochrome in any tissue, including the compound eyes and ocelli (Tearle 1991). This particular transmembrane transporter is involved in the transport of ommochrome precursor proteins from the cytoplasm into pigment granules (Mackenzie *et al*. 2000). As a

consequence, loss of function of the *white* gene prevents the pigmentation of eye and ocelli, resulting in abnormal white eyes in *Drosophila*.

Stern *et al.* tested the φ-C31-mediated integration efficiency for *Dsim#2176* and reported a rather high integration efficiency of 1/21. Albeit integration efficiency is also dependent on the donor plasmid and the source of the integrase, therefore integration efficiency is not strictly comparable in heterologous settings (Stern *et al.* 2007, Bischof *et al.* 2007).

7. The *lacZ* enhancer-reporter vector, pBlueRabbit

The plasmid pBlueRabbit (pBR) (Housden *et al.* 2012), which was kindly provided by Kat Millen, was chosen as vector for the enhancer-reporter assays. The components of pBR (see Figure 4) are discussed in the following: The reporter gene *lacZ* is expressed from an *Hsp70* minimal promoter (as discussed in the last section) to enable analysis and detection of enhancer activity (Housden *et al.* 2012). *lacZ* encodes the enzyme *β*-galactosidase, transcription of the gene can be either measured directly by RNA *in situ* hybridization, by reverse-transcribed quantitative PCR (RT-qPCR) (Stahlberg *et al.* 2004, Overbergh *et al.* 2003) or indirectly by means of colorimetric analysis (by exposure to X-Gal, X-Gal will be cleaved by *β*-galactose, and further oxidized to the insoluble dye 5-Bromo-4-chloro-indigo, which can be measured). The spatial and temporal expression of the gene can be assessed both quantitative and qualitative (Janssen *et al.* 2006, Serebriiskii & Gelomis 2000, Kvon 2015).

pBR carries a version of the *white* gene, in which large parts of the first intron and regulatory region are missing, appropriately this gene is termed *mini-white* (or *mini-w*) (Pirrotta 1988). Despite the reduced size, *mini-white* largely remains the functionality of *white*, and if *mini-white* is introduced into *white* deficient *Drosophila* flies, it rescues the mutant phenotype (white eye phenotype) and re-establishes eye and ocelli pigmentation (resulting in yellow or orange or red eyes). The degree of pigmentation induced by *mini-white* varies in between the color spectrum of yellow to red, dependent on the genomic integration locus of *mini-white* and the characteristics of the vector by which *mini-white* is delivered (Pirrotta 1988, Silicheva

et al. 2010). Hence the dominant marker *mini-white* (w^+) present in pBR greatly expedites identification of integration events in *Dsim*#2176, because the transformation of pBR into the *Dsim*#2176 genome will be demonstrated by red eye phenotype (instead of white eye phenotype that characterizes *Dsim*#2176).

Furthermore, pBR contains the kanamycin (*kan*) resistance gene for bacterial selection, an attB site for φ-C31 mediated integration, a loxP site that allows removal of *kan*, and *gypsy* insulator elements (Housden *et al.* 2012).

The *gypsy* insulator sites positioned on both sides of the enhancer-reporter sequence have been shown to confer robust transgene expression and to decrease the magnitude of position effects. Insulator sites frequently lead to increased transgene activity, as effects on transgene expression are observed to be predominantly repressive. The *gypsy* insulators are known to block the effect of over 20 *cis*-regulatory elements. No aberrant basal activity and ectopic expression resulting from *gypsy* insulator sites have been observed yet (Markstein *et al.* 2008, Barolo *et al*. 2000). Additionally, the vector is arranged in a way that the *mini-white* gene, the *kan* gene, and the vector backbone also flank the enhancer-reporter sequence, with the purpose to further minimize the effect of neighboring sequences upon genomic integration of the plasmid (Housden *et al.* 2012).

The vector pBR showed no basal expression in *Drosophila melanogaster* (in larval brains and trachea) and reported a known enhancer expression pattern (enhancer of the *Notch* gene) accurately and reproducibly. Additionally, no ectopic gene expression was observed when pBR was integrated into a genomic location prone to position effects (Housden *et al.* 2012). Considering all these characteristics, pBR seems to be a suitable vector for enhancer-reporter assays.
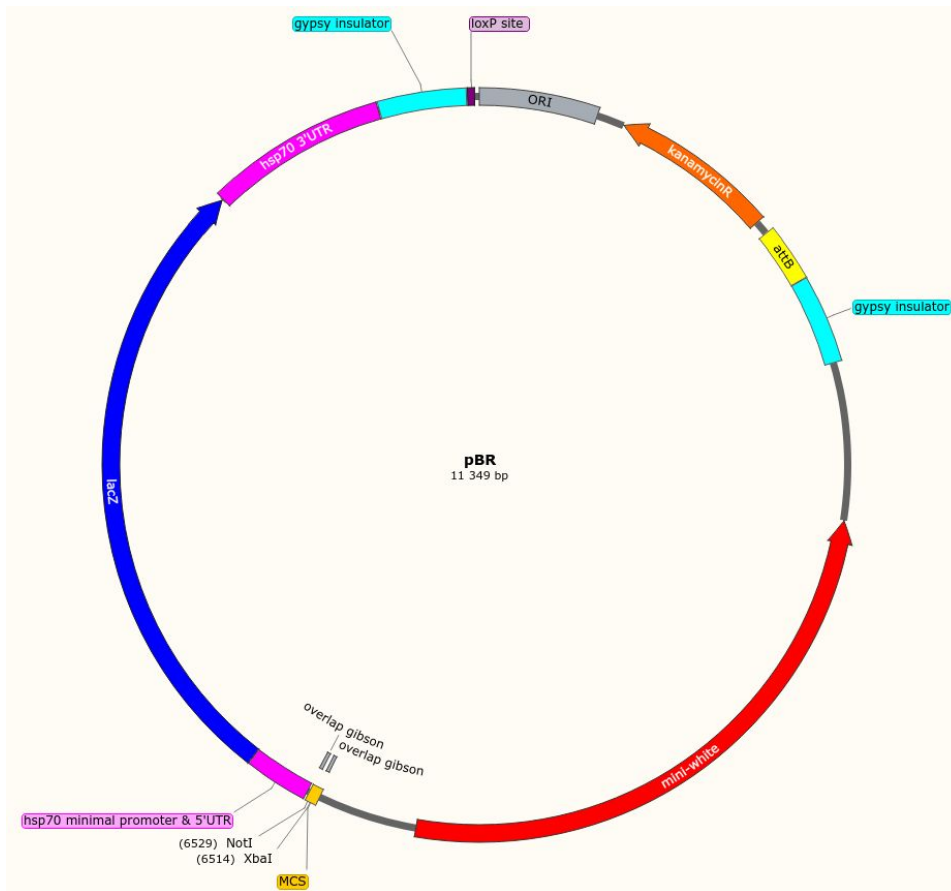
Figure 4 All elements of pBlueRabbit (pBR) annotated: origin of replication (ORI), kanamycin resistance gene (kanamycinR), φ-C31 bacterial attachment site (attB), *gypsy* insulator sites, *mini-white* marker, multiple cloning site (MSC), the recognition sites for the enzymes that were used for linearization (XbaI and NotI) are indicated, the sequence overlap in the MSC for Gibson assembly also shown (overlap gibson), *Hsp70* minimal promoter with 5'UTR, *β*-galactosidase gene (*lacZ*), *Hsp70* 3'UTR and the loxP site.

8. The helper plasmid vas-φ-C31(3xP3-EPFG)

To facilitate transformation we coinject, additionally to the enhancer-reporter vectors that are to be transformed, the plasmid vas-φ-C31(3xP3-EGFP) that provides the φ-*C31* integrase. It has been demonstrated previously that this plasmid is suitable for microinjection purposes and transfers a high integration success in *Drosophila melanogaster* (Bischof *et al*. 2007, Zhang *et al*. 2014). In this plasmid, the expression of the φ-*C31* integrase is governed by the promoter and the 5' UTR of the *vasa* gene, which is known to drive zygotic transcription in primordial germ-cells of *D. melanogaster* (Sano *et al*. 2002, Bischof *et al*. 2007, Zhang *et al*. 2014).

9. Impacts on transcription regulation

One parameter that influences the regulatory activity of a CRE is the composition of the core promoter from which expression is driven, as the core promoter is an active regulatory component in transcription regulation (Butler & Kadonaga 2001). Selective communication between enhancers and core promoters is an often observed phenomenon: enhancers require the presence of specific sequence motifs in the core promoters; different enhancers have distinct preferences for various sequence motifs; different core promoters harbor various sequence motifs; it has been shown that not all enhancers function efficiently with all core promoters (Pfeiffer *et al.* 2008, Smale & Kadonaga 2003, Zabidi & Stark 2016, Butler & Kadonaga 2001). This selective communication is considered as an evolutionary means to further increase the gene expression repertoire, again, in a combinatorial manner (Smale & Kadonaga 2003, Smale 2001, Butler & Kadonaga 2001). The selective communication suggests that factors bound to enhancers must directly or indirectly interact with factors bound to the core promoter (Smale 2001).

The difference between the two (putative) enhancer versions which we are examining is compounded by the 12 major SNPs and also by a vast amount of hitchhiking, non-significant SNPs (328 in number). I assume that the spectrum of TFs, that exhibits motif binding activity in the enhancer, is required to shift in order for the enhancer to change its core promoter preference. However, selective communication between enhancers and core promoters remains largely obscure and experimental data is required for the validation of enhancer preferences (Zabidi *et al.* 2014).

In pBR the reporter gene *lacZ* is expressed from the minimal promoter of the housekeeping gene *Hsp70*. The sequence that we are using contains the core promoter region up to position - 50 relative to the initiation start site (Hiromi & Gehring 1987). The TATA-box present in the *Hsp70* acts as a weak promoter, the sequence region that is required for heat-shock inducible expression is not included in this fragment (Hiromi & Gehring 1987, Pelham 1985, Smith *et al.* 1993). The *Hsp70* promoter is a prominent and standard promoter for transgenic expression in *Drosophila* (Barolo *et al.* 2000, Zabidi *et al.* 2014, Stern *et al.* 2017).

Transgene expression (and gene expression in general) is known to be highly affected by surrounding sequences and the local chromatin configuration (summarized as position effects) (e.g. Groth *et al.* 2004, Bischof *et al.* 2007, Housden *et al.* 2012, Markstein *et al.* 2008, Barolo *et al.* 2000, Fish *et al.* 2007, Shlyueva *et al.* 2014, Long *et al.* 2016). Enhancer activity correlates with diverse histone modifications which surrogate chromatin accessibility and relaxed nucleosome density (Shlyueva *et al.* 2014, Kim & Shiekhattar 2015, Jory *et al.* 2015, Kvon *et al.* 2014). Particular integration loci have been shown to result in ectopic reporter gene expression (Kvon 2015, Housden *et al.* 2012). The effect of a particular chromosomal position can be tested prior via the integration of a reporter gene, accompanied solely by the minimal promoter of choice (Pfeiffer *et al.* 2008). Another possibility is to integrate the reporter gene with an enhancer whose regulatory profile is already approved (Housden *et al.* 2012). However, if uniform enhancer-reporter vectors, that only differ in enhancer inserts, are integrated precisely into the same genomic location (i.e. site-specific), then the influence of the genomic environment is considered constant. Consequently, peculiarities in reporter gene expression pattern can be accredited as inherent regulatory features of the respective enhancer insert. This greatly facilitates the comparison of the regulatory profile of DNA sequences (Kvon 2015, Groth *et al.* 2004, Jenett *et al.* 2012, Housden *et al.* 2012, Fish *et al.* 2007).

Site-specific recombination may be accomplished by a variety of recombinase systems, the most prominent are the Flp, Cre, and φ-C31 recombinases (Oloruniji *et al.* 2016). We choose φ-C31 as integrase system as it enables non-reversible, high-frequency integration events, with unspecific integration events being reported as rather rare (Bischof *et al.* 2007). The φ-C31 integrase catalyzes sequence-directed recombination between a bacterial (attB) and a phage (attP) attachment site (Barolo *et al.* 2000, Bischof *et al.* 2007, Grandchamp *et al.* 2014). The integration event is non-reversible as recombination between the two sites consumes the attB and attP site, and results in the generation of two hybrid sites (attL and attR) flanking the integrated fragment. These hybrid sites do no longer allow φ-C31 mediated relocation of the transgene (Fish *et al.* 2007, Grandchamp *et al.* 2014).

Additionally to the shared loci of integration, the enhancer-reporter constructs will also encounter the same *trans*-regulatory environment upon integration into the genome of an isogenic *D. simulans* strain (*Dsim*#2176). Eventually, as we set the parameters of position

effects, *trans*-regulatory environment, and core promoter constant, a direct comparison between the ancestral and the evolved version of the putative enhancer is feasible.

10. Aim and hypothesis

Based on E&R data from South African *Drosophila simulans* populations and their follow-up analysis from Anna Maria Langmüller, we do know that a particular "evolved" haplotype-block at the chromosome 3L (position 4,568,235 to 4,571,357) results in an increased fitness value of *D. simulans* flies in a stressful, hot-temperature environmental regime. Yet, the explicit fitness benefit resulting from this selected genomic variant is not known. This haplotype-block happens to be located in the protein-coding gene *GD13851*, moreover, exactly half of the 12 candidates SNPs which characterize the haplotype-block are located in a putative CRE (identified with http://jaspar.genereg.net/) of *GD13851*. In contemplation of available literature on the evolutionary fate of CREs, we evaluate the probability that this intronic region is indeed an enhancer as high. We further hypothesize that the top-candidate SNPs in this (putative) enhancer region cause a change in spatial and/or temporal transcription pattern.

No experimental validated function has been reported for *GD13851* yet, but the gene belongs to the glycosyl family 31, a group of glycosyltransferases that is known to play a crucial role in neuronal development, signaling and physiology. Considering this, we believe that the hypothesized modulation of transcriptional regulation occurs (also) in the fly brain. To test these hypotheses we aim to perform an enhancer-reporter assay with either ancestral or selected haplotype-block, and examine the spatiotemporal gene expression pattern of either putative enhancer version in the fly brain.

Materials & Methods

Table 1. Devices and technical material

Material or Device	Manufacturer
Universal hood II, Molecular Imager Gel Doc XR system	Bio-Rad Laboratories Gmbh
ProFlex base	Applied Biosystems, Thermo Fisher Scientific
CFX Connect Optics Module	Bio-Rad Laboratories Gmbh
2720 thermal cycler	Applied Biosystems, Thermo Fisher Scientific
Centrifuge 5425	Eppendorf AG
PIPETMAN classic p2	Gilson Incorporated
PIPETMAN classic p10	Gilson Incorporated
PIPETMAN classic p20	Gilson Incorporated
PIPETMAN classic p100	Gilson Incorporated
PIPETMAN classic p200	Gilson Incorporated
PIPETMAN classic p1000	Gilson Incorporated
PIPETBOY acu 2	INTEGRA Biosciences GmbH
GD 100-p12	Grant Instrument™
PSC-20	Grant Bio™
MSH basic Yellow-line	IKA®-Werke GmbH & Co. KG
L13701 Shaking Incubator	GFL Gesellschaft für Labortechnik GmbH
Universal lab incubator inb50	Memmert GmbH + Co. KG
Test-tube-rotator 34528	Snijder Labs
vortex-genie touch mixer 1	Scientific Industrie, Inc.
MW 2235 CW	Bomann
MP-250V Power Supply	Major Science
Horizon 58 Agarose Gel Electrophoresis Apparatus	Apogee Electrophoresis
Horizon 11-14	Life Technologies Inc.
Dr-36vl	CLF Plant Climatics GmbH
M-152 Manipulator	NARISHIGE Group
Femtojet 5247	Eppendorf AG
Stereomikroskop ms5	Leica Mikrosysteme Handelsges.m.b.H.
Trinokulartubus	Leica Mikrosysteme Handelsges.m.b.H.
IMC 521234 Schieber Integrated Modulation Contrast	Leica Mikrosysteme Handelsges.m.b.H.
Kl 200, Cold Lightsource	SCHOTT AG
LaboStar® Ultra pure water and reverse osmosis	Evoqua Water Technologies GmbH

systems

VX-120, Autoclave	Systec GmbH
MDF-U7386S, HCFC-Free Ultra-Low Temperature Freezer	SANYO Electric Co., Ltd.
Centrifuge 5804	Eppendorf AG
Centrifuge 5430R	Eppendorf AG
Centrifuge 5424 R	Eppendorf AG
multiply® - µStrip 0.2 mL chain, qPCR strip	SARSTEDT AG & Co. KG
8 Lid chain flat	SARSTEDT AG & Co. KG
PCR 8er-capstrips, farblos, doomed	Biozym Scientific GmbH
Pipette tip 200 µL gelb	SARSTEDT AG & Co. KG
Pipette tip 20µL farblos	SARSTEDT AG & Co. KG
Pipette tip 1000µL blue	SARSTEDT AG & Co. KG
CELLSTAR® Serological Pipettes 2 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 5 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 10 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 25 mL	Greiner Bio-One International GmbH
96 Well Microplatte	Greiner Bio-One International GmbH
Embryo Collection Cage-Large	Genesee Scientific Corporation
Micro Tubes 1.5 mL Neutral	SARSTEDT AG & Co. KG
Safeseal micro tube 2mL PP	SARSTEDT AG & Co. KG
CELLSTAR® tubes 50 mL, PP, graduated	Greiner Bio-One International GmbH
CELLSTAR® tubes 15 mL, PP, graduated	Greiner Bio-One International GmbH
Test Tubes PS Round 100 x 16mm	Greiner Bio-One International GmbH
DURAN® Laboratory glass bottles	DWK Life Sciences GmbH
Erlenmeyer Flasks	KAVALIERGLASS, a.s.
Measuring cylinders	VWR International, LLC.
Inoculation Loops	SARSTEDT AG & Co. KG
Microscope slide, cut edges, frosted end	Carl Roth GmbH + Co. KG
Colour-fixed indicator sticks	Carl Roth GmbH + Co. KG
pH 1000L, phenomenal	VWR International, LLC.
Microloader Pipette tips, 20 µL	Eppendorf AG

Table 2. Chemicals and Reagents

Reagent	Manufacturer	Art. No.
Formaldehyd	Carl Roth GmbH + Co. KG	CP10.1
Propanol	Carl Roth GmbH + Co. KG	CP41.4
Ethanol Absolute	Scharlab,S.L.	et00051000
PUFFERAN® TRIS	Carl Roth GmbH + Co. KG	AE15.1
ROTIPURAN® ortho-Phosphoric 85 %	Carl Roth GmbH + Co. KG	6366.1
LB Broth (lennox)	Carl Roth GmbH + Co. KG	X964.2
LB broth base (Lennox L Broth Base)	Invitrogen	12780-052
Peptone/Trypsin aus casein	Carl Roth GmbH + Co. KG	6681.1
Peptone	Carl Roth GmbH + Co. KG	2366.1
Glycerol	Carl Roth GmbH + Co. KG	3783.1
Agarose for gel electrophoresis	Carl Roth GmbH + Co. KG	3810.4
Ethidium bromide (1 % lsg in H2O)	Carl Roth GmbH + Co. KG	221.2
Acetic Acid	Carl Roth GmbH + Co. KG	3738.5
EDTA disodium salt dihydrate	Carl Roth GmbH + Co. KG	8043.2
LB-Agar (lennox)	Carl Roth GmbH + Co. KG	X965.2
Agar-Agar, Kobe I	Carl Roth GmbH + Co. KG	5210.5
Saccharose	Carl Roth GmbH + Co. KG	4621.2
100 % apfelsaft aus apfelsaftkonzentrat	Spar	
tryptone	Carl Roth GmbH + Co. KG	8952.3
Sodium chloride	Carl Roth GmbH + Co. KG	3957.2
Yeast extract	Carl Roth GmbH + Co. KG	2363.1
Potassium chloride	Merck KGaA	1.049.361.000
Bleach	DanKlorix	
Silica gel orange	Carl Roth GmbH + Co. KG	T199.4
Oil 10S, Poly(chlortrifluorethylen) 800, VOLTALEF®	VWR International, LLC.	24.627.188
Dry yeast, fermipan red, *Saccharomyces cerevisiae*	Casteggio Lieviti SRL.	
2- Propanol, ROTISOLV®	Carl Roth GmbH + Co. KG	7590.1
Tris-HCl	Carl Roth GmbH + Co. KG	9090.1
Sodium dodecyl sulphate	Carl Roth GmbH + Co. KG	2326.1
Potassium acetat	Carl Roth GmbH + Co. KG	4986.1
Diethyl pyrocarbonate	Carl Roth GmbH + Co. KG	K028.1
Heptane	Carl Roth GmbH + Co. KG	7725.2
Magnesium chloride	Carl Roth GmbH + Co. KG	KK36.1
Nipagin	Thermo Fisher Scientific Inc.	126960025
XbaI	New England Biolabs GmbH	R0145S

NotI - HF®	New England Biolabs GmbH	R3189S
KpnI	New England Biolabs GmbH	R0142S
HindIII	New England Biolabs GmbH	R0104S
CutSmart® Buffer	Solis Biodyne	B7204S
dNTPs	Solis Biodyne	02-21-00400
2.1 Buffer	New England Biolabs GmbH	B7202S
FIREPol® DNA Polymerase	Solis BioDyne OÜ	01-01-0500
10 x Reaction Buffer B	Solis Biodyne	01-01-0500
25 mM MgCl2	Solis Biodyne	01-01-0500
10x Solution S	Solis Biodyne	01-01-0500
100 bp DNA ladder	New England Biolabs GmbH	N3231S
1 kb DNA ladder	New England Biolabs GmbH	N3232L
Gel Loading Dye, Purple (6X)	New England Biolabs GmbH	B7024S
T4 DNA Ligase	New England Biolabs GmbH	M0202S
Q5® High-Fidelity DNA Polymerase	New England Biolabs GmbH	M0491S

Table 3. Kits

Kit	Manufacturer	Art. No.
QIAquick PCR purification kit	QIAGEN GmbH	28104
GeneJet Plasmid Miniprep Kit	Thermo Fisher Scientific Inc.	K0502
Plasmid midi kit (100)	QIAGEN GmbH	12145
GeneJET Gel Extraction Kit	Thermo Fisher Scientific Inc.	K0692
NEBuilder® HiFi DNA Assembly Cloning Kit	New England Biolabs GmbH	E5520S

Table 4. Solutions

Solution	Preparation
SOB	950 mL milliQ + 20 g tryptone + 5 g yeast extract + 0.5 g NaCl
SOC	200 μL Glucose + 50 μL MgCl2 + 10mL SOB medium
Nipogin mix	2.25 g Nipagin + 15 mL Ethanol absolute
Apple Juice Agar	13 g Agar Agar + 8 g Saccharose + 250 mL milliQ + 83 mL apple juice + 3 mL Nipogin mix
TAE Stock Solution 50x	242 g Tris base + 15.5 mL Phosphoric Acid + 100 mL of 0.5 M EDTA (pH=8)
TAe Solution 1X	100 mL TAE Stock Solution 50x + 4900 mL deionized H2O

Table 5. Biological and genetic material

Organism	Supplier/Obtained from	Art. No. / Reference
NEB® Stable Competent *E. coli* (High Efficiency)	New England Biolabs GmbH	C2987I
Invitrogen™One Shot™ TOP10 Chemically Competent *E. coli*	Thermo Fisher Scientific Inc.	C404003
Invitrogen™Subcloning Efficiency™ DH5α Competent *E.coli*	Thermo Fisher Scientific Inc.	18265017
Dsim #2176 (*y*[1] *w*[1]; pBac{3XP3::EYFP-,attP}) (*Drosophila simulans*)	David Stern	Stern *et al.* 2017
Bloomington #32218 (*w*[*]; P{*y*[+t7.7] *w*[+mC]=10XUAS-IVS-mCD8::RFP}attP2) (*Drosophila melanogaster)*	Bloomington *Drosophila* Stock Center	
wOR(P) gypsy-lacZ (*Drosophila melanogaster*)		Sarot *et al.* 2004
w^{1118} (*Drosophila melanogaster*)	Vienna *Drosophila* Resource Center	
Plasmids	**Obtained from**	**Reference**
pBlueRabbit (pBR)	Ben E. Housden	Housden *et al.* 2012
vas-φ-C31(3xP3-EGFP)	Frank Schnorrer	Zhang *et al.* 2014

1. Institution

All experiments were performed at the Institute for Population Genetics, Department of Biomedical Sciences of the University of Veterinary Medicine Vienna, Veterinärplatz 1, 1210 Vienna.

2. Primer design & supplier

We designed all the primers for the smaller fragments (1S, 2S, 1A, 2A) using the primer basic local alignment search tool (BLAST) from the NCBI website (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) with default settings, except for the

organism parameter in the primer pair specificity parameter section (changed to *Drosophila simulans*) and for the primer melting temperatures (Min=60°C, Opt=63°C, Max=67°C).

I used the NEBuilder v.2.2.5 from the *New England Biolabs Inc.* website (https://nebuilder.neb.com/) to design the primers for the larger fragments (3S-1, 3S-2, 3N-1, 3N-2) for which we have performed Gibson cloning. All DNA oligos have been ordered from Sigma Aldrich, desalted and dry, storage and handling were performed according to the manufacturer.

Table 6. Primers

Primer	Sequence (5' to 3')
LS1	AGCGGCCGCAGCAGGTGTCAACAGGTCGC
LS2	ATCTAGAGCCCATTCGGGCCACGATAA
LS3	ATCTAGA CCGCCAGTCTGCGAAACTCA
LS4	CCGGATCCCCCGGTACCCGCAGCAGGTGTCAACAGGTC
LS5	GCCCATTCGGGCCACGAT
LS6	TTATCGTGGCCCGAATGG
LS7	CTTGGCTGCAGGTCGCGACTCAGTGAGCCCATCAAGGC
EnhColonyFor	TCGCGCACGTTTCTTATTGCG
EnhColonyRev	AACGCTGGCGACTTCTTGGG
lacZ-For	GATACACTTGCTGATGCGGTGCTGATT
lacZ-Rev	CTGTAGCGGCTGATGTTGAACTGGAAG
lacZ - seq	GTTCAATGATGTCCAGTGCAG
attP-For	CCCAGGTCAGAAGCGGTTTTCG
attP-Rev	TACGTGTCCACCCCGGTCACAA

3. PCR for *GD13851* haplotype-block amplification

I performed PCR from isolated genomic DNA of single flies from either the isofemale line 42 (evolved haplotype) or from the isofemale line 46 (ancestral haplotype), using the Q5 High-Fidelity DNA Polymerase (New England Biolabs GmbH) from New England Biolabs. I received the samples from Lauri Torma, who performed the DNA extraction himself from single flies of the isofemale line 42 ("evolved" haplotype-block) and from the isofemale line 46 (ancestral haplotype-block). The final concentration of fly genomic DNA in the PCR reaction was adjusted in between the concentration level that is recommended for PCR from

yeast (0.1µg/mL) and mammalian (10 µg/mL) genomic template DNA (according to Sambrook & Russell 2001: 8.20). I performed PCR according to the manufacturer's protocol (M0491, NEB, https://international.neb.com/protocols/2013/12/13/pcr-using-q5-high-fidelity-dna-polymerase-m0491, 20.11.2019).

Table 7. PCR from genomic DNA for haplotype amplification

	Iterations	Temperature	Duration
Initial Denaturation	1 time	98°C	5 min
Amplification	35 cycles	98°C	10 sec
		66°C / 70°C	30 sec
		72°C	30 sec per 1 kb + 15 sec
Final extension	1 time	72°C	2 min
Hold		24°C	infinite

Table 8. Amplicons & PCR components

Amplicon	PCR Components (DNA template + primers)	Amplicon size [bp]
1S	DNA isolated from single fly from isofemale line 42 + LS1 +LS2	2667
2S	DNA isolated from single fly from isofemale line 42 + LS1 +LS3	4240
3S-1	DNA isolated from single fly from isofemale line 42 + LS4 +LS5	2667
3S-2	DNA isolated from single fly from isofemale line 42 + LS6 +LS7	3056
1A	DNA isolated from single fly from isofemale line 46 + LS1 +LS2	2667
2A	DNA isolated from single fly from isofemale line 46 + LS1 +LS3	4240
3A-1	DNA isolated from single fly from isofemale line 46 + LS4 +LS5	2667
3A-2	DNA isolated from single fly from isofemale line 46 + LS6 +LS7	3056
Colony PCR	Transformed *E.coli* + EnhColonyFor + EnhColonyRev	288
Screening for lacZ	Pooled DNA of *Dsim*#2176 F1 flies (parents microinjected) + lacZ-For + lacZ-Rev	407

Screening for attP	Pooled DNA of Dsim#2176 flies + attP-For + attP-Rev	120

4. Isolation of Amplicons

Isolation of PCR amplicons was performed using the Qiagen, QIAquick PCR purification kit, and according to the manufacturer's protocol (QIAquick PCR Purification Kit and QIAquick PCR & Gel Cleanup Kit Quick-Start Protocol, https://www.qiagen.com/lu/resources/resourcedetail?id=e0fab087-ea52-4c16-b79f-c224bf760c39&lang=en, 20.11.2019).

5. Agarose-Gel electrophoresis

Agarose content of the agarose gel was individually adjusted to the length of the DNA fragments that we aimed to separate, and varied between 0.7 % and 1.4 % (Sambrook & Russell 2001: 5.5). The DNA samples were run by applying a voltage of 1-5 V per cm distance between the positive and negative electrode, to the gel (Sambrook & Russell 2001: 5.13). Ethidium Bromide was incorporated into the agarose gel at a concentration of 0.5 µg / mL (Sambrook & Russell 2001: 5.14-5.15), for detection of single-strand and double-strand nucleic acid in the gel.

6. Restriction digest

All restriction enzymes were purchased from NEB and handling, storage and restriction digest was performed following NEB's recommendations (Optimizing Restriction Endonuclease Reactions, https://international.neb.com/protocols/2012/12/07/optimizing-restriction-endonuclease-reactions, 20.11.2019). Buffers compatible with the restriction enzymes were chosen according to NEB's recommendations (NEBuffer Activity/Performance Chart with Restriction Enzymes, https://international.neb.com/tools-and-resources/usage-guidelines/nebuffer-performance-chart-with-restriction-enzymes, 20.11.2019).

7. Restriction Ligation

Ligation was performed using T4 DNA Ligase (New England Biolabs GmbH) according to the manufacturer's protocol (M0202 from NEB's website: https://international.neb.com/protocols/0001/01/01/dna-ligation-with-t4-dna-ligase-m0202, 20.11.2019). Before ligation, the plasmid pBR and all the amplicons that were to be ligated (1S, 2S, 1A, 2A) got digested with XbaI and NotI. After the ligation, I performed an additional purification with Qiagen, QIAquick PCR purification kit, again according to the manufacturer's protocol. This is a feasible option because the fragments generated by restriction digestion are approximately 10-15 bp in length, and the QIAquick PCR purification kit isolates only DNA fragments with a length of 100 bp to 10 kb (https://www.qiagen.com/lu/products/discovery-and-translational-research/dna-rna-purification/dna-purification/dna-clean-up/qiaquick-pcr-purification-kit/#technicalspecification, 20.11.2019)

8. Gibson Assembly

Gibson Assembly was performed in order to ligate the subfragments (3S-1 & 3S-2 and 3A1 & 3A-2) that compound my largest enhancer fragments (3S and 3A), with each other and with the vector backbone. We used the NEBuilder® HiFi DNA Assembly Cloning Kit and followed the manufacturer's protocol (NEBuilder HiFi DNA Assembly Reaction Protocol from NEB's website: https://international.neb.com/protocols/2014/11/26/nebuilder-hifi-dna-assembly-reaction-protocol, 20.11.2019). Before assembly, the vector pBR has been digested with XbaI and NotI. This renders the vector into a linear conformation, making pBR a suited target for the exonuclease used in Gibson Assembly. The amplicons used in this protocol (3S-1, 3S-2, 3A-1, 3A-2) are designed in a fashion to create an overlap in the multiple cloning site (MCS) of pBR after digestion. After the ligation, I again performed an additional purification with Qiagen, QIAquick PCR purification kit, according to the manufacturer's protocol.

Table 9. Enhancer-reporter constructs ligation components and protocol

Plasmid	Ligation components	Protocol
p1S	1S + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p2S	2S + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p3S	3S-1 + 3S-2 + pBR	NEBuilder HiFi DNA Assembly Reaction Protocol
p1A	1A + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p2A	2A + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p3A	3A-1 + 3A-2 + pBR	NEBuilder HiFi DNA Assembly Reaction Protocol

9. Transformation

Transformations using either NEB® Stable Competent *Escherichia coli* (*E. coli*) (High Efficiency), Invitrogen™ One Shot™ TOP10 Chemically Competent *E. coli*, or Invitrogen™ Subcloning Efficiency™ DH5α Competent Cells were performed, respectively according to the manufacturer's recommendations, listed below:

1)   Protocol: High Efficiency Transformation for NEB® Stable Competent *E. coli* (C3040I,   C3040H), (https://international.neb.com/protocols/2014/08/12/high-efficiency-transformation-protocol-c3040i, 20.11.2019), but SOC medium used instead of NEB 10-beta/Stable Outgrowth Medium

2)   Protocol: USER   GUIDE   -   One   Shot®   TOP10   Competent   Cells, (https://www.thermofisher.com/order/catalog/product/C404003#/C404003, 20.11.2019)

3)   Protocol: Subcloning   Efficiency   DH5alpha   Chemically   Competent   *E.   coli* (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/subcloningefficiencydh5alpha_man.pdf)

After gel-extraction of p1S, p1A, p2S, and p3S we transformed NEB® Stable Competent *E. coli* with the gel-extracted plasmids, this time following the Protocol for cloning DNA containing   repeat   elements   (C3040,   NEB) (https://international.neb.com/protocols/2013/10/30/protocol-for-cloning-dna-containing-repe

, 20.11.2019). In this protocol, the overnight incubation temperature is decreased (from 37°C to 30°C) with the intention to reduce the rate of recombination events.

10. Selective agar plates and medium for transformed *E. coli*

The agar plates (20 g LB Agar (Lennox) per 1 L milliQ) and the medium (20 g LB Broth (Lennox) per 1 L milliQ) used for cultivation and selection of transformed *E. coli* contained the broad-spectrum antibiotic Kanamycin in a concentration of 50 µg/mL as advised by Sambrook & Russell (2001: A2.6). The antibiotic was added after autoclaving and after the media decreased in temperature, to preserve the integrity of thermolabile antibiotics (Sambrook & Russell 2001: A2.5), even though kanamycin is stated to be autoclavable by Carl Roth (Carl Roth: Produkt Datenblatt LB-Agars mit Antibiotika. Available from: https://www.carlroth.com/downloads/ba/en/8/BA_8861_EN.pdf). For guidelines on agar plate preparation see Sambrook & Russell, 2001 (A2.5).

11. *E. coli* Overnight cultivation

Overnight cultivation of transformed *E. coli* cultures was performed in selective medium (kanamycin, LB broth, 3 mL or 50 mL cultures) for 12 hours at 37° C in a shaking incubator (L13701 GFL).

12. Colony PCR

For the colony PCR, we took agar plates with transformed *E. coli* that have been incubated overnight. I gently touched individual colonies with the tip of a plastic pipette tip (Pipette tip 20µL, SARSTEDT AG & Co. KG), and stuck the plastic pipette tip in a single tube of a multiply® - µStrip 0.2 mL chain, qPCR strip (SARSTEDT AG & Co. KG), filled with milliQ so that the tip is submerged. The pipette tip remained in the tube of the multiply® - µStrip 0.2 mL chain, qPCR strip, while I picked additional colonies. As soon as I picked as many colonies as needed (from 40 - 80), I collectively mixed the pipette tips in their respective qPCR tubes for 5 sec. Then I used the same pipette tip, still stuck in the qPCR tube, to transfer 2 µL of the 30µL colony-milliQ dilution to the PCR mix.

I used the FIREPol® DNA (Taq-) Polymerase from Solis BioDyne and the primers "ColonyFor" and "ColonyRev" for identification of colonies that harbor plasmids with the correct inserts.

I proceeded with the further colony PCR protocol according to the manufacturer's recommendations

(https://www.solisbiodyne.com/pics/7511_Data_Sheet_FIREPol_DNA_Polymerase.pdf, 20.11.2019).

Table 10. Colony PCR

Colony PCR		temperature	time
Initial Denaturation	1 time	95°C	10 min
Amplification	25 cycles	95°C	30 sec
		56°C	30 sec
		72°C	30 sec
Final extension	1 time	72°C	1 min
Hold		24°C	infinite

13. Purification of plasmid DNA from overnight cultures

Isolation of plasmid DNA from 3 mL LB broth overnight cultures was performed using the GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific Inc.) according to the manufacturer's protocol (PRODUCT INFORMATION Thermo Scientific GeneJET Plasmid Miniprep Kit #K0502, #K0503, from Thermo Fisher Scientific's website: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0012655_GeneJET_Plasmid_ Miniprep_UG.pdf, 20.11.2019)

Isolation of plasmid DNA from 50 mL LB broth overnight cultures was performed using the Plasmid Midi Kit (100) (Qiagen GmbH), according to the manufacturer's protocol (Quick-Start Protocol QIAGEN ® Plasmid Mini, Midi, and Maxi Kits March 2016, from Qiagen's website:

https://www.qiagen.com/be/resources/resourcedetail?id=c164c4ce-3d6a-4d18-91c4-f5763b6d4283&lang=en, 20.11.2019).

14. Gel extraction

I performed gel extraction using the GeneJet Gel Extraction Kit (Thermo Fisher Scientific Inc.) according to the manufacturer's instructions (PRODUCT INFORMATION: Thermo Scientific GeneJET Gel Extraction Kit#K0691, #K0692; from Thermo Fisher Scientific's website:

https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0012661_GeneJET_Gel_Extraction_UG.pdf, 20.11.2019)

15. Isopropanol precipitation

Was performed according to Sambrook and Russell (2001: 6.30) or according to Qiagen's protocol (Quick-Start Protocol QIAGEN ® Plasmid Mini, Midi, and Maxi Kits March 2016, from Qiagen's website:
https://www.qiagen.com/be/resources/resourcedetail?id=c164c4ce-3d6a-4d18-91c4-f5763b6d4283&lang=en, 20.11.2019)

16. Sequencing

To validate the integrity of all enhancer-reporter constructs Sanger sequencing was performed by LGC Genomics GmbH, Ostendstrasse 25, 12459 Berlin, Germany. Isolated plasmid content from transfected *Escherichia coli* cultures was sequenced with the primer pair "lacZ-seq" and "EnhColonyRev". DNA sample preparation was performed according to the company guidelines (Online ordering guide, from LGC's website: https://shop.lgcgenomics.com/documents/Flyer_Online_ordering_sequencing_LGC_Genomics.pdf, 20.11.2019)

17. Storage of plasmids, DNA, primers

All the genetic material and enzymes were stored at -20°C, competent cells at -80°C.

18. *Drosophila* maintenance

Drosophila simulans flies were maintained at 25°C (humidity 75 %) if not stated otherwise. The storage unit was the incubator model dr-36vl (clf plant climatics). All anesthesia of *Drosophila* during the study was conducted by CO_2. The bottles and vials used for adult *Drosophila* maintenance contained either one of two standard *Drosophila* media, one was used for adult maintenance, one for the maintenance of microinjected larvae (softer standard food variant).

19. Apple Agar plates

Apple agar plates were used for the embryo collection cage as they are adequate for large and rapid collections of embryos (Kiehart *et al*. 2000). We mixed and autoclaved 13 g Agar-Agar, 8 g Saccharose, and 250 mL milliQ. Afterwards, 83 mL hot apple juice and 3 mL nipagin were added, the medium was mixed thoroughly and always 30 - 35 mL poured into 90 mm Petri dishes. After solidification, the Petri dishes were stored inverted at 4°C.

20. Injection mixes

Multiple individual *Escherichia coli* colonies that were validated (restriction digest) to possess proper donor plasmids have been used to inoculate 50 mL liquid overnight cultures. Plasmid purification from these 50 mL cultures was executed using Plasmid Midi Kits (100) (Qiagen GmbH) according to the manufacturer's recommendation (Quick-Start Protocol QIAGEN ® Plasmid Mini, Midi, and Maxi Kits March 2016), isolated DNA was dissolved in milliQ.

The integrity of the enhancer-reporter construct, including the orientation of the enhancer fragment, was further validated by restriction digest (KpnI, HindIII) and DNA sequencing with the primers "lacZ-seq" and "EnhColonyRev" (LGC Genomics).

Injection mixes were prepared by the addition of the helper plasmid p.vas-ϕC31(3xP3-EGFP) (Zhang *et al*. 2014) (dissolved in milliQ), to each donor plasmid (p1S, p2S, p3S, p1A, p2A, p3A), in a suited concentration ($c_{helperPlasmid}$= 190-250 µg/mL, $c_{donorPlasmid}$= 250 µg/mL).

DNA concentration of donor and helper plasmid has been observed before not to be much of a critical factor for the success of genomic integration in *Drosophila* embryo microinjections. With concentrations between 190 and 250 µg/mL, we are working in a range that has been

observed to yield high transformation efficiencies before (Gompel & Schröder 2015, Bischof *et al*. 2007, Zhang *et al*. 2014, Spradling & Rubin 1982).

Furthermore, the injection mixes were centrifuged for 15 min (4°C, 15 000 rpm) and the most upper phase of liquid was withdrawn and used for injections. This aims to sediment and hence exclude any matter that otherwise may clog the injection needle (Kiehart *et al*. 2000, Spradling & Rubin 1982).

21. Heptane glue

Heptane glue was prepared by adding 5 m one-sided tape to 200 mL heptane and letting it spin overnight on the test-tube rotator.

22. *Drosophila* embryo microinjections

Techniques and protocols for are discussed in Spradling & Rubin (Science 1982), Kiehart *et al*. (*Drosophila* Protocols 2000), Gompel & Schröder (*Drosophila* Germline Transformation 2015, from the Nicolas Gompel's lab website (http://gompel.org/methods): http://gompel.org/wp-content/uploads/2015/12/Drosophila-transformation-with-chorion.pdf, 21.11.2019), and Fish *et al*. (Nature Protocols 2007). A brief description follows.

200 - 300 flies were transferred to an egg-lay cage, 3-4 days prior to injections, to let the flies accustom to their new environment (Gompel & Schröder 2015). Apple-agar plates with little yeast paste (dry yeast, fermipan red) spread in the middle, were used to promote oviposition of *Drosophila* flies (Becher *et al*. 2012).These plates were changed twice a day in this acclimation phase to keep the flies well fed and also conditioned to getting tapped to the ground (Kiehart *et al*. 2000). For injections, a harvest of ~150 embryos each 30 min is required.

On the Injection day, I changed the apple-agar plates plus yeast at least 3 times before starting the first round of injections, to assure harvesting embryos that are not older than 30 min. I loaded 2-3 needles before starting the injections, to account for any breakage or clogging of needles, which happens usually. The loading was performed with microloader (Eppendorf) tips, by holding the injection needles in a 45° angle (to the ground), and by letting an injection

mix droplet run to the tip, thereby avoiding any bubble formation. The loaded injection needles can be stored up to 4 days at 4°C (Nicolas Gompel 2015: http://gompel.org/wp-content/uploads/2015/12/Drosophila-transformation-with-chorion.pdf, 05.12.2019). Then the first needle was mounted on the micromanipulator, fixed and sealed in a capillary holder with grip head, to which a silicon tube was connected. The other end of the silicon tube was connected to a Femtojet 5247 (Eppendorf) microinjector, that was controlled by a computer mouse. The microinjector was adjusted to apply an injection pressure (p_i) of 658 hectopascal (hPa) to the injection needle in the capillary holder whenever the right mouse button was pressed.

The Injection needle was fixed in an approximately 15° angle to the microscope stage. Multiple layers of heptane glue had to be arranged in 15 min intervals, ensuring a heptane glue layer depth of about 5 layers on each microscope slide.

Each 30 min I harvested embryos by gently tapping the flies to the bottom of the egg-laying cage and replacing the apple-agar plate with yeast. The egg-laying cage was transferred back to at 25°C until the next harvest, while the rest of the following procedure was performed at ~19°C. To my observations, *Drosophila simulans* deposits its embryos preferentially in the yeast paste. I took the harvested apple-agar plate and put 15 mL 60 % bleach (danKlorix) directly into the plate. I used a brush to swirl the yeast paste into the bleach until the yeast dissolved completely. Then I stopped using the brush and swayed the petri dish gently instead. This whole step took exactly 2:30 min and aimed to de-chorionate the embryos. Subsequently, the bleach is decanted into a fine mesh filter, the embryos remain in the filter and are washed with an H_2O squeeze bottle (milliQ) for approximately 30 seconds. The washed embryos (now without chorion) are transferred with a brush to a fresh apple-agar plate, that has been cut in the diameter. Then I aligned the embryos right at the straight edge of the agar plate with a metal pick. The embryos were adjusted into the same dorsal/ventral and posterior/anterior orientation, to minimize the need to refocus during the injections. Thereafter the line of embryos was fixed onto a microscope slide with heptane glue layers on it, by slowly lowering the slide in a slight angle (~10° to the agar plate) onto the embryos, until they get gently pushed into the apple-agar plate. Upon lifting of the slide, all embryos stuck to the slide. If not, it is possible to set up a second line behind the first one with a

second try. Next, the slide with the embryos was transferred into a small Tupperware box containing dry silica gel and remained there for 6 min. The desiccation step aims to relieve the turgor pressure that is released when the injection needle penetrates the vitelline membrane and consequently minimizes cytoplasmic leakage (Spradling & Rubin 1982). This step is crucial to the survival of injected embryos (Fish *et al.* 2007). After appropriate desiccation, the line of embryos got covered under a thin layer of halocarbon oil 10S (VWR chemicals), to 1) stop further desiccation, 2) allow for oxygen exchange (Fish *et al.* 2007).

The microscope slide with the immobilized, dechorionated, and desiccated embryos was put onto the microscope stage. For penetration with the injection needle, the embryos were moved with their posterior ends against the fixed microinjection needle tip. These gentle and plain movements were performed only by moving the microscope stage. The needle tip was inserted into the posterior end and then withdrawn back to the posterior pole, as far as possible while still remaining in the embryo. This aims to maximize the spatial overlap of nuclei that are destined to become germ cells and the exogenous plasmid DNA. Regulated via the right mouse button, an amount of injection mix, appropriate to the desiccation level was injected via the microinjector Femtojet 5247 (injection pressure = 658 hPa). This resulted in the injection of approximately 40 pL injection mix into the embryos (volume ~ 2 nL) (Spradling & Rubin 1982), the entering of solution was visible. Afterward, the needle was quickly withdrawn to minimize any damage to the embryo. No or very little fluid leakage should appear from the injected embryos. All uninjected, damaged or improperly aged embryos got removed, to circumvent the crossing of uninjected embryos.

This whole procedure took exactly 30 min, so any embryo was exactly 60 min old when he got injected, of those 60 min each embryo developed 30 min at 25°C and 30 min at ~19°C. This means that embryos were approximately at nuclear cycle 7-8 when they got injected. In these embryonic stage cellularization did not occur yet and the embryo is present as a syncytial blastoderm instead. The nuclei are still concentrated in the interior of the embryo in cycle 7. In telophase of nuclear cycle 8, the majority of nuclei migrate to the surface of the embryo. In early cycle 9, chromosomes of the primordial germ cells reach the posterior pole surface, while the somatic chromosomes reach the cortex at ~ nuclear cycle 10. In nuclear cycle 11 the chromosomes of the future germ cells form separate pole cells, cellularization finally occurs in nuclear cycle 14 (Raff & Glover 1989). To reach the maximum integration

rate into chromosomes of prospective germ cells, I injected the exogenous plasmid DNA in cycle 7-8, before the chromosomes reached the posterior pole.

After Injections embryos remained glued to the microscope slide under oil, in humid conditions (in 90 mm Petri dishes with wet paper towel fixed onto the top) at 19°C for 48 hours. As soon as 1st instar larvae hatched, these were transferred from the microscope slides to food vials by "rolling" them with a metal pick in the oil. Large groups of larvae were gathered in the oil and formed to little spheres consisting of oil and larvae by gently rolling this little sphere across the halocarbon oil "puddle". These oil-larvae spheres were then transferred to standard food vials. The transfer was performed like this to decrease the risk of physical harm to any single larvae to a minimum. Food vials containing rescued larvae were thereafter transferred back to 25°C. Between 40 - 80 larvae were maintained together in a single vial, to allow for collective food processing (termed "social digestion", Louis & Polavieja 2017). ~3 days after injection, filter paper was added to the food vials to supply the *D. simulans* flies with an adequate location to pupate, to avoid any loss of injected G0 flies due to burial in agar. As soon as adult flies hatched, I separated them and backcrossed them to males and virgins of their original strain (*Dsim* #2176). I always paired one male "survivor" with 5 female virgins, and 5 female "survivors" with 5 fresh males. I flipped those vials then every 4 days for 4 times in total. If germline integration took place in some flies then I should obtain red-eyed (w^+) F1 flies from these crosses.

23. PCR from pooled flies of F1 generation and control strains

The isolation of genomic DNA from the pooled F1 flies (*Dsim*#2176 flies of whom one parent has been microinjected) and from pooled flies of control strains (w^{1118}, Bloomington #32218, *wOR*(*P*) *gypsy-lacZ,* and uninjected *Dsim*#2176) has been performed according to the protocol: "Rapid small scale isolation of Drosophila DNA and RNA" by Hermann Steller, from the Rubin Lab Manual (1990). The final concentration of fly genomic DNA in the PCR reaction (approx. 0.4 µg/mL) was adjusted in between the concentration level that is recommended for PCR from yeast (0.1µg/mL) and mammalian (10 µg/mL) genomic template DNA (according to Sambrook & Russell 2001: 8.20). I used the FIREPol® DNA (Taq-) Polymerase from Solis BioDyne and the primer pairs "lacZ-For", "lacZ-Rev" and "attP-For", "attP-Rev" for the amplification of any *lacZ* gene and attP site that may be present in the flies.

I executed the PCR according to the manufacturer's protocol (https://www.solisbiodyne.com/pics/7511_Data_Sheet_FIREPol_DNA_Polymerase.pdf, 20.11.2019).

Table 11. PCR from pooled genomic DNA

	Iterations	Temperature	Duration
Initial Denaturation	1 time	98° C	10 min
Amplification	38 cycles	95° C	30 sec
		62° C	30 sec
		72° C	45 sec
Final extension	1 time	72° C	7 min
Hold		24°C	infinite

24. Basic Local Alignment Search Tool (BLAST)

To identify a putative ortholog of *vasa* in *Drosophila simulans* the tool Standard Protein BLAST (NCBI Resource Coordinators 2018, Altschul *et al*. 1990) from NCBI was used (blastp, default settings. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi). The amino acid sequence of the Vasa isoform vas-PB was used as input and searched against the *D. simulans* genome (sequence information from FB2019_05, Thurmond *et al*. 2019).

To access the degree of conservation of the non-coding regions upstream of the TSS of *vasa* and *GD23992* the tool BLAST from Flybase was used (FB2019_05, Thurmond *et al*. 2019). The *vasa* promoter and 5'UTR sequence present in the plasmid vas-φ-C31(3xP3-EGFP) was used as query sequence, and was searched against both, the *D. melanogaster* (Dmel, r=6.31) and the *D. simulans* (Dsim, r=2.02) genome. The low complexity filter was turned off, otherwise default settings were used.

Results

1. Construction of enhancer-reporter vectors

I received two genomic DNA samples from Lauri Torma, that originated from two *Drosophila simulans* isofemale lines (one from isofemale line 42, one from isofemale line 46). These isofemale lines are two out of the 1278 isofemale lines that were used in total to initiate the South African E&R experimental cages that were analyzed later by Anna Maria Langmüller. The "evolved" haplotype-block of the *GD13851* locus (see Figure 5 B) was isolated from an individual fly from isofemale line 42. This haplotype-block that has been shown to rise in frequency within the population as a response to a stressful hot-environmental regime (referred to as "evolved" or "selected" (putative) enhancer version or haplotype-block in the following) (see Figure 1). The "ancestral" haplotype-block of the *GD13851* locus (see Figure 5 A) was isolated from an individual fly from isofemale line 46. The SNPs that characterize this haplotype-block were present in high frequency in the base population, before the population was subjected to the shift in selection regime (referred to as "ancestral" (putative) enhancer version or haplotype-block in the following).

We amplified 3 sequence fragments from both (putative) enhancer versions by polymerase chain reaction (PCR). These amplicons differ in length, do overlap, and are located in the intronic region between the first and the third exon, the second exon is included in all amplicons. These fragments span a length of about 2.6 kb, 4.2 kb, and 5.7 kb, respectively, and start from a sequence base (2.6 kb fragment) and increase in size in steps from that base (see Figure 5). This overlapping design was chosen to catch cooperative and additive interactions between TFs, which bind to distant motifs (Long *et al*. 2016, Small *et al*. 1992), that are potentially existing and crucial for regulatory activity of this region. So in total, we generated 6 different fragments, three from the "selected" haplotype-block (harboring the evolved version of the candidate SNPs, these fragments are referred to as 1S, 2S, 3S, S = "selected", see Figure 5B & 6B & 6C) and three from the "ancestral" haplotype-block (with the ancestral version of the candidate SNPs, fragments referred to as A1, A2, A3, A =

"ancestral", see Figure 5A & 6A). I divided the large 5,7 kb fragment into two amplicons (3S-1 & 3S-2 and 3A-1 & 3A-2, ~2.7 and ~3 kb respectively) that were further ligated by Gibson assembly (see Figure 5).
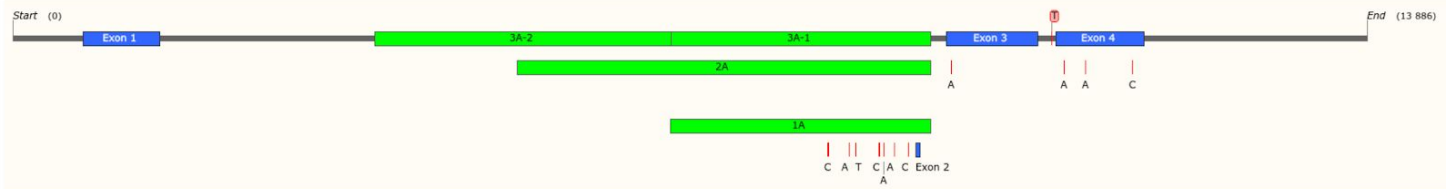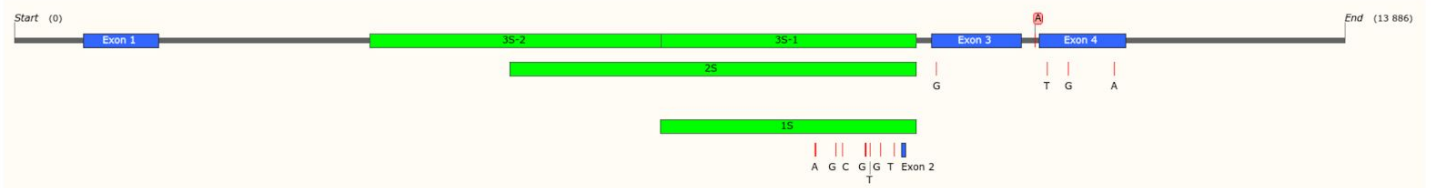

Figure 5 Amplification of *GD13851* intronic regions: (A) The ancestral haplotype-block of *GD13851* and (B) the evolved haplotype-block of *GD13851*. The exons are depicted in blue, the introns between the exons as black line, the position and the variant of the candidate SNPs are marked in red. All amplicons (in A: 1A, 2A, 3A-1, 3A-2 and in B: 1S, 2S, 3S-1, 3S-2) are shown in green. All of them contain the predicted enhancer region, and also the small exon number 2. The largest enhancer fragments (3A and 3S) were subdivided into two amplicons (3A-1, 3A-2 and 3S-1 and 3S-2) were ligated via Gibson assembly.

Figure 6 Electrophoresis of the amplicons from either haplotype-block: Agarose gel electrophoresis of amplicons amplified by PCR from genomic DNA of single flies from (A) isofemale line 46 (ancestral) or (B, C) isofemale line 42 (evolved). (B,C) The PCR from the evolved haplotype-block was performed twice because at first (B) only the fragment 3S-1 worked. At the second try (C) the remaining fragments were amplified (for 1S-PCR an aliquot of the first failed round of injection was used as input instead of genomic DNA). PCR and gel electrophoresis executed as described in the methods.

2. Insertion into pBR

Hence all sequence fragments from both haplotype-blocks were inserted into the vector pBlueRabbit that has been shown to be appropriate for an enhancer-reporter assay.

The two shorter fragments from both haplotypes (1S, 2S, 1A, 2A) were ligated into the MCS of pRB by restriction cloning. We used the restriction enzymes XbaI and NotI from NEB. Therefore we added XbaI and NotI recognition sites to the 5' ends of primer pairs. The multiple cloning site (MSC) is located adjacent to the 5' end of the *Hsp70* promoter. The putative enhancer end that is located close to exon 3 (at the original genomic locus) is immediately joined to the *Hsp70* minimal promoter, while the end of the sequence that is closer to the exon 1 is closer to *mini-white* gene (see Figure 7 and Figure 8).
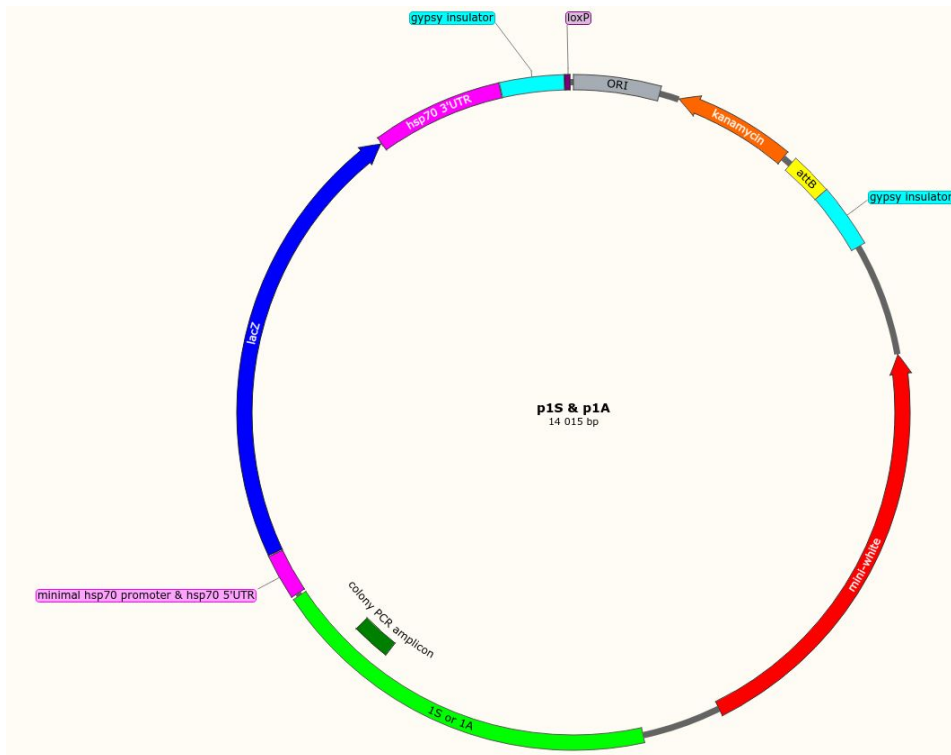

Figure 7 The plasmid p1S or p1A: The finished donor plasmid p1S or p1A after restriction cloning with the sequence fragment 1S or 1A (indicated in green) and pBR, respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.
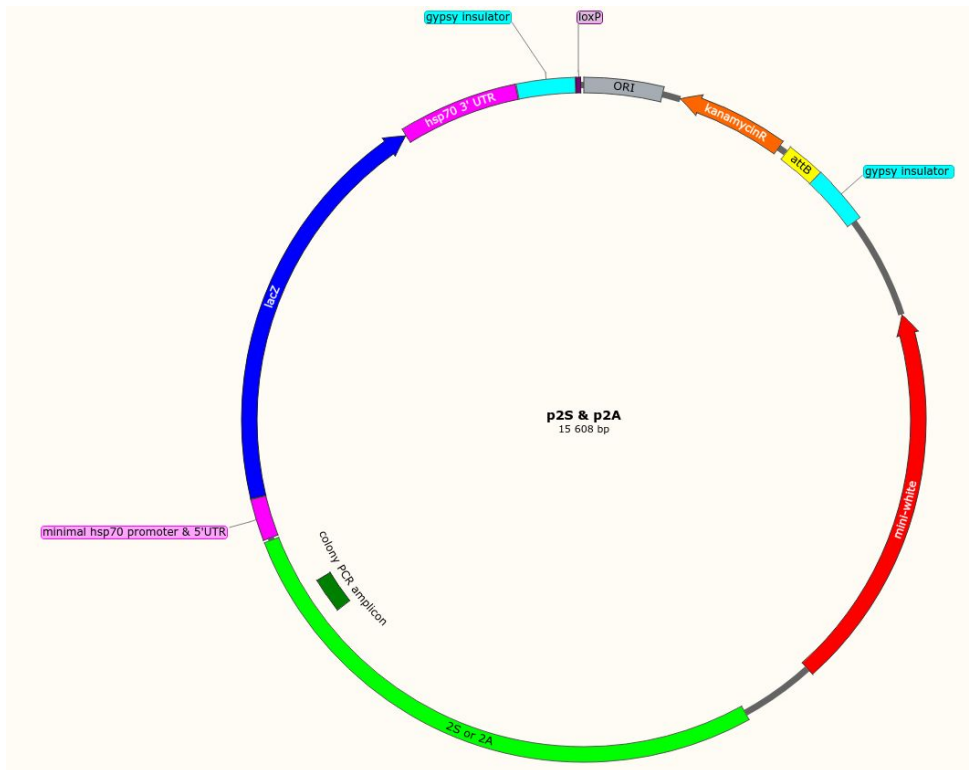
Figure 8 The plasmid p2S or p2A: The finished donor plasmid p2S or p2A after restriction cloning with the sequence fragment 2S or 2A (indicated in green), respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.

Even though New England Biolabs (NEB) states on their website that the Q5 High-Fidelity DNA Polymerase is capable of amplifying DNA fragments up to 10 kb from complex genomic DNA, we chose to split the largest fragments (3S and 3N) that span a length of almost 6 kb into two amplicons (~2,7 kb and ~3 kb), to preclude any complications beforehand. These two amplicons exhibit terminal DNA sequence overhangs with each other and with the sequence of the MCS of pBR. The Gibson overlaps were designed using the NEBuilder v.2.2.5 (New England Biolabs Inc.).

These two amplicons and the plasmid were consequently ligated using a 5' exonuclease, a polymerase, and a ligase, this methodology is known as Gibson assembly (Gibson *et al.* 2009). The sequence region that is close to exon 3 is again immediately joined to the minimal *Hsp70* promoter.
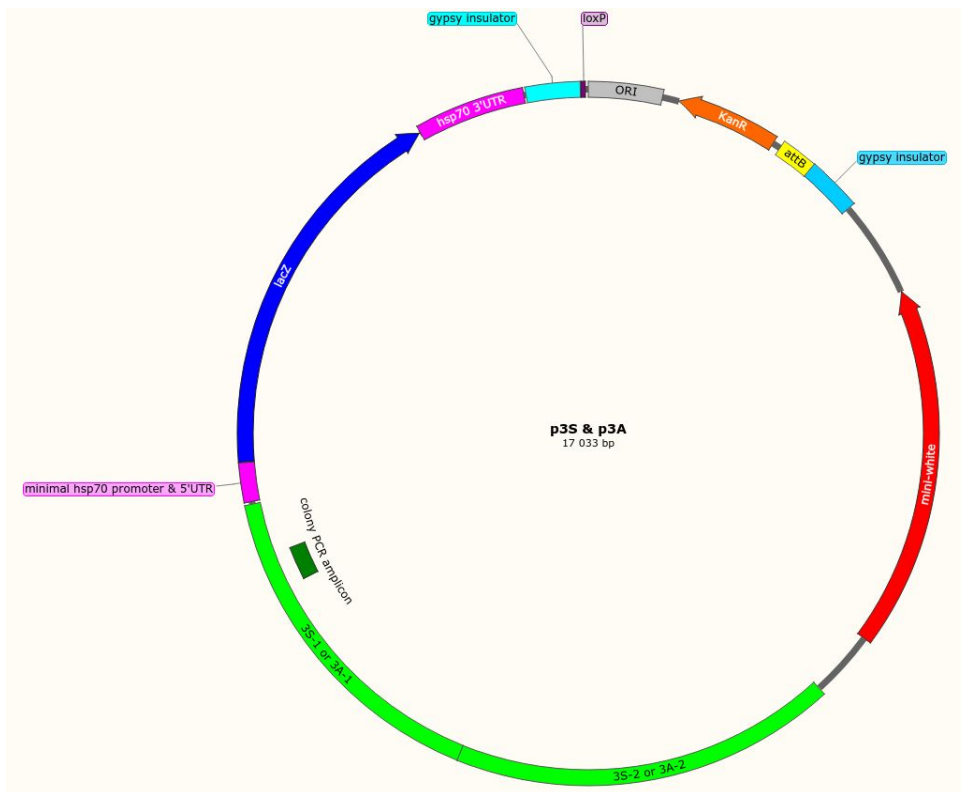
Figure 9 The plasmid p3S or p3A: The finished donor plasmid p3S or p3A after Gibson assembly with the sequence fragments 3S-1 and 3S-2 or 3A-1 and 3A-2 (indicated in green), respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.

The enhancer-reporter vectors were called p1S (containing fragment 1S), p2S (containing fragment 2S), p3S (containing fragment 3S), p1A (containing fragment 1A), p2A (containing fragment 2A), p3A (containing fragment 3A).

3. Amplification of the enhancer-reporter vectors

3.1 Transformation of TOP10 and DH5α competent cells

Replication of the final enhancer-reporter vectors was executed via chemical transformation of Invitrogen™ Subcloning Efficiency DH5α *E. coli* for the plasmids p1S, p2S, p1A and p2A, and Invitrogen™ One Shot™ TOP10 Chemically Competent *E. coli* for the plasmids 3S and 3A. Selection for successfully transfected clones was achieved by the presence of the antibiotic kanamycin, and screening for the presence of (putative) enhancer fragments in pBR

was performed by colony PCR with the primer pair "EnhColonyFor" and "EnhColonyRev". The integrity of enhancer-reporter vectors was examined by restriction digestion (see Figure 10).

This first transformation was highly successful as we acquired masses of transformed clones and in retrospective, was the first indicator that something failed. When these colonies were screened by colony PCR, almost all colonies were negative, yet they seemed to harbor the kanamycin resistance gene, as they were growing on selective media. To investigate this issue, we randomly choose negative colonies (negative for colony PCR), plus the only four positive (in colony PCR) colonies and inoculated 3 mL LB broth overnight cultures with them. After isolation of plasmid DNA from these overnight cultures, restriction digest (HindIII, KpnI) was performed to validate the integrity of the plasmid. Instead of the appropriate enhancer-reporter vectors, almost only 2 kb fragments were observed (see Figure 10).
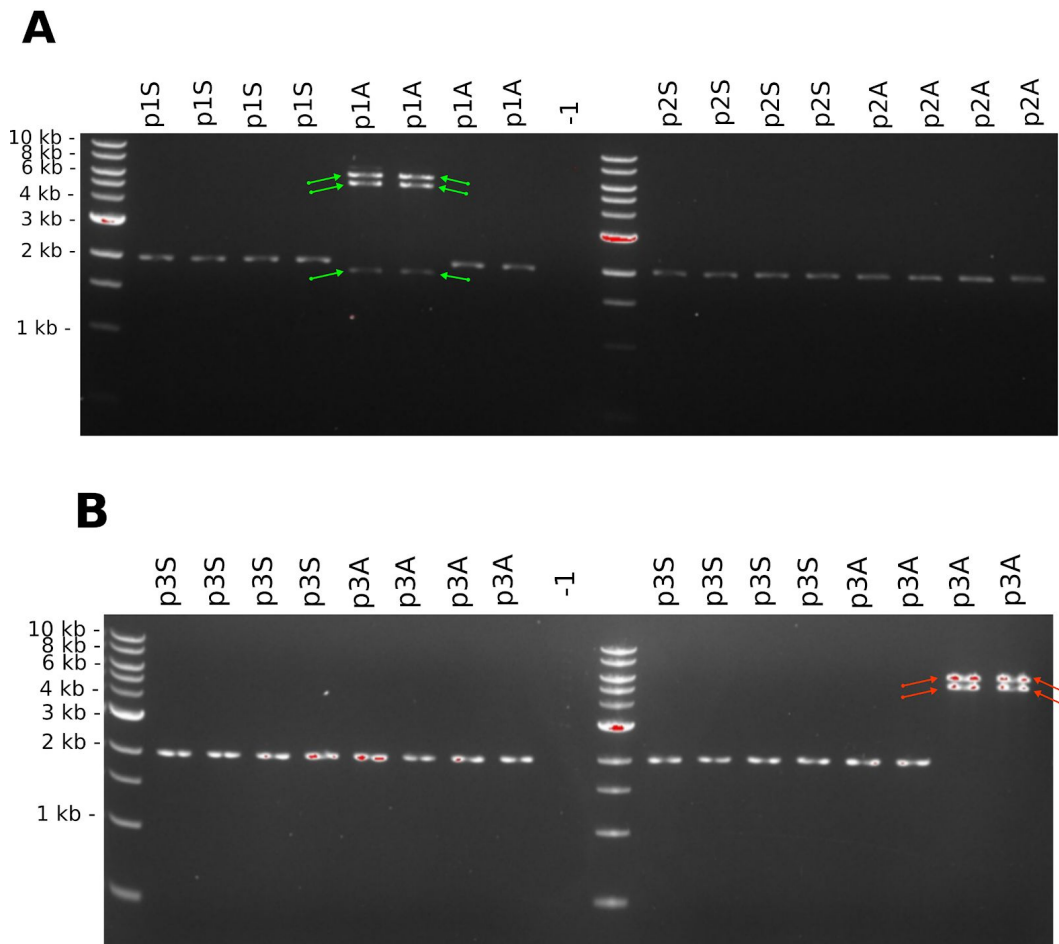
Figure 10 An unexpected 2 kb band emerges in high frequency: (A, B) An unintended ~ 2 kb plasmid emerges in DH5α cells and top10 competent cells transfected with appropriate enhancer-reporter constructs. Agarose gel electrophoresis of the isolated plasmid content from 3 mL overnight cultures after digestion with HindIII and KpnI. Only 4 clones carry intelligible plasmids: (A) two colonies harbor the correct plasmid p1S (marked with green arrows), (B) two colonies carry the empty vector pBR (marked with red arrows). Cultivation, transformation, plasmid isolation, restriction digest and agarose gel electrophoresis performed as described in Material & Methods.

We suspect these 2 kb fragments to be generated as a consequence of intramolecular plasmid recombination (Fishel *et al.* 1981, Laban & Cohen 1981, Smith 1989). We consider the recombination to occur between the *gypsy* insulator sites. Recombination between these elements would result in a fragment with a sequence length of about 2,400 bp, so about the size in which this fragment migrates through the gel. This fragment harbors the kanamycin resistance gene and therefore would allow bacterial growth in the presence of Kanamycin.

Furthermore, the origin of replication (ORI) would be retained in such a recombined "sub"-plasmid, rendering it replication competent (see Figure 11).
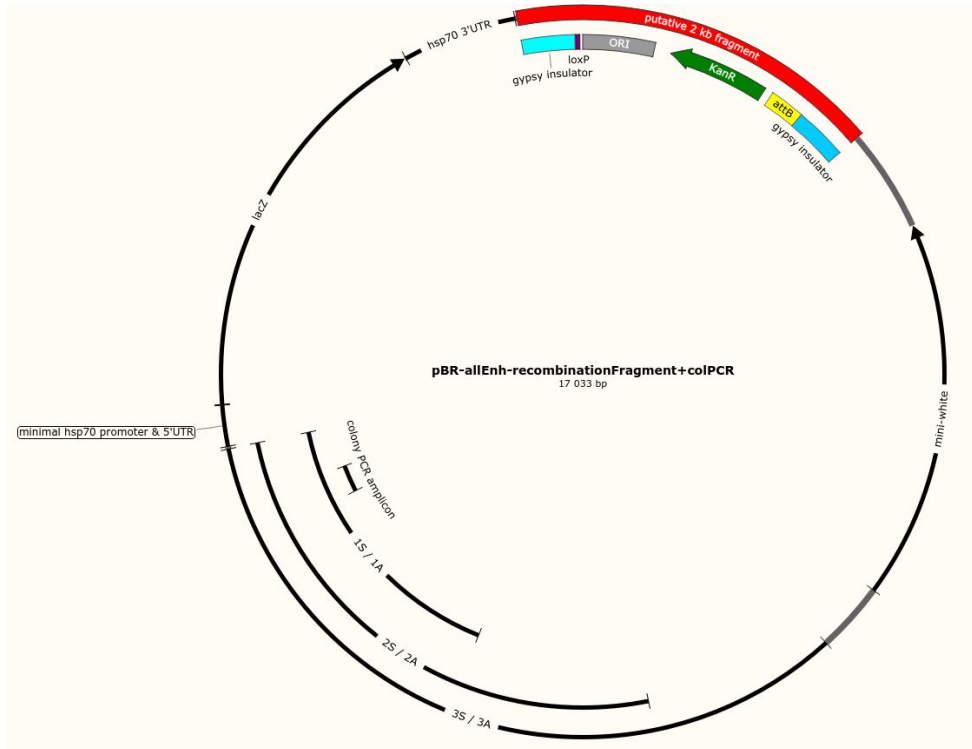
Figure 11 The product of potential intramolecular plasmid recombination between the *gypsy* sequences: This subfragment (marked in red, as "putative 2 kb fragment") would be 2400 bp in length, would contain the ORI (grey), the kanamycin resistance gene (green), the loxP (purple) and the attB (yellow) site.

3.2 Transformation of NEB® stable competent *E. coli*

Based on our suspicion, we tried to circumvent this issue with an additional transfection of NEB® Stable Competent *E. coli*. This particular strain carries a loss of function (null) mutation (recA1) in the recA gene, which has a central role in recombination pathways in E. coli (Umezu & Kolodner 1994), hence recombination in this strain is diminished. However, Invitrogen™ Subcloning Efficiency DH5α and Invitrogen™ One Shot™ TOP10 Chemically Competent *E. coli* carry the same mutation (recA1). After the transformation, screening was performed by colony PCR and positively identified clones were used for the inoculation of 3 mL LB broth overnight cultures. The plasmid content of these cultures was isolated and used as input for agarose gel electrophoresis either digested or undigested (KpnI, HindIII) (see

Figure 12). Curiously, recombination was only slightly impeded and the 2 kb fragment still appeared in addition to the plasmid of interest in some colonies (see Figure 12).
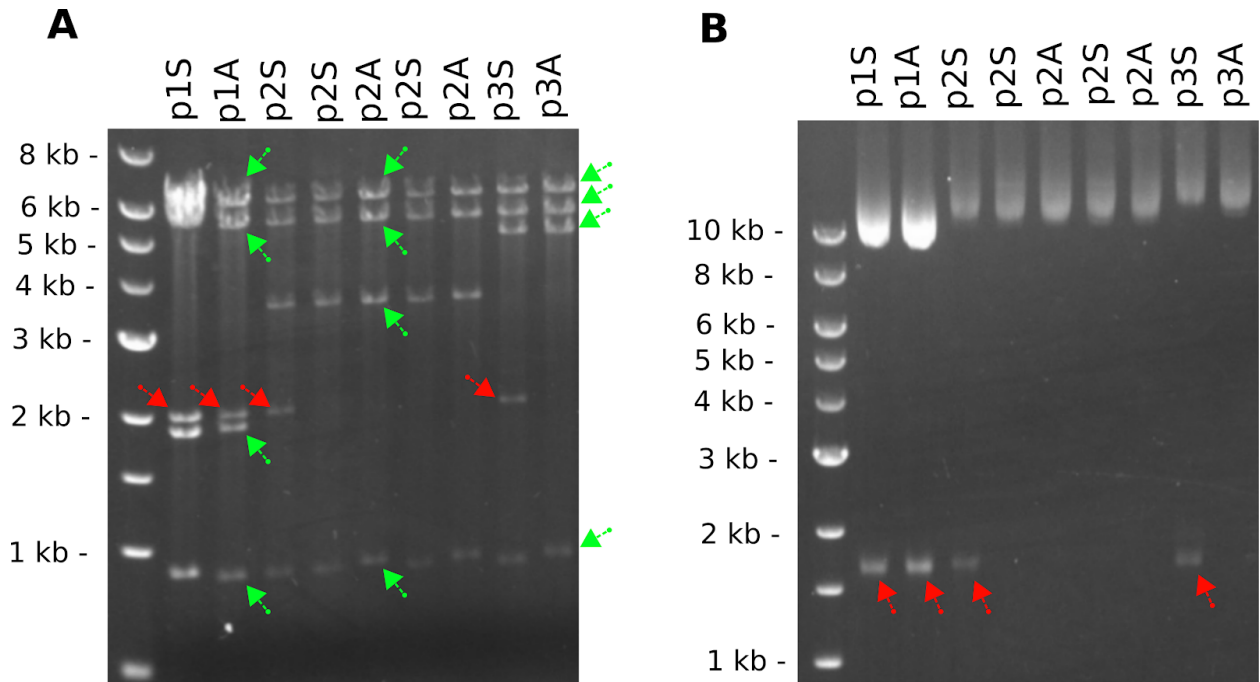

Figure 12 The 2 kb fragment keeps appearing in *recA* deficient *E. coli*: Plasmid content isolated from transfected stable competent *E. coli* (NEB) 3 mL cultures was used for agarose gel electrophoresis either digested with KpnI and HindIII (A) or undigested (B). (A) The expected restriction fragment length is indicated in one lane for each of the three different enhancer-reporter constructs by green arrows. (A, B) The 2 kb contamination (indicated by red arrows) is still emerging additionally to the correct enhancer-reporter constructs, this is evident when the plasmid content is digested (A) or undigested (B). Cultivation, transformation, plasmid isolation, restriction digest and agarose gel electrophoresis was executed as described in Material & Methods.

As a consequence, we performed gel-extraction of the undigested plasmids p1S, p1A, and p3S, and transfected NEB® Stable Competent *E. coli* again with either these gel-extracted constructs, or with isolated plasmids from the last overnight culture for p2S, p2A, and p3A (see Figure 15). This should rule out the possibility that the 2 kb contamination was present in the ligation product. Additionally, we changed the protocol of the transformation (to NEB, C3040: "Protocol for cloning DNA containing repeat elements"). This was done with the intention to minimize the occurrence of recombination as a consequence of the decreased overnight incubation temperature used in this protocol (from 37°C to 30°C). As before, the screening for the enhancer sequence after the transfections was performed by colony PCR,

positive clones were chosen to inoculate 3 mL LB broth overnight cultures, and the plasmid DNA content of these cultures was isolated, digested and used as input for agarose gel electrophoresis (see Figure 13). In the isolated plasmid DNA from these overnight cultures, the 2 kb fragment is not detectable (see Figure 13).
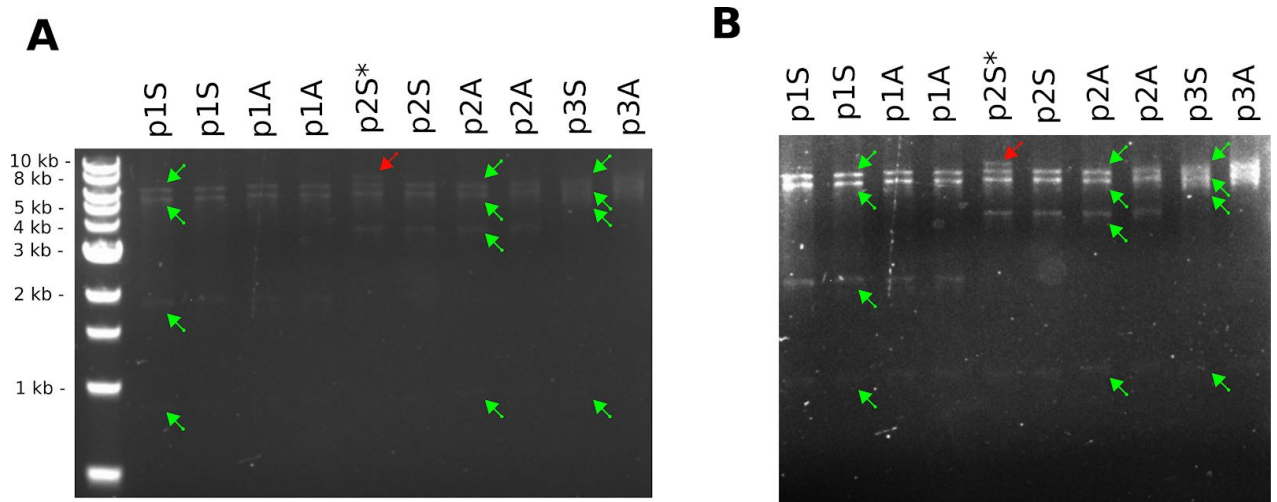

Figure 13 The 2 kb contamination is not present or at much lower levels: NEB® Stable Competent *E. coli* clones, transfected with constructs obtained from gel-extraction (p1S, p1A, p3S) or isolated from overnight cultures (p2S, p2A, p3A) were used to inoculate 3 mL overnight cultures. (A, B) The isolated plasmid content from these cultures was digested (KpnI, HindIII), and used as input (~ 30 ng DNA per lane) for agarose gel electrophoresis. (B) the same electrophoresis as in (A) but with increased exposure. The 2 kb contamination is not detectable in any clone. The expected restriction fragment lengths are indicated in one lane by green arrows for each different construct. Only one colony (labeled p2S*) displayed aberrant restriction fragment lengths (unexpected fragment indicated by a red arrow) and was excluded from any further use. Cultivation, plasmid isolation, restriction digest and agarose gel electrophoresis performed as described in Material & Methods.

Hence we used these cultures that have been validated to carry appropriate enhancer-reporter vectors (the cultures from which the isolated plasmid DNA has been used in Figure 13; see Figure 15 for workflow), to inoculate 50 mL LB broth overnight cultures. After 12 hours of cultivation, the plasmid content from these 50 mL cultures was isolated and used as input for agarose gel electrophoresis, digested or undigested. The 2 kb fragment was present again in all cultures except one (see Figure 14A and 14B).

The quantity of DNA input for this electrophoresis was very high (input ~ 300 ng per lane, see Figure 14A and Figure 14B). The DNA input in the electrophoresis with isolated plasmid

DNA from the cultures that have been used for inoculation was much lower (total DNA input ~ 30 ng per lane, see Figure 13A and 13B). Maybe the contamination was already present in the cultures used for inoculation, but the DNA input for electrophoresis (Figure 13, total DNA input ~ 30 ng) was not enough to allow for detection of low-frequency contamination. Sambrook and Russell state that bands containing as little as ~10 ng DNA can be detected in the presence of ethidium bromide (c = 0.4 μg/mL) in agarose gel (Sambrook & Russell 2001: 5.14), a quantity that was probably not reached.

Therefore, wanted to scrutinize if the 2 kb contamination was already present, although at very low frequency, in the clones used for inoculation of the 50 mL cultures. We used the remaining isolated plasmid content from these clones (so exactly the same plasmid DNA isolate as in Figure 13) to run an additional electrophoresis, this time undigested and with double the amount of total DNA input (~60 ng per lane). Despite the increased DNA quantity, still no contamination is detectable (see Figure 14C) in the clones used for inoculation. This suggests that the re-establishment of the presence of the 2 kb contamination is a consequence of a further round of overnight propagation of the *E.coli* cultures.

Figure 14 Prolonged cultivation re-established the presence of the 2 kb contamination: (A and B) Isolated plasmid content from 50 mL cultures used as input (~ 300 ng DNA per lane) for electrophoresis, digested (first lane of 1S-3A) or undigested (second lane 1S - 3A). (A) and (B) show the same electrophoresis, (B) with higher exposure. The expected restriction fragment lengths are indicated in one lane by green arrows for each construct. The 2 kb contamination (indicated with red arrows) re-appeared in almost all cultures (1S, 1A, 2S, 2A, 3A). (C) The isolated, undigested plasmid DNA content from cultures that were used for inoculation, was used as input (~ 60 ng DNA per lane) for electrophoresis. Still, there is no contamination from the putative recombined "sub"-plasmid detectable. Cultivation, plasmid isolation, electrophoresis and digest performed as described in Material & Methods.

Figure 15 Flowchart of transformations with and propagation of the enhancer-reporter constructs: The workflow that led to the final enhancer-reporter vectors that were used for microinjections. Selection of clones that carry the plasmid pBR (with or without insert) was achieved via kanamycin containing media, screening for clones in which also the various enhancer inserts are present via colony PCR. A detailed description of each individual step is available in Material & Methods.

After all these steps we undertook, we can now conclude that 1) this 2 kb fragment was not present in the initial competent cell aliquots (DH5α cells, top10 cells, stable competent *E. coli*) that were used for transformation, 2) the 2 kb contamination appears as a result of propagation of correct enhancer-reporter vectors in *E. coli*, 3) the plasmid recombines in *E. coli* cells via a pathway that is not hindered through the knockdown of *recA*, and 4) the recombination takes place in sites flanking the Kanamycin resistance gene.

To investigate this unexpected behavior further, we could have isolated the plasmid via gel-extraction and sequenced it, to identify the sites of recombination, and maybe elucidate the pathway by which it recombines. After consultation with my supervisor, we chose not to pursue this issue further for time reasoning. Instead, as the 2 kb plasmid is only present in low

concentrations, we choose to inject the donor plasmids from the overnight culture (see Figure 14A & 14B, see Figure 15), having reduced the 2 kb fragment as much as technically possible in the amount of time. The quality of donor plasmid DNA is supposedly high, as we isolated the DNA from 50 mL cultures with a midi plasmid purification kit (Qiagen).

These (unforeseeable) problematic transformations and propagations consumed a lot of time but eventually I replicated all the enhancer-reporter vectors (p1S, p2S, p3S, p1N, p2N, p3N). In order to gain all six of them, we screened over 975 colonies by colony PCR, so for each construct in average 162,5 colonies had to be screened to find a single colony with the appropriate enhancer-reporter construct (for the workflow see Figure 15). We verified the integrity of all enhancer-reporter constructs via restriction digest and Sanger sequencing.

4. Microinjections and screening for transformants

The next step was the integration of our final constructs into germ-line chromosomes of *Drosophila simulans* (strain #2176) flies.

Microinjections were performed according to standard protocols (see Materials & Methods). Donor and helper plasmids were both present in the injection mix, diluted in milliQ, in concentrations of 250 µg/mL and approximately ~200 µg/mL, respectively. The injections were performed over 2 weeks, injections for each particular enhancer-reporter construct were performed on a single day, in a timeframe of about 7h. Per construct, around 400 embryos should get injected, for the plasmid p1S two days of injections were executed (approximately 650 embryos got injected in total), because of the unsatisfactory quality of the first round of p1S injections. After 48 hours at 19°C, the hatched larvae were rescued and transferred back to 25°C. Microinjections, maintenance and crossing procedure is described in detail in Material & Methods. After microinjections we backcrossed the G0 adults (which have been microinjected as embryos) to *D. simulans w⁻* (*Dsim*#2176, mutation *w1*) flies. If stable integration took place in the germline, the white eye (*w⁻*) phenotype of *Dsim*#2176 will be rescued eventually in the F1 generation, by expression of the dominant *mini-white* (*w⁺*) marker that is present in pBR. Therefore identification of integration events was

straight-forward by screening of the F1 generation for flies with red/orange eyes. Microinjection is a rather harmful procedure to embryos that and often causes irreparable damage to the prospective reproductive system, as a consequence a fraction of the injected G0 flies is sterile.

Table 12. Statistics of microinjections, from the top to the bottom: Record of donor and helper plasmid concentrations, total number of embryos injected, total number of larvae harvested & the percentage of injected embryos that reached larval stage, total number of flies that developed into adult flies & the percentage of larvae that developed into adults, the number of female adult flies, number of male adult flies, and the number of sterile male adult flies & percentage of sterility among males. Because female injected G0 flies were always pooled in groups of 5, it was not accessible if individual adult females were sterile.

	p1S	p1A	p2S	p2A	p3S	p3A	p1S repeat
c_{donor} [ng/μL]	250	250	250	250	250	250	250
c_{helper} [ng/μL]	190	211	200	233	240	242	190
embryos injected (#)	250	381	535	420	345	444	400
larvae yield (# / %)	118 / 47	127 / 33	134 / 25	53 / 13	65 / 19	75 / 17	111 / 28
adults (# / %)	58 / 49	71 / 56	83 / 62	25 / 47	19 / 29	41 / 55	69 / 62
♀ adults (#)	24	25	35	11	5	16	28
♂ adults (#)	34	46	48	14	14	25	41
♂ sterile (# / %)	19 / 56	30 / 65	30 / 63	6 / 43	8 / 57	13 / 52	26 / 63
eyes (white/red)	58/0	71/0	83/0	25/0	19/0	41/0	69/0

Unfortunately, we did not obtain any red-eyed F1 offspring in the course of the study, this suggests that no integration events of pBR occurred in the germline of microinjected G0 flies. The most plausible reasons for this ill success are: 1) Absence of attP sites in the injected fly strain *D. simulans* #2176, 2) poor protocol or execution of microinjections, 3) successful integration of the exogenous enhancer-reporter vector, but transcriptional silencing 4) inadequacy of the *vasa* promoter from *D. melanogaster* to drive transgene expression in the germline of *D. simulans*.

To assay transcriptional silencing of successful integration events, we performed a PCR to amplify the *lacZ* gene from pooled genomic DNA of *w*⁻ F1 flies (who's parents had been microinjected). To validate the presence of the attP site in the strain, we additionally executed

a PCR to amplify the endogenous attP site from pooled genomic DNA of *Dsim*#2176 flies. We isolated genomic DNA from pooled individuals of 1) the strain #32218 from Bloomington (*w*[*];P{*y*[+t7.7]*w*[+mC]=*10XUAS-IVS-mCD8::RFP*}attP2), which is confirmed to possess an endogenous attP site (positive control attP), 2) the *D. melanogaster* strain *wOR*(P) *gypsy-lacZ* (Sarot *et al.* 2004, positive control *lacZ*), 3) the strain *w*[1118] from the Vienna *Drosophila* Resource Center (VDRC) which is validated to possess neither a *lacZ* nor an attP site (negative control for *lacZ* and attP), 4) uninjected flies from the strain *Dsim* #2176, and finally, 5) a fraction of the F1 generation whose G0 parents have been injected. We performed PCR as described in Material & Methods, with primers for 1) the attP site, 2) the *lacZ* gene present in the enhancer-reporter vector.

The results confirmed that the attP site is present in the strain *Dsim*#2176 and that integration of the vector did not occur, at least not in the fraction we tested.

Figure 16 PCR to validate the attP site and examine for *lacZ* site presence in a fraction of the F1 generation: For the first seven lanes PCR reaction was performed using the primer pair "lacZFor" (GATACACTTGCTGATGCGGTGCTGATT) and "lacZRev" (CTGTAGCGGCTGATGTTGAACTGGAAG). The second position is empty. For the last three lanes, the primer pair "attP-For" (CCCAGGTCAGAAGCGGTTTTCG) and "attP-Rev" (TACGTGTCCACCCCGGTCACAA) was used. Control lanes: 1) the empty vector backbone of pBR (positive control *lacZ*), 2) *w*[1118] (negative control for attP and *lacZ*), 3) *wOR(P) gypsy-lacZ*, from Sarot *et al.* 2004 (positive control *lacZ*), and 4) Bloomington #32218 (positive control attP). The F1 Fraction A, B, and C each contain pooled genomic DNA from approximately 60 F1 flies, those 60 derive from about 4 different injected G0 parents. The sample *Dsim*#2176 contains pooled genomic DNA from ~60 flies from the *Dsim*#2176 stock. We do not have an explanation for why the strain *w*[1118] is *lacZ* positive, but we suspect contamination. This PCR validates that no silenced integration took place in the F1 flies fraction tested (see first seven lanes), and that the strain used for injections (*Dsim*#2176) harbors a endogenous attP site (see last 3 lanes).

Discussion

Unfortunately, due to the complications during the amplification process and the failing of the *Drosophila* transformations, the molecular nature of the selective advantage that the favourable haplotype-block confers still remains to be elucidated. We were not able to examine whether the intronic region of *GD13851*, harboring the candidate SNPs, is indeed a CRE and further if these SNPs cause a change in transcription regulatory profile. The functional validation of selection targets identified by E&R studies is usually problematic, this time for technical issues. However, the difficulties we encountered will be discussed in the following.

1. The recombination of pBlueRabbit

It is quite surprising that all *E. coli* strains that we used recombined pBR, given the fact that all these strains are deficient in the *recA* gene (*recA1* mutation, https://international.neb.com/products/c3040-neb-stable-competent-e-coli-high-efficiency#Product%20Information, 28.11.2019). The *recA* gene encodes for a DNA-dependent ATPase, comprised by a filament of proteins, that binds single-stranded DNA (ssDNA) and catalyzes homology search, strand exchange and branch migration (Aguilera & Rothstein 2007). Furthermore, it plays a central role in general recombination processes, such as recombinational DNA repair, and induces the SOS response consequently to DNA damage (Aguilera & Rothstein 2007). Three pathways which enable recombination are known in *E. coli*: the recBCD-dependent recombination pathway, the recF-dependent recombination pathway, and the recE-dependent recombination pathway. The RecBCD-dependent pathway has been described as the main pathway and the recF- and recE-dependent pathway being required mainly only when the recBCD-dependent pathway is malfunctioning (Smith 1989, Smith 1988, Gillen *et al.* 1981). These pathways show a high redundancy in activity and functionality, a strong overlap in the repertoire of enzymes involved, and similar substrate preferences (Smith 1988). The recA protein resides at the central core of these recombination pathways in *E. coli,* and *recA* is essential for all three of them (at least for conjugational

recombination) (Smith 1989, Smith 1988, Gillen *et al.* 1981). Multiple studies revealed that the enzymatic machinery requirement is conditional on the substrates that are to be recombined (Smith 1989, Smith 1988, Gillen *et al.* 1981, Laban & Cohen 1981).

The *recE* gene encodes for the exonuclease VIII (ExoVIII), this exonuclease digests one strand of double-strand DNA (dsDNA), works from 5' to 3' and has a preference for dsDNA ends (Smith 1988). *recE* has been identified by the contrast of *recB- recC-* double mutants, that are characterized by low survival upon exposure to ultraviolet light (UV) or ethyl methane sulfonate (EMS), to *recB- recC- sbcA* triple mutants. An additional mutation in *sbcA* rescues the *recB- recC-* double mutant phenotype and re-establishes survival rates almost to wild type level (Barbour *et al.* 1970). *sbcA* is positioned in close proximity to *recE* and the *sbcA* mutation induces/enhances the expression of *recE*. *sbcA* and *recE* reside on the defective lambda-like prophage *rac*, and it is proposed that *sbcA* is either located in a repressor of *recE*, or that *sbcA* generates a *de-novo* promoter sequence (Kolodner *et al.* 1994, Smith 1988, Laban and Cohen 1981, Chu *et al.* 1989). The derepression of ExoVIII by *sbcA* has been shown to increase the occurrence of plasmidic recombination (Fishel *et al.* 1981, Laban & Cohen 1981). Intramolecular plasmid recombination increases from *recB recC* mutants to *recB recC sbcA* mutants 40-fold, and more than a 1000-fold from *recA* mutants to *recB recC sbcA* mutants (Laban & Cohen 1981).

Conjugational recombination in *E. coli* via the *recE* pathway obligates on *recA*, as well as recombination of two or more plasmids (i.e. intermolecular plasmid recombination) via the *recE* pathway. Remarkably, recombination of elements in a single plasmid (i.e. intramolecular plasmid recombination) via the *recE* pathway is not contingent on *recA* (Smith 1989). It has been shown by Laban & Cohen (1981) that the knockdown of *recA* additionally to the *sbcA* mutation decreases intermolecular plasmid recombination 40-fold, but does not affect the rate of intramolecular plasmid recombination.

Hence, intramolecular plasmid recombination via the *recE* pathway does not require the integrity of the recA protein and therefore is still functional in NEB stable competent cells. The *recE* pathway is an option by which the plasmid pBR could have been constantly recombined in course of its propagation in *E. coli*.

2. The strain *Dsim*#2176 and the donor plasmid pBR

In the *Drosophila simulans* strain *Dsim*#2176 no ectopic, but robust transgene expression has been observed upon the integration of an enhancer-reporter construct into the endogenous attP site (Stern *et al.* 2007). This has been tested by the detection of *lacZ* reporter gene expression regulated from the *D. melanogaster even-skipped* stripe 2 enhancer (Stern *et al.* 2007). The results from Stern *et al.* (2007) suggest the genomic location of the attP site in *Dsim*#2176 to be subjected to negligibly position effects. Additionally, the plasmid pBR has been reported to show no basal expression, to accurately and reproducibly duplicate a known expression pattern (*Notch*), and to exhibit no ectopic reporter gene expression in a genomic location characterized by strong position effects (Housden *et al.* 2012). Although this has been reported for pBR integration into a *D. melanogaster w⁻ y⁻* genome, and no reports for the use of pBR in *D. simulans* are available.

3. The helper plasmid and the *vasa* promoter

The helper plasmid vas-φ-C31(3xP3-EGFP) which supplies the φ-*C31* integrase is tested by Bischof *et al.* (2007) and later by Zhang *et al.* (2014) also in a *Drosophila melanogaster y⁻ w⁻* background. The coding region of the φ-*C31* integrase gene in that plasmid is adapted to the *D. melanogaster* codon usage, this codon-optimization has been shown to further increase integration efficiency in *D. melanogaster* (Bischof *et al.* 2007). The codon usage of *Drosophila simulans* and *Drosophila melanogaster* has been reported before to show a highly similar codon usage pattern (Vicario *et al.* 2007). Therefore, I speculate that the codon usage is more alike between *D. melanogaster* and *D. simulans* than between *Streptomyces* (original host of the φ-C31 phage) and *D. simulans*, and thus further assume that *D. simulans* should also benefit from codon optimized φ-*C31* integrase.

The expression of the φ-*C31* integrase is controlled by the upstream regulatory (promoter) region and the 5'UTR of the *vasa* gene of *D. melanogaster*. This *cis*-regulatory region has been shown to cause germ-line specific zygotic gene expression, which is initiated soon after gastrulation and continues throughout oogenesis and further development (in *D.*

melanogaster) (Sano *et al*. 2002). This sequence has been proven to be suited (in *D. melanogaster*) to drive φ-C31 transgene expression in primordial germ cells, thereby resulting in high rates of integration of donor plasmids into the germline (Bischof *et al*. 2007, Zhang *et al*. 2014). This helper plasmid has been used for the first time in *D. simulans,* at the Institute of Population Genetics of the University of Veterinary Medicine Vienna (Lauri Torma, unpublished) and resulted in a successful transformation events once, but failed in multiple subsequent transformation attempts. However, no information has been reported yet in literature elucidating to what degree the functionality of this promoter and 5'UTR is conserved from *D. melanogaster* to *D. simulans*.

No official orthologue of the *vasa* gene is reported for *D. simulans* (FB2019_05, Thurmond *et al*. 2019, https://flybase.org/reports/FBgn0283442.html#orthologs, 18.11.2019), but using the protein basic local alignment search tool (BLAST) (BLASTP, NCBI Resource Coordinators 2018, Altschul *et al*. 1990) of the translated protein of the *D. melanogaster vasa* gene (isoform *vas-PB*) against the *D. simulans* genome identifies the gene *GD23992* (isoform D) with a sequence analogy of ~82 % over a query cover of ~79 % (BLASTP. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2019]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi) as putative ortholog of *vasa*. Furthermore, GD23992 is a putative syntenic ortholog of the *D. melanogaster vasa* gene, as all genes flanking GD23992 are identified by Flybase (FB2019_05, Thurmond *et al*. 2019) as syntenic orthologs (see Figure 17).

Figure 17 The *GD23392* locus of *D. simulans* and the *vasa* locus of *D. melanogaster* show shared synteny: Display of the genomic locus of *GD23992* (top) and *vasa* (bottom) at the chromosome 2L of *D. simulans* and *D. melanogaster,* respectively. Genomic coordinates are plotted to the rulers. The *vasa* and the *GD23992* gene are highlighted in yellow. All surrounding protein-coding genes are displayed in light and dark blue. Genes that are classified by Flybase as syntenic orthologs are connected with a thin orange line. Information on orthology and screenshot obtained from FB2019_05, Thurmond *et al*. 2019, gBrowse. Screenshot processed via the open-source graphics editor Inkscape.

Even more strikingly when we use BLAST to align the *vasa* promoter and 5'UTR region that is present in the plasmid vas-ϕ-C31(3xP3-EGFP) (total sequence length of this region: 2151 bp) to the *D. simulans* genome (BLAST from FB2019_05, Thurmond *et al*. 2019, default settings, low complexity filter off), a large fragment (1235 bp) is aligned with high analogy (95,3 %), and two smaller sequence fragments are aligned with decreased sequence analogy (86,1 % and 84,2 %, 193 bp and 157 bp, respectively) to the region immediately before the transcription start site (TSS) of *GD23992* (see Figure 18).

Figure 18   The alignment of the *D. melanogaster* and *D. simulans vasa* region: The query (labeled: Drosophila_melanogaster_vasa_promoter_&_5'UTR) is the complete *vasa* promoter and 5'UTR region that drives φ-C31 integrase expression in the plasmid vas-φ-C31(3xP3-EGFP). This sequence was blasted against the genome of *D. melanogaster* (Dmel, r6.31) (positive control) and *D. simulans* (Dsim, r2.02). The *D. melanogaster vasa* gene spans from 2L:15,061,656 to 2L:15,074,311. A large fragment with high analogy, and two smaller sequence fragments with intermediate analogy to the *vasa* promoter have been identified immediately upstream of the putative *D. simulans vasa* ortholog *GD23992* (*GD23992* located on Scf_2L:14,574,667..14,587,306). However, the alignment to *D. simulans* is discontinuous, contains large gaps and decreases in analogy in proximity to the TSS of the *GD23992*. Screenshot obtained from the tool BLAST from FB2019_05. The genomic start- and end-coordinates are added and irrelevant sub-alignments are veiled via the open-source graphics editor Inkscape.

Also, the crucial 40 bp sequence region in the promoter close to the 5'UTR, that have been observed to be sufficient, and more importantly, obligatory for transgene expression in primordial germ cells during embryogenesis (Sano *et al*. 2002), is covered in this alignment with an analogy of 38 out of 40 bases (see Figure 19).

```
>gnl|dsim|Scf_2L  type=golden_path_region; loc=Scf_2L:1..23539531; ID=Scf_2L; dbxref=GB:CM002910;
MD5=4db334c02c86dfa856dc1a48c595acf1; length=23539531; release=r2.02; species=Dsim;

HSP # = 2 , Score = 165.028 bits (83) , Expect = 1.19969e-38
Identities = 167 / 194 (86.1%) , Positives = 167 / 194 (86.1%) , Gaps = 2 / 194 (1%)
Strand = Plus / Plus
```

| GBrowse | Subject FASTA |

```
vasa  promoter & 5'UTR:  1866    ATCACTTAGGTTGCTTGAATATTACAATTTCATTAATAGCTAAATCACATTTGATGTGTT    1925
                                 |||||| |||| | |||||| ||||  |||||| |||||||||||||||||||||||||||
Dsim\Scf_2L:  14574517           ATCACTAAGGTGGTTTGAACATTAGTATTTCGTTAATAGCTAAATCACATTTGATGTGTC    14574576

vasa  promoter & 5'UTR:  1926    AGTGGAAAACGGCTATATATATAAATTATCGAAATTGTGAATATCGAATTGCGATAGCAC    1985
                                 ||||||||||||||| ||||| | ||||||||||||| |||| | | | ||||||||||||
Dsim\Scf_2L:  14574577           AGTGGAAAACGGCTGTATATGGATATTATCGAAATTTTGAAAA--GTAATGCGATAGCAC    14574634

vasa  promoter & 5'UTR:  1986    AATGGGAAATTCCACCACTAGATTTTTGGTACTTTTAACAGATCCTTTTCGGTTTTGCGT    2045
                                 |||||| |||||||||||||| | |  ||||||||||||||||||||||||||||||||||
Dsim\Scf_2L:  14574635           CATGGGATATTCCACCACTAGCTCTGGATTACTTTTAACAGATCCTTTTCGGTTTTGCGT    14574694

vasa  promoter & 5'UTR:  2046    TGCGCGAAGTGATC    2059
                                 ||||||| ||||||
Dsim\Scf_2L:  14574695           TGCGCGAGGTGATC    14574708
```

Figure 19 The crucial sequence region of the *D. melanogaster vasa* promoter is covered in the alignment: This Figure shows a part of the alignment of the *vasa* promoter & 5'UTR of the plasmid vas-φ-C31(3xP3-EGFP) (indicated on the left: *vasa* promoter & 5'UTR) to the *D. simulans* genome (indicated on the left: Dsim\Scf_2L; r2.02) on sequence level (same alignment as in Figure 18). The shown sequence Dsim\Scf_2L: 14574517...14574708 constitutes the first of the two regions (from the left) in the alignment (Figure 18) with a key of 80-200 (first one of the two purple fragments in Figure 18). The 40 bp sequence that is reported to be crucial for germ-line specific transgene expression (Sano *et al*. 2002) is indicated in red. This 40 bp sequence fragment is conserved (38 of 40 bases are analog) in close proximity to the TSS of the putative *vasa* ortholog *GD23992*. Screenshot from FB2019_05, BLAST, Thurmond *et al*. 2019, default settings, low complexity filter off. via the open-source graphics editor Inkscape.

The high sequence conservation of the crucial 40 bp region (38 out of 40 are analog) would suggest that the functionality of this region is conserved as well. *Cis*-regulatory elements are known to have a high sequence turnover in *Drosophila* species. Furthermore, low sequence conservation at CRE's, high functional conservation, and regulatory incompatibility has been reported between *Drosophila species* (Arnold *et al*. 2014, Coolon *et al*. 2014, Wittkopp & Kalay 2011, Lemos *et al*. 2008, Taher *et al*. 2011). However, the observed level of sequence conservation of the complete *vasa* promoter region does not permit any inferences about functional conservation. Experimental evidence via an *in vivo* enhancer-reporter assay is required for any conclusion about the cross-species functionality of this sequence fragment.

Conclusion

1. The recombination of pBlueRabbit

To my knowledge, no study has been conducted yet to investigate the rate of intramolecular plasmid recombination via the *recE* pathway in *recA* deficient cells without *sbcA* mutation. The lab of Gillen *et al.* (1981) stated that the *recE* pathway is only active and detectable if a mutation in *recB* or *recC* or both is present additional to the *sbcA* mutation. If that is indeed true, this would of course counter the possibility that the competent cells we used recombined pBR via the *recE* pathway.

For one strain with the genotype *recB-, recC-* and *recE-* (mutation *rac-0*) and diminished ExoVIII activity (< 3 %) has been reported (*E. coli* strain JC 5519: Gillen *et al.* 1981, Strain: The Coli Genetic Stock Center (CGSC) #: 5114, https://cgsc2.biology.yale.edu/Strain.php?ID=8882, 01.12.2019, EcoliWiki: https://ecoliwiki.org/colipedia/index.php/Category:Strain:JC5519, 01.12.2019). Another strain carries the same *rac-0* mutation but is otherwise wild type for both genes *recB* and *recC* (CGSC #: 1157, https://cgsc.biology.yale.edu/Strain.php?ID=4509, 01.12.2019, EcoliWiki: https://ecoliwiki.org/colipedia/index.php/Category:Strain:AB1157, 01.12.2019). It would be quite revealing to render these two strains transformation competent, and to transfect them with our enhancer-reporter constructs and check for the occurrence of any recombination.

Another quicker and more feasible option would be to use another vector that is suited for enhancer-reporter assays, instead of pBR, which does not contain the suspected loci of recombination (*gypsy* insulator sites).

2. The *vasa* promoter

Because the functionality of the *D. melanogaster vasa* promoter and 5'UTR is not functionally validated in *D. simulans,* the change of the helper plasmid would a promising proposition to tackle the issue of failed microinjections. Probably the best option would be to use the helper plasmid pBS130 (Gohl *et al. 2011*) that is used in David L. Stern's lab, Janelia Farm, USA. This plasmid carries a heat-shock inducible (heat-shock protein 70-promoter) source of the φ-*C31* integrase and is standardly used for transformations of *D. simulans*. Embryos are incubated by Stern *et al.* one hour after microinjections for one hour at 37° C, and a high integration rate is regularly observed with this protocol using pBS130 (1/21 in *Dsim*#2176, see Stern *et al.* 2017).

Another possibility would be to construct a plasmid that expresses the φ-*C31* integrase from the promoter of the *D. melanogaster nanos* gene. The *nanos* promoter has been shown to drive germ-line specific expression in *D. simulans* and *D. melanogaster* and furthermore, to drive a similar expression pattern in in both species during embryonic development (Holtzman *et al.* 2010).

Summary

The lab of Christian Schlötterer conducted experimental evolution studies with a South African population of *Drosophila simulans* for 60 non-overlapping generations under stressful, hot-cycling (18°C/28°C) environmental conditions. The most significant and focal allele frequency change was observed for SNPs located in a putative *cis*-regulatory structures of the gene *GD13851*. This suggests that modulated gene expression of *GD13851* may be the causative adaptive change leading to increased thermal fitness in hot environments.

To test if the SNPs in the particular *cis*-regulatory region cause a change in the spatiotemporal transcription regulation of *GD13851*, enhancer-reporter assays in *D. simulans* were performed.

Therefore multiple enhancer-reporter vectors were generated, in which the reporter gene *lacZ* is driven from a minimal promoter, by the putative *cis*-regulatory sequence of either the "evolved" or the ancestral haplotype. Additionally, this enhancer-reporter plasmid contains an attB site to enforce stable, site-specific genomic integration via the ϕ-C31 integrase system. As a result position effects on gene expression are considered constant, and differences in spatio-temporal expression of the reporter gene can be considered as an inherent feature of the different enhancer fragments.

The propagation of these constructs was performed using NEB® Stable Competent *Escherichia coli* cells, a strain that is deficient in the *recA* gene, a gene that has been proven to be central to recombination processes in *E. coli*. Surprisingly, prolonged cultivation resulted in this strain in intramolecular plasmid recombination within sub-elements of our enhancer-reporter vectors. The *recE* pathway could be a mean to facilitate recombination in *recA* mutant *E. coli*.

Microinjections into an attP site carrying *D. simulans* (*y⁻*,*w⁻*) strain were executed. A *mini-white* gene present in the enhancer-reporter construct renders screening and identification of successful integration events straightforward in the *w⁻* genetic background.

Unfortunately, no transformed F1 offspring was received for any of the 6 injected constructs. The presence of the attP site in the *D. simulans* strain was validated by PCR. The *vasa* promoter from *D. melanogaster* that was used in order to drive ϕ-C31 integrase is suspected to be non-functional in *Drosophila simulans*.

Abbreviations

bp	-	base pair
CRE	-	*cis*-regulatory element
E&R	-	evolve and resequence
EE	-	experimental evolution
kb	-	kilobases
Mbp	-	megabase pairs
PCR	-	polymerase chain reaction
RT-qPCR	-	reverse transcription quantitative polymerase chain reaction
TF	-	transcription factor
TSS	-	transcription start site

Glossary

Basal apparatus - the complex of transcription factors that assemble at the promoter before RNA polymerase is bound

Beneficial allele - an allele that is associated with an increase in fitness of the carrier

Clonal interference - competition among beneficial alleles that occur in (asexual) clones (Elena & Lenski 2003)

Coactivators - required for the function of DNA-binding activators, but not for basal transcription per se, and do not show site-specific binding by themselves

Deleterious allele - an allele that is associated with an decrease in fitness of the carrier

Fixation - the replacement of all genetic variants at a particular locus by an (beneficial) allele, resulting in genetic monomorphism at the locus

Gene - a gene is defined as a DNA segment that contributes to phenotype or function. In the absence of demonstrated function a gene may be characterized by sequence, transcription of homology (Wain *et al.* 2002)

Glycan - synonymous with polysaccharides, compounds consisting of large numbers of monosaccharides linked glycosidically (Moss *et al.* 1995)

Hard selective sweep - if the beneficial allele originates from a single mutation, then all ancestral variation that is linked to the allele locus will be eliminated by hitchhiking (Hermisson & Pennings 2005)

Linkage disequilibrium - nonrandom association between alleles of two loci in a population (Schlötterer *et al.* 2015)

Locus - refers to a genomic map position (no synonym for gene) (Wain *et al.* 2002)

Long-range hitchhiking - linkage of a neutral genetic variant to a beneficial allele in *cis*, over long genomic distances

Protein-coding gene - gene coding for a protein

Selective sweep - reduction or elimination of sequence variation near a beneficial allele that increases in frequency and eventually reaches fixation

Soft selective sweep - if a allele is present in the population in multiple copy and then renders beneficial, a part of the variation linked to the allele locus will be preserved (Hermisson & Pennings 2005)

Standing genetic variation - is used to refer to a population that harbors a substantial amount of genetic polymorphisms

Transcription factor binding motif - degenerate sequence pattern which summarize the binding preferences of a particular TF

References

Alberts B, Bray D, Hopkin K, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2014. Essential Cell Biology. Fourth Edition. New York: Garland Science, Taylor and Francis Group.

Altschul SF, Gish W, Miller W, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol., 215:403-410.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. Nature Genetics 46(7) : 685-692.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?. Journal of cellular biochemistry, 94:890-898.

Barbour SD, Nagaishi H, Templin A, Clark AJ. 1970. Biochemical and genetic studies of recombination proficiency in *Escherichia coli*, II. rec+ revertants caused by indirect suppression of rec-mutations. Proceedings of the National Academy of Science, 67(1):128-135.

Barghi N, Tobler R, Nolte V, Jaksic AM, Mallard F, Otte KA, Dolezal M, Taus T, Kofler R, Schlötterer C. 2019. Genetic redundancy fuels polygenic adaptation in *Drosophila*. PLOS Biology, 17(2):e3000128.

Barghi N, Tobler R, Nolte V, Schlötterer C. 2017. *Drosophila simulans*: A species with improved resolution in evolve and resequence studies. Genes Genomes Genetics, 7:2337-2343.

Barolo S, Carver LA, Posakony JW. 2000. GFP and β-Galactosidase Transformation Vectors for Promoter/Enhancer Analysis in *Drosophila*. BioTechniques, 29:726-732.

Barton NH, Briggs DG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. New York: Cold Spring Harbor Laboratory Press.

Becher PG, Flick G, Rozpedowska E, Schmidt A, Hagman A, Lebreton S, Larsson MC, Hansson BS, Piskur J, Witzgall P, Bengtsson M. 2012. Yeast, not fruit volatiles mediate *Drosophila melanogaster* attraction, oviposition and development. Functional Ecology, 26:822-828.

Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for *Drosophila* using germ-line-specific Φ-C31 integrases. PNAS, 104:3312-3317.

Butler JEF & Kadonaga JT. 2001. Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs. Genes and Development, 15:2515-2519.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of a morphological evolution. Cell, (134):24 - 36.

Chu CC, Templin A, Clark AJ. 1989. Suppression of a frameshift mutation in the *recE* Gene of *Escherichia coli* K-12 occurs by gene fusion. Journal of Bacteriology, 171(4):2101-2109.

Colley KJ, Varki A, Kinoshita T. 2017. Chapter 4: Cellular Organization of Glycosylation. In: Varki A *et al*., editors. Essentials of Glycobiology. Third Edition. New York: Cold Spring Harbor Laboratory Press.

Crocker J, Ilsley GR. 2017. Using synthetic biology to study gene regulatory evolution. Current Opinion in Genetics & Development, 47:91-101.

De Jong G and Bochdanovits Z. 2003. Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, and the insulin-signaling pathway. Journal of Genetics, 82(3):207-223.

Drosophila 12 Genome Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature, 450:203-218.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. PNAS, 99(25):16144-16149.

Eddison M, Belay AT, Sokolowski MB, Heberlein U. 2012. A genetic screen for olfactory habituation mutations in *Drosophila*: analysis of novel foraging alleles and an underlying neural circuit. Plos One, 7(12):e51584.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature Reviews Genetics, 4: 457-469.

Finlay BJ, Esteban GF. 2009. Can Biological Complexity Be Rationalized?. BioScience, 59(4):333-340.

Fishel RA, James AA, Kolodner R. 1981. *recA*-independent general recombination of plasmids. Nature, 294:184-186.

Fish MP, Groth AC, Calos MP, Nusse R. 2007. Creating transgenic Drosophila by microinjecting the site-specific ΦC31 integrase mRNA and a transgene-containing donor plasmid. Nature Protocols, 2(10) : 2325 -2331.

Franssen SU, Barton NH, Schlötterer C. 2016. Reconstruction of haplotype-blocks selected during experimental evolution. Mol. Biol. Evol., 34(1):174-184.

Gaudet P, Livstone M, Thomas P. 2011. Gene Ontology annotation inferences using phylogenetic trees. GO Reference Genome Project. Brief. Bioinform., 12(5):449-462.

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA III, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods, 6(5) : 343 - 345.

Gillen JR, Willis DK, Clark AJ. 1981. Genetic analysis of the recE pathway of genetic recombination in *Escherichia coli* K-12. Journal of Bacteriology, 145(1):521-532.

Glater EE, Megeath LJ, Stowers RS, Schwarz TL. 2006. Axonal transport of mitochondria requires milton to recruit kinesin heavy chain and is light chain independent. The Journal of Cell Biology, 173(4):545-557.

Gompel N, Schröder EA. 2015. *Drosophila* Germline Transformation. From the Nicolas Gompel's lab website: http://gompel.org/methods. url:

http://gompel.org/wp-content/uploads/2015/12/Drosophila-transformation-with-chorion.pdf,
21.11.2019.

Grandchamp N, Altémir D, Philippe S, Ursulet S, Pilet H, Serre MC, Lenain A, Serguera C, Mallet J, Sarkis C. 2014. Hybrid lentivirus-ϕ-C31-int-NLS vector allows site-specific recombination in murine and human cells but induces DNA damage. PLOS ONE, 9(6):e99649.

Gray S, Szymanski P, Levine M. 1994. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes & Development, 8:1829-1838.

Greer LF. 2017. Editorial: Genomics of experimental evolution. Frontiers in Genetics, 8:93.

Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of Transgenic *Drosophila* by Using the Site-Specific Integrase from phage ϕ-C31. Genetics, 166:1775-1782.

Haley CS & Birley AJ. 1983. The genetical response to natural selection by varied environments, II. Observations on replicate populations in spatially varied laboratory environments. Heredity, 51(3):581-606.

Harper M, Boyce JD, Adler B. 2006. *Pasteurella multocida* pathogenesis: 125 years after Pasteur. FEMS Microbiol Lett, 265:1-10.

Hart GW, Slawson C, Ramirez-Correa G, Lagerlof O. 2011. O-GlcNAcylation and phosphorylation: roles in signaling, transcription and chronic disease. Annu. Rev. Biochem., 80:825-58.

Hermisson J, Pennings PS. 2005. Soft Sweeps : Molecular population genetics of adaptation from standing genetic variation. Genetics, 169: 2335-2352.

Hiromi Y and Gehring WJ. 1987. Regulation and function of the *Drosophila* segmentation gene *fushi tarazu*. Cell, 50:963-974.

Holtzman S, Miller D, Eisman RC, Kuwayama H, Niimi T, Kaufman TC. 2010. Transgenictools for members of the genus *Drosophila* with sequenced genomes. Fly, 4(4):349-362.

Housden BE, Millen K, Bray SJ. 2012. *Drosophila* reporter vectors compatible with φC31 integrase transgenesis techniques and their use to generate new notch reporter fly lines. Genes Genomes Genetics, 2:79-82.

Janssen H, Hou S, Jaeger J, Kim AR, Myasnikova E, Sharp D, Reinitz J. 2006. Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster even skipped* gene. Nature Genetics, 38(10):1159-1165.

Jenett A, Rubin GM, Ngo TT, Shepherd D, Murphy C, Dionne H, Pfeiffer BD, Cavallaro A, Hall D, Jeter J, Iyer N, Fetter D, Hausenfluck JH, Peng H, Trautman ET, Svirskas RR, Myers EW, Iwinski ZR, Aso Y, DePasquale GM, Enos A, Hulamm P, Lam SC, Li HH, Laverty TR, Long F, Qu L, Murphy SD, Rokicki K, Safford T, Shaw K, Simpson JH, Sowell A, Tae S, Yu Y, Zugates CT. 2012. A Gal4-Driver Line Resource for *Drosophila* Neurobiology. Cell Reports 2:991-1001.

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Caroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. Cell, 132 : 783-793.

Jónás Á, Taus T, Kosiol C, Schlötterer C, Futschick A. 2016. Estimating the effective population size from temporal allele frequency changes in experimental evolution. Genetics, 204:(723-735).

Jory A, Estella C, Giorgianni MW, Slattery M, Laverty TR, Rubin GM, Mann RS. 2012. A survey of 6,300 genomic fragments for *cis*-regulatory activity in the imaginal discs of *Drosophila melanogaster*. Cell Reports, 2:1014-1024.

Judd BH. 1995. Mutations of *zeste* that mediate transvection are recessive enhancers of position-effect variegation in *Drosophila melanogaster*. Genetics, 141:245-253.

Justice N, Jan YN. 2002. Variations on the notch pathway in neural development. Current Opinion in Neurobiology, 12:64-70.

Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC. 2012. Experimental evolution. Trends in Ecology and Evolution, 27(10):547 - 560.

Kiehart DP, Crawford JM, Montague RA. 2000. Quantitative Microinjection of *Drosophila* embryos. In: Sullivan W, Ashburner M, Hawley RS, editors. *Drosophila* Protocols. New York:

Cold Spring Harbor Laboratory Press.

Kim TK, Shiekhattar R. 2015. Architectural and functional commonalities between enhancers and promoters. Cell, 162 : 948-959.

King MC and Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188 (4148) : 107 - 116.

Kofler R and Schlötterer C. 2014. A guide for the design of evolve and resequence studies. Mol. Biol. Evol. 31(2):474-483.

Kolodner R, Hall SD, Luisi-DeLuca C. 1994. Homologous pairing proteins encoded by the *Escherichia coli recE* and *recT* genes. Molecular Microbiology, 11(1):23-30.

Kolss M, Vijendravarma RK, Schwaller G, Kawecki TJ. 2009. Life-history consequences of adaptation to larval nutritional stress in *Drosophila*. Evolution, 63(9):2389-2401.

Krebs JE, Kilpatrick S, Goldstein E. 2014. Lewin's Genes XI, 11th Edition. Burlington: Jones & Bartlett Learning.

Kuhstoss S, Rao RN. 1991. Analysis of the integration function of the streptomyces bacteriophage φ-C31. J. Mol. Biol., 222(897-908).

Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A. 2014. Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. Nature, 512:91-95.

Kvon EZ. 2015. Using transgenic reporter assays to functionally characterize enhancers in animals. Genomics 106:185-192.

Laban A, Cohen A. 1981. Interplasmidic and intramolecular plasmid recombination in *Escherichia coli* K-12. Mol Gen Genet, 184 : 200-207.

Latchman DS. 1997. Transcription factors An overview. The International Journal of Biochemistry & Cell Biology, 29(12):1305–1312.

Lee D, Redfern O, Orengo C. 2007.Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology. 8:995-1005.

Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. PNAS, 105(38) : 14471-14476.

Levine M and Tijan R. 2003. Transcription Regulation and animal diversity. Nature 424:147-151.

Lin YR, Reddy BVVG, Irvine KD. 2008. Requirement for a core 1 galactosyltransferase in the *Drosophila* nervous system. Developmental Dynamics, 237:3703-3714.

Li W, De Schutter K, Van Damme EJM, Smagghe G. 2019. Synthesis and biological roles of O-glycans in insects. Glycoconjugate Journal, doi: 10.1007/s10719-019-09867-1.

Lloyd VK, Sinclair DAR, Alperyn M, Grigliatti TA. 2002. Enhancer of *garnet*/δAP-3 is a cryptic allele of the *white* gene and identifies the intracellular transport system for the white protein. Genome, 45:296-312.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nature Reviews Genetics, 16(10):567-582.

Long HK, Prescott Sl, Wysocka J. 2016. Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell, 167 : 1170-1186.

Louis M, Polavieja G. 2017. Collective Behaviour: Social Digging in *Drosophila* Larvae. Current Biology, 27:R1002-R1023.

Mackenzie SM, Howells AJ, Cox GB, Ewart GD. 2000. Sub-cellular localisation of the White/Scarlet ABC transporter to pigment granule membranes within the compound eye of *Drosophila melanogaster*. Genetica, 108:239-252.

Markstein M, Pitsouli C, Villalta C, Celniker SE, Perrimon N. 2008. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nature Genetics, 40:476-483.

Matthews KA, Kaufman TC, Gelbart WM. 2005. Research resources for *Drosophila*: the expanding universe. Nature Reviews Genetics, 6:179-193.

McManus CJ, Coolon JD, O'Duff M, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. Genome Research, 20(6):816-825.

Michel B, Baharoglu Z, Lestini R. 2007. Genetics of recombination in the model bacterium *Escherichia coli*. In: Aguilera A, Rothstein R, editors. Molecular Genetics of Recombination. Berlin Heidelberg : Springer Verlag, 17:1-26.

Mohr SE, Hu Y, Kim K, Housden BE, Perrimon N. 2014. Resources for functional genomic studies in *Drosophila melanogaster*. Genetics, 197:1-18.

Morgan TH. 1910. Sex limited inheritance in *Drosophila*. Science, 32(812):120-122.

Moss GP, Smith PAS, Tavernier D. 1995. Glossary of class names of organic compounds and reactive intermediates based on structure. Pure & Appl. Chem., 67(8/9):1307-1375.

Naylor LH. 1999. Reporter gene technology: the future looks bright. Biochemical Pharmacology, 58 : 749-757.

NCBI Resource Collaborators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 46(D1):D8-D13.

Nouhaud P, Tobler R, Nolte V, Schlötterer C. 2016. Ancestral population reconstitution from isofemale lines as a tool for experimental evolution. Ecology and Evolution, 6(20): 7169-7175.

Nuzhdin SV, Keightley PD, Pasyukova EG. 1993. The use of retrotransposons as markers for mapping genes responsible for fitness differences between related *Drosophila melanogaster* strains. Genetic Research, 62:125-131.

Olorunniji FJ, Rosser SJ, Stark WM. 2016. Site-specific recombinases: molecular machines for the genetic revolution. Biochem. J., 473 : 673-684.

Orozco-terWengel P, Kapun M, Nolte V, Kofler R, Flatt T, Schlötterer C. 2012. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Molecular Ecology, 21:4931-4941.

Overbergh L, Giulietti A, Valckx D, Decallonne B, Bouillon R, Mathieu C . 2003. The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression. Journal of Biomolecular Techniques, 14:33-43.

Palmieri N, Nolte V, Chen J, Schlötterer. 2014. Genome assembly and annotation of a *Drosophila simulans* strain from madagascar. Molecular Ecology Resources, 15:372-381.

Papadopoulos D, Schneider, Meier-Eiss J, Arber W, Lenski RE, Blot M. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. Proceedings of the National Academy of Science USA, 96:3807-3812.

Pekkurnaz G, Trinidad JC, Wang X, Kong D, Schwarz TL. 2014. Glucose regulates mitochondrial motility via milton modification by O-GlcNAc Transferase. Cell, 158:54-68.

Pelham H. 1985. Activation of heat-shock genes in eukaryotes. Trends in Genetics, 1:31–35.
Pfeiffer et al. 2008. Tools for neuroanatomy and neurogenetics in *Drosophila*. PNAS, 105(28) : 9715-9720.

Pirrotta V. 1988. Vectors for P-mediated transformation in *Drosophila*. Biotechnology, 10:437-456.

Plotkin SA & Plotkin SL. 2011. The development of vaccines: how the past led to the future. Nature Reviews Microbiology, 9:889 - 893.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. PNAS, 104:8605-8612.

Raff JW, Glover DM. Centrosomes, and not nuclei, initiate pole cell formation in *Drosophila* embryos. Cell, 57:611-619.

Remolina SC, Chang PL, Leips J, Nuzhdin SV, Hughes KA. 2012. Genomic basis of aging and life history evolution in *Drosophila melanogaster*. Evolution, 66(11):3390-3403.

Rengifo J, Gibson CJ, Winkler E, Collin T, Ehrlich BE. 2007. Regulation of the inositol 1,4,5-trisphosphate receptor type I by O-GlcNAc glycosylation. The Journal of Neuroscience, 27(50):13813-13821.

Reuter JA, Spacek D, Snyder MP. 2015. High-Throughput Sequencing Technologies. Mol Cell, 58(4):586-597.

Sabl JF, Birchler JA. 1993. Dosage dependent modifiers of *white* alleles in *Drosophila melanogaster*. Genet. Res., 62:15-22.

Sambrook J, Russel DW. 2001. Molecular Cloning, A Laboratory Manual. Third edition. New York: Cold Spring Harbor Laboratory Press. 5.14 - 5.15, 8.20, 5.5, 5.13, A2.5 - A2.6, 6.30.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acid Research, 32:D91-D94.

Sano H, Nakamura A, Kobayashi S. 2002. Identification of a transcriptional regulatory region for germline-specific expression of vasa gene in *Drosophila melanogaster*. Mechanism of development, 112:129-139.

Sano H, Nakamura A, Kobayashi S. 2002. Identification of a transcriptional regulatory region for germline-specific expression of vasa gene in *Drosophila melanogaster*. Mechanism of development, 112:129-139.

Santos M, Borash DJ, Joshi A, Bounlutay N, Mueller LD. 1997. Density-dependent natural selection in *Drosophila*: evolution of growth rate and body size. Evolution, 51(2):420-432.

Sarot E, Payen-Groschêne G, Bucheton A, Pélisson A. 2004. Evidence for a *piwi*-dependent RNA silencing of the *gypsy* endogenous retrovirus by the *Drosophila melanogaster flamenco* gene. Genetics, 166(3):1313-1321.

Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity, 114:431-440.

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nature Reviews Genetics, 15(11):1-15.

Schneider D, Lenski RE. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Research in Microbiology, 155:319-327.

Schwientek T, Keck B, Levery SB, Jensen MA, Pedersen JW, Wandall HH, Stroud M, Cohen SM, Amado M, Clausen H. 2002. The *Drosophila* gene brainiac encodes a glycosyltransferase putatively involved in glycosphingolipid synthesis. J. Biol. Chem., 277(36):32421-32429.

Seeman NC, Rosenberg JM, Rich A. 1976. Sequence-specific recognition of double helical acids by proteins. Proc. Nat. Acad. Sci., 73(3):804-808.

Serebriiskii IG, Golemis EA. 2000. Uses of *f* to study gene function: Evaluation of β-Galactosidase assays employed in the yeast two-hybrid system. Analytical Biochemistry, 285:1-15.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics, 15:272-286.

Signor SA and Nuzhdin SV. 2018. The evolution of gene expression in *cis* and *trans*. Trends Genet., 34(7):532-544.

Silicheva M, Golovin A, Pomerantseva E, Parshikov A, Georgiev P, Maksimenko O. 2010. *Drosophila mini-white* model system: new insights into positive position effects and the role of transcriptional terminators and *gypsy* insulators in transgene shielding. Nucleic Acids Research, 38(1):39-47.

Smale & Kadonaga. 2003. The RNA polymerase II core promoter. Annu. Rev. Biochem., 72:449-479.

Smale ST. 2001. Core promoters active contributors to combinatorial gene regulation. Genes & Development, 15:2503-2508.

Small S, Arnosti DN, Levine M. 1993. Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. Development, 119:767-772.

Small S, Blair A, Levine M. 1992. Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. The EMBO Journal, 11(11):4047-4057.

Smith GR. 1988. Homologous recombination in prokaryotes. Microbiological reviews, 52(1) : 1-28.

Smith GR. 1989. Homologous recombination in *E. coli*: multiple pathways for multiple reasons. Cell, 58 : 807-809.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res., 23:23-35.

Smith D, Wohlgemuth J, Calvi BR, Franklin I, Gelbart WM. 1993. *hobo* enhancer trapping mutagenesis in *Drosophila* reveals an insertion specificity different from *P* elements. Genetics, 135:1063-1076.

Spradling AC, Rubin GM. 1982. Transposition of cloned p elements into *Drosophila* germ line chromosomes. Science, 218(4570):341-347.

Stahlberg A, Hakansson J, Xian X, Semb H, Kubista M. 2004. Properties of the reverse transcription reaction in mRNA Quantification. Clinical Chemistry, 50(3):509-515.

Stanley P and Okajima T. 2010. Roles of glycosylation in Notch signaling. Current Topics in Developmental Biology, 92:131-164.

Staudacher E. 2015. Mucin-type O-glycosylation in invertebrates. Molecules, 20:10622-10640.

Steller H. 1999. Rapid small scale isolation of *Drosophila* DNA and RNA. In: Rubin L, editor. The Rubin lab manual. Berkeley.

Stern DL, Crocker J, Ding Y, Frankel N, Kappes G, Kim E, Kuzmickas R, Lemire A, Mast JD, Picard S. 2017. Genetic and transgenic reagents for *Drosophila simulans*, *D. mauritiana*, *D. yakuba, D. santomea*, and *D. virilis*. Genes Genomes Genetics, 7:1339-1347.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays, 29(3):288-299.

Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. Genome Res., 21(7):1139-1149.

Taus T, Futschik A, Schlötterer C. 2017. Quantifying selection with pool-seq time series data. Molecular Biology and Evolution, 34(11):3023-3034.

Tearle R. 1991. Tissue specific effects of ommochrome pathway mutations in *Drosophila melanogaster*. Genet. Res., 57:257-266.

Teotónio, H, Chelo IM, Bradić M, Rose MR, Long AD. 2009. Experimental evolution reveals natural selection on standing genetic variation. Nature Genetics, 41(2):251–257.

Thanos D, Maniatis T. 1995. Virus induction of Human IFNβ gene expression requires the assembly of an enhanceosome. Cell, 83:1091-1100.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, Kaufman TC, Calvi BR and the FlyBase Consortium. 2019. FlyBase 2.0: the next generation. Nucleic Acids Res., 47(D1):D759–D765.

Tobler R, Franssen SU, Kofler R, Orozco-terWengel P, Nolte V, Hermisson J, Schlötterer C. 2014. Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. Mol. Biol. Evol. 31(2):364 - 375.

Tolstorukov MY, Jernigan RL, Zhurkin VB. 2004. Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. Journal of molecular Biology, 337:65-76.

Turner TL, Miller PG. 2012. Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. Genetics, 191:633-642.

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM. 2011. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genetics, 7(3):e1001336.

Umezu K, Kolodner RD. 1994. Protein interactions in genetic recombination in *Escherichia coli*. Journal of biological chemistry, 269(47):30005-30013.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evolutionary Biology, 7:226.

Von Hippel PH. 2007. From "simple" DNA-protein interactions to the macromolecular machines of gene expression. Annu. Rev. Biophys. Biomol. Struct., 36:79-105.

Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. 2002. Guidelines for human gene nomenclature. Genomics, 79(4):464-470.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in *cis* and *trans* gene regulation. Nature, 430 : 85-88.

Wittkopp PJ and Kalay G. 2012. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics, 13 : 59 - 69.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. Nature, 8:206-216.

Yamamoto-Hino M, Yoshida H, Ichimiya T, Sakamura S, Maeda M, Kimura Y, Sasaki N, Aoki-Kinoshita KF, Kinoshita-Toyoda A, Toyoda H, Ueda R, Nishihara S, Goto S. 2015. Phenotype based clustering of glycosylation-related genes by RNAi-mediated silencing. Genes to Cell, 20:521-542.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2014. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature, 518(7540):556-559.

Zabidi M, Stark A. 2016. Regulatory enhancer-core-promoter communication via transcription factors and cofactors. Trends Genet. 32(12):801-814.

Zachar Z, Bingham PM. 1982. Regulation of *white* locus expression: the structure of mutant alleles at the *white* locus of *Drosophila melanogaster*. Cell, 30:529-541.

Zhang L, Zhang Y, Ten Hagen KG. 2008. A mucin-type O-glycosyltransferase modulates cell adhesion during *Drosophila* development. The Journal of biological chemistry, 283(49):34076-34086.

Zhang X, Koolhaas WH, Schnorrer F. 2014. A versatile two-step crispr- and rmce-based strategy for efficient genome engineering in *Drosophila*. Genes Genomes Genetics, 4: 2409-2418.

Zhou D, Udpa N, Gersten M, Visk DW, Bashir A, Xue J, Frazer KA, Posakony JW, Subramaniam S, Bafna V, Haddad GG. 2011. Experimental selection of hypoxia-tolerant *Drosophila melanogaster*. PNAS, 108(6):2349-2354.

Figures and Tables

Figure 1
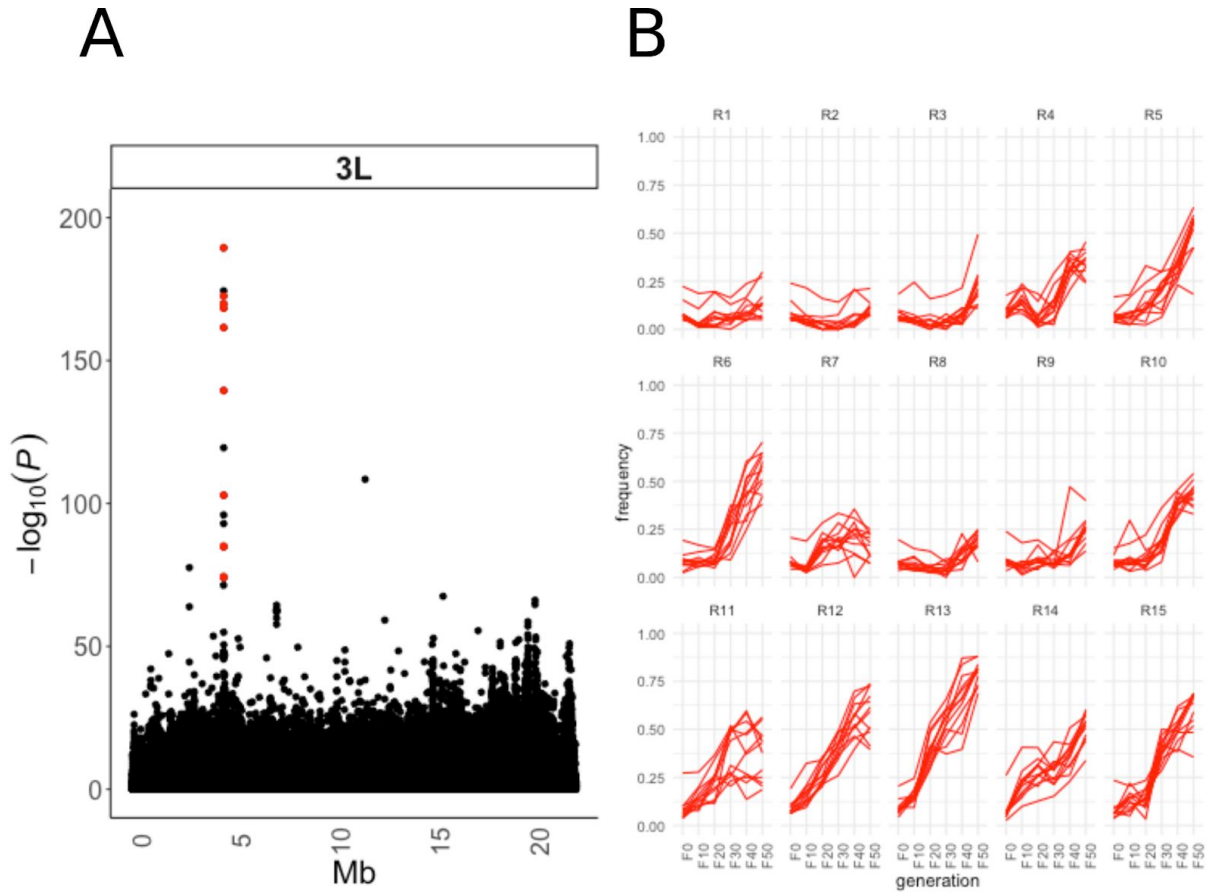
Figure 1 SNP cluster on 3L and trajectories of individual SNPs: (A) Manhattan plot showing the negative log10 transformed *p*-values from Cochran-Mantel-Haenszel test contrasting the ancestral (F0) with evolved (F50) populations, across positions on the chromosome 3L. The top-candidate, significant SNPs that were reconstituted as haplotype-block (*haploReconstruct*) are indicated in red. (B) Individual allele frequency trajectories of all top-candidate SNPs displayed over all 50 generations and across all replicates.

Figure 2
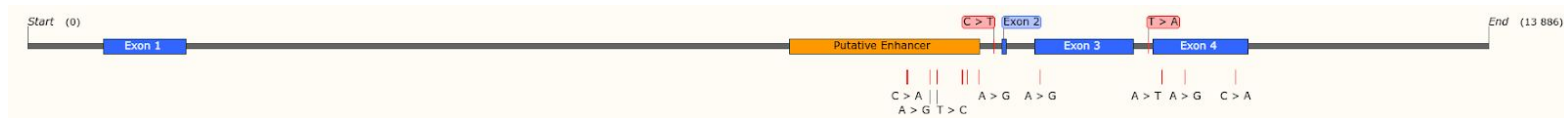
Figure 2 Approximately 15 kilobases (kb) of the genomic region illustrated, in which *GD13851* is located: The gene is depicted in light blue with yellow accentuation. All transcripts are portrayed, exons are marked by grey and pinkish boxes, introns indicated by a black line only. The Screenshot was obtained from FB2019_05, Thurmond *et al*. 2019, gBrowse: https://flybase.org/cgi-bin/gbrowse2/dsim/?Search=1;name=FBgn0185548 (Accessed: 23.11.2019, 09:00)

Figure 3

Figure 3 The locus of selection: The exons of *GD13851* are depicted in blue, introns are depicted as a grey line in 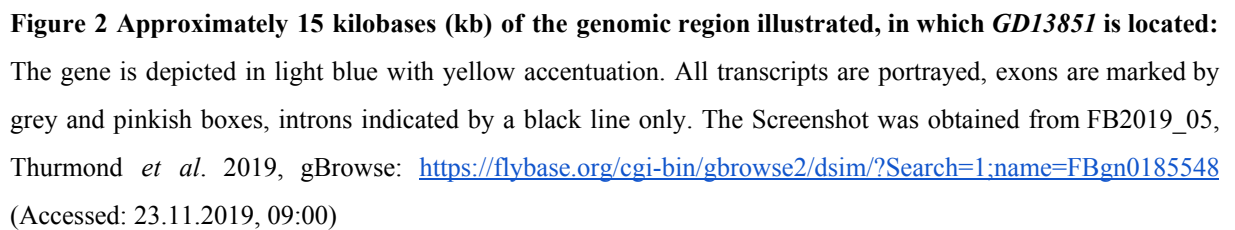between the exons. The region that was predicted by the software JASPAR (http://jaspar.genereg.net/) as an enhancer region is indicated in orange. The position of the SNPs that characterize both haplotype-blocks (ancestral and evolved) are shown in red, always in this order: "(ancestral > evolved)".

Figure 4
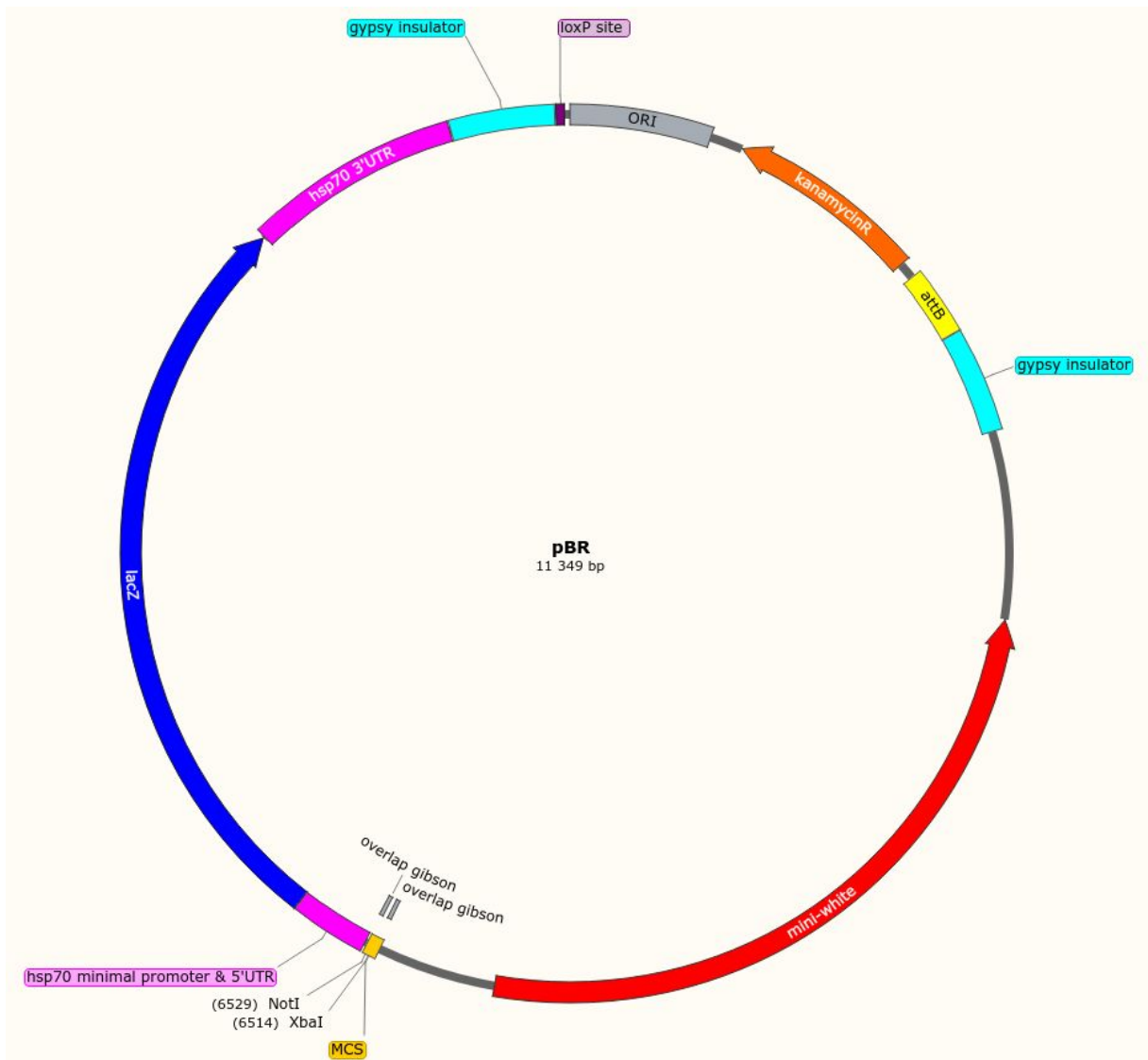
Figure 4 All elements of pBlueRabbit (pBR) annotated: origin of replication (ORI), kanamycin resistance gene (kanamycinR), φ-C31 bacterial attachment site (attB), *gypsy* insulator sites, *mini-white* marker, multiple cloning site (MSC), the recognition sites for the enzymes that were used for linearization (XbaI and NotI) are indicated, the sequence overlap in the MSC for Gibson assembly also shown (overlap gibson), *Hsp70* minimal promoter with 5'UTR, *β*-galactosidase gene (*lacZ*), *Hsp70* 3'UTR and the loxP site.

Figure 5

A
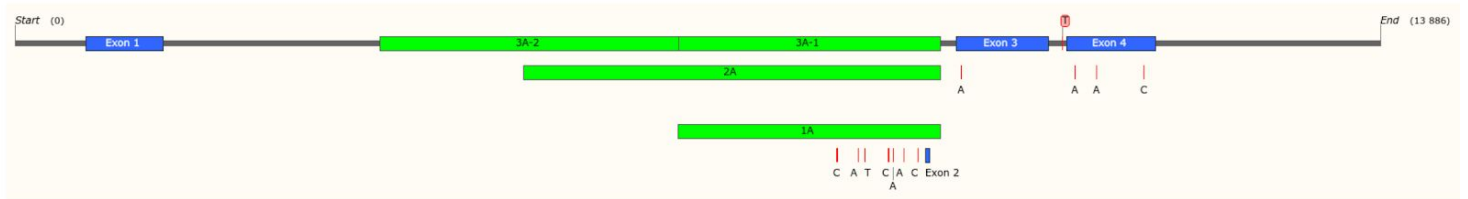
B
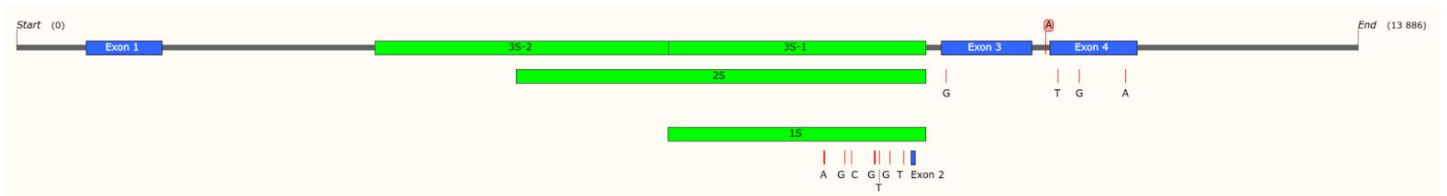
Figure 5 Amplification of *GD13851* intronic regions: (A) The ancestral haplotype-block of *GD13851* and (B) the evolved haplotype-block of *GD13851*. The exons are depicted in blue, the introns between the exons as black line, the position and the variant of the candidate SNPs are marked in red. All amplicons (in A: 1A, 2A, 3A-1, 3A-2 and in B: 1S, 2S, 3S-1, 3S-2) are shown in green. All of them contain the predicted enhancer region, and also the small exon number 2. The largest enhancer fragments (3A and 3S) were subdivided into two amplicons (3A-1, 3A-2 and 3S-1 and 3S-2) were ligated via Gibson assembly.

Figure 6

Figure 6 Electrophoresis of the amplicons from either haplotype-block: Agarose gel electrophoresis of amplicons amplified by PCR from genomic DNA of single flies from (A) isofemale line 46 (ancestral) or (B, C) isofemale line 42 (evolved). (B,C) The PCR from the evolved haplotype-block was performed twice because at first (B) only the fragment 3S-1 worked. At the second try (C) the remaining fragments were amplified (for 1S-PCR an aliquot of the first failed round of injection was used as input instead of genomic DNA). PCR and gel electrophoresis executed as described in the methods.

Figure 7
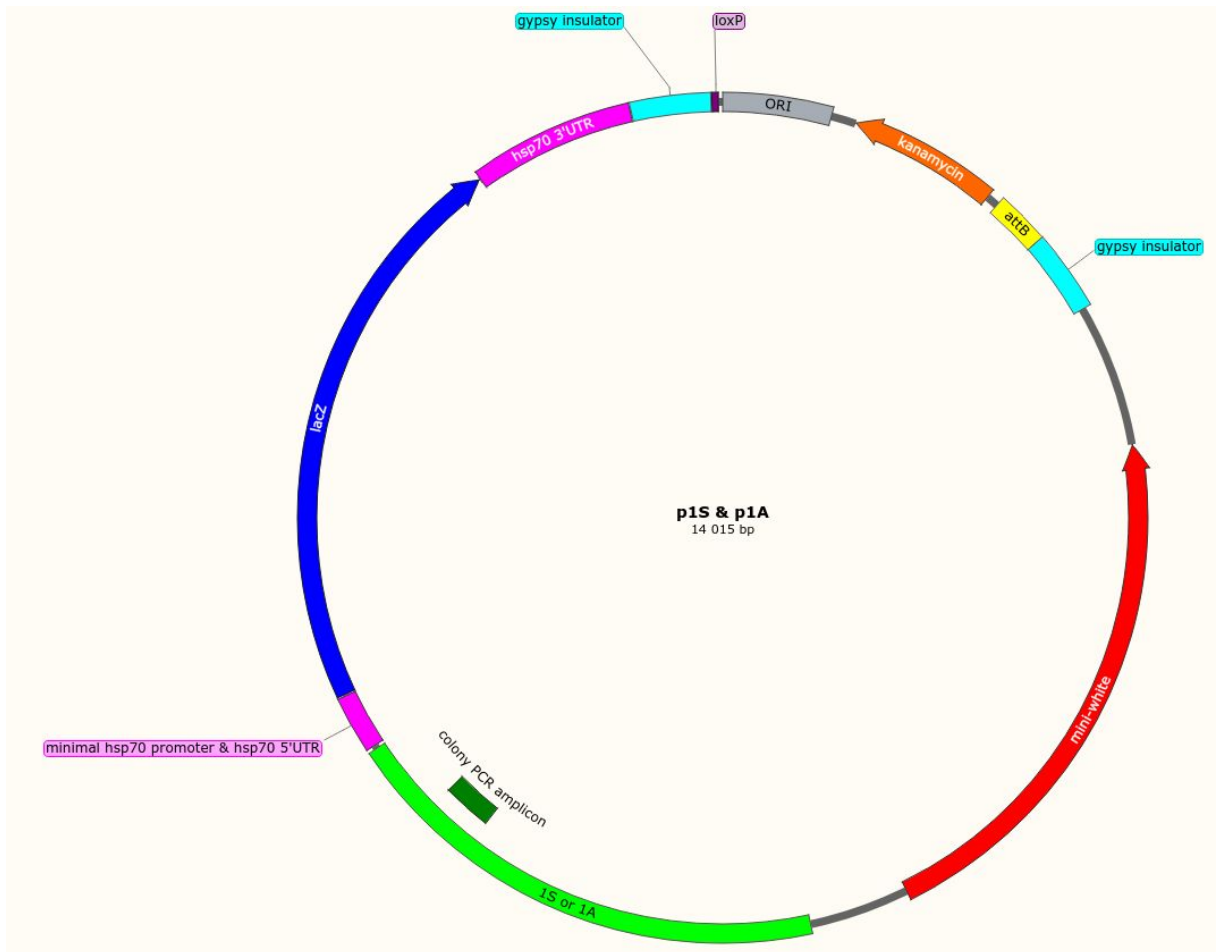
Figure 7 The plasmid p1S or p1A: The finished donor plasmid p1S or p1A after restriction cloning with the sequence fragment 1S or 1A (indicated in green) and pBR, respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.
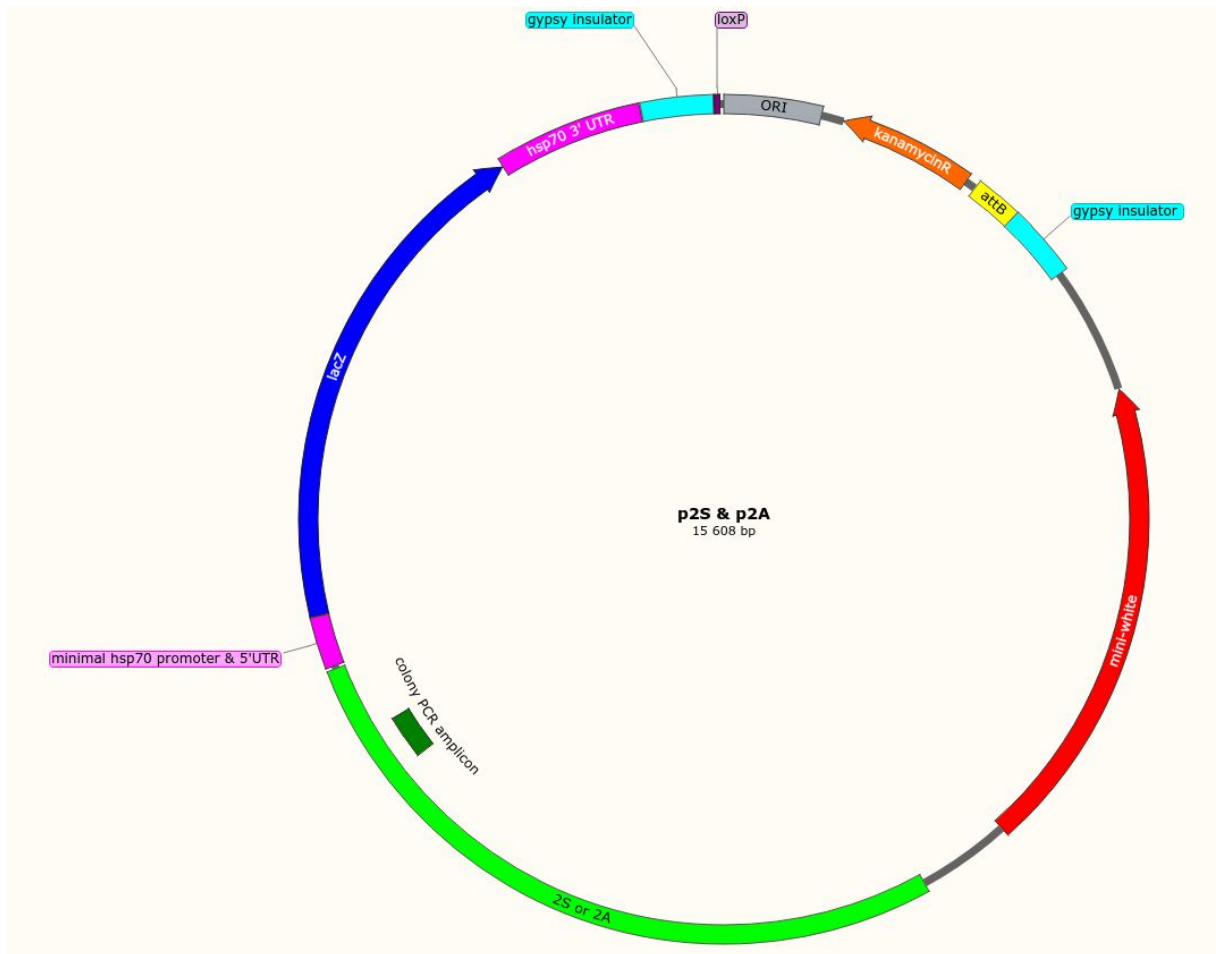
Figure 8

Figure 8 The plasmid p2S or p2A: The finished donor plasmid p2S or p2A after restriction cloning with the sequence fragment 2S or 2A (indicated in green), respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.

Figure 9
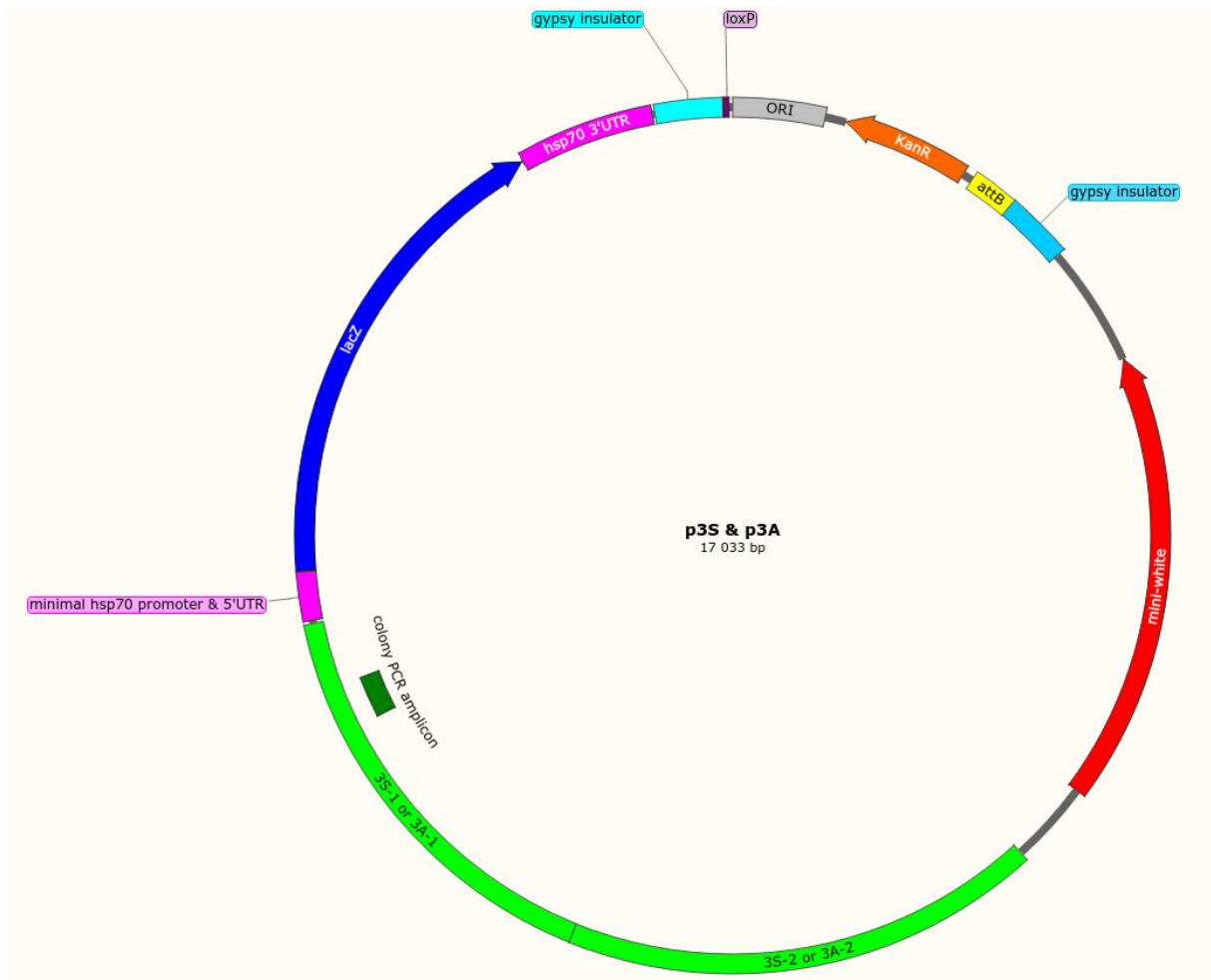
Figure 9 The plasmid p3S or p3A: The finished donor plasmid p3S or p3A after Gibson assembly with the sequence fragments 3S-1 and 3S-2 or 3A-1 and 3A-2 (indicated in green), respectively. The sequence marked in dark green is amplified by Colony PCR to confirm the presence of the putative enhancer insert in pBR.

Figure 10

Figure 10 An unexpected 2 kb band emerges in high frequency: (A, B) An unintended ~ 2 kb plasmid emerges in DH5α cells and top10 competent cells transfected with appropriate enhancer-reporter constructs. Agarose gel electrophoresis of the isolated plasmid content from 3 mL overnight cultures after digestion with HindIII and KpnI. Only 4 clones carry intelligible plasmids: (A) two colonies harbor the correct plasmid p1S (marked with green arrows), (B) two colonies carry the empty vector pBR (marked with red arrows). Cultivation, transformation, plasmid isolation, restriction digest and agarose gel electrophoresis performed as described in Material & Methods.

Figure 11

Figure 11 The product of potential intramolecular plasmid recombination between the *gypsy* sequences: This subfragment (marked in red, as "putative 2 kb fragment") would be 2400 bp in length, would contain the ORI (grey), the kanamycin resistance gene (green), the loxP (purple) and the attB (yellow) site.

Figure 12

Figure 12 The 2 kb fragment keeps appearing in *recA* deficient *E. coli*: Plasmid content isolated from transfected stable competent *E. coli* (NEB) 3 mL cultures was used for agarose gel electrophoresis either digested with KpnI and HindIII (A) or undigested (B). (A) The expected restriction fragment length is indicated in one lane for each of the three different enhancer-reporter constructs by green arrows. (A, B) The 2 kb contamination (indicated by red arrows) is still emerging additionally to the correct enhancer-reporter constructs, this is evident when the plasmid content is digested (A) or undigested (B). Cultivation, transformation, plasmid isolation, restriction digest and agarose gel electrophoresis was executed as described in Material & Methods.

Figure 13 The 2 kb contamination is not present or at much lower levels: NEB® Stable Competent *E. coli* clones, transfected with constructs obtained from gel-extraction (p1S, p1A, p3S) or isolated from overnight cultures (p2S, p2A, p3A) were used to inoculate 3 mL overnight cultures. (A, B) The isolated plasmid content from these cultures was digested (KpnI, HindIII), and used as input (~ 30 ng DNA per lane) for agarose gel electrophoresis. (B) the same electrophoresis as in (A) but with increased exposure. The 2 kb contamination is not detectable in any clone. The expected restriction fragment lengths are indicated in one lane by green arrows for each different construct. Only one colony (labeled p2S*) displayed aberrant restriction fragment lengths (unexpected fragment indicated by a red arrow) and was excluded from any further use. Cultivation, plasmid isolation, restriction digest and agarose gel electrophoresis performed as described in Material & Methods.

Figure 14

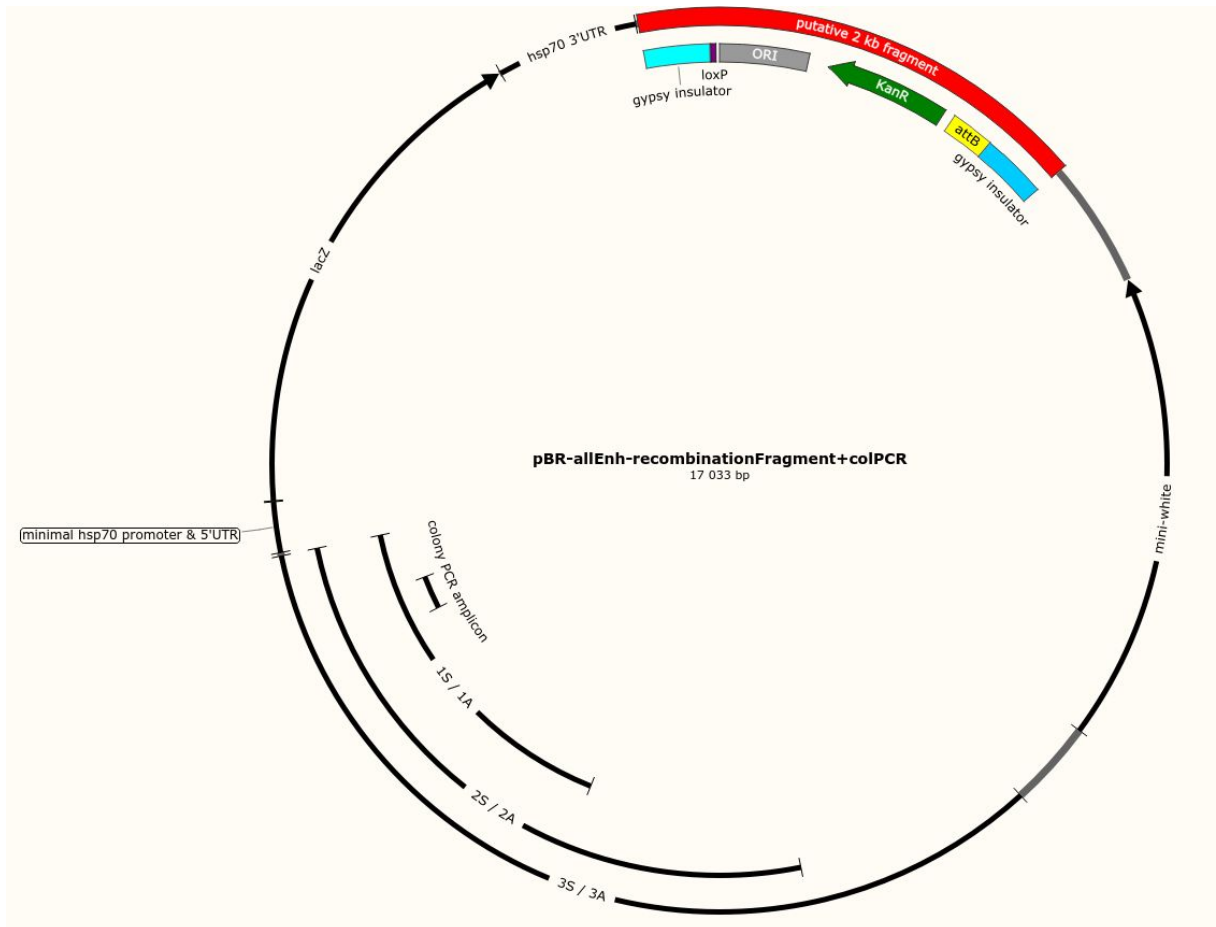**Figure 14 Prolonged cultivation re-established the presence of the 2 kb contamination:** (A and B) Isolated plasmid content from 50 mL cultures used as input (~ 300 ng DNA per lane) for electrophoresis, digested (first lane of 1S-3A) or undigested (second lane 1S - 3A). (A) and (B) show the same electrophoresis, (B) with higher exposure. The expected restriction fragment lengths are indicated in one lane by green arrows for each construct. The 2 kb contamination (indicated with red arrows) re-appeared in almost all cultures (1S, 1A, 2S, 2A, 3A). (C) The isolated, undigested plasmid DNA content from cultures that were used for inoculation, was used as input (~ 60 ng DNA per lane) for electrophoresis. Still, there is no contamination from the putative recombined "sub"-plasmid detectable. Cultivation, plasmid isolation, electrophoresis and digest performed as described in Material & Methods.

Figure 15

Figure 15 Flowchart of transformations with and propagation of the enhancer-reporter constructs: The workflow that led to the final enhancer-reporter vectors that were used for microinjections. Selection of clones that carry the plasmid pBR (with or without insert) was achieved via kanamycin containing media, screening for clones in which also the various enhancer inserts are present via colony PCR. A detailed description of each individual step is available in Material & Methods.

Figure 16

Figure 16 PCR to validate the attP site and examine for *lacZ* site presence in a fraction of the F1 generation: For the first seven lanes PCR reaction was performed using the primer pair "lacZFor" (GATACACTTGCTGATGCGGTGCTGATT) and "lacZRev" (CTGTAGCGGCTGATGTTGAACTGGAAG). The second position is empty. For the last three lanes, the primer pair "attP-For" (CCCAGGTCAGAAGCGGTTTTCG) and "attP-Rev" (TACGTGTCCACCCCGGTCACAA) was used. Control lanes: 1) the empty vector backbone of pBR (positive control *lacZ*), 2) w^{1118} (negative control for attP and *lacZ*), 3) *wOR(P) gypsy-lacZ*, from Sarot *et al.* 2004 (positive control *lacZ*), and 4) Bloomington #32218 (positive control attP). The F1 Fraction A, B, and C each contain pooled genomic DNA from approximately 60 F1 flies, those 60 derive from about 4 different injected G0 parents. The sample *Dsim*#2176 contains pooled genomic DNA from ~60 flies from the *Dsim*#2176 stock. We do not have an explanation for why the strain w^{1118} is *lacZ* positive, but we suspect contamination. This PCR validates that no silenced integration took place in the F1 flies fraction tested (see first seven lanes), and that the strain used for injections (*Dsim*#2176) harbors a endogenous attP site (see last 3 lanes).

Figure 17

Figure 17 The *GD23392* locus of *D. simulans* and the *vasa* locus of *D. melanogaster* show shared synteny: Display of the genomic locus of *GD23992* (top) and *vasa* (bottom) at the chromosome 2L of *D. simulans* and *D. melanogaster,* respectively. Genomic coordinates are plotted to the rulers. The *vasa* and the *GD23992* gene are highlighted in yellow. All surrounding protein-coding genes are displayed in light and dark blue. Genes that are classified by Flybase as syntenic orthologs are connected with a thin orange line. Information on orthology and screenshot obtained from FB2019_05, Thurmond *et al*. 2019, gBrowse. Screenshot processed via the open-source graphics editor Inkscape.

Figure 18

Figure 18 **The alignment of the *D. melanogaster* and *D. simulans vasa* region:** The query (labeled: Drosophila_melanogaster_vasa_promoter_&_5'UTR) is the complete *vasa* promoter and 5'UTR region that drives φ-C31 integrase expression in the plasmid vas-φ-C31(3xP3-EGFP). This sequence was blasted against the genome of *D. melanogaster* (Dmel, r6.31) (positive control) and *D. simulans* (Dsim, r2.02). The *D. melanogaster vasa* gene spans from 2L:15,061,656 to 2L:15,074,311. A large fragment with high analogy, and two smaller sequence fragments with intermediate analogy to the *vasa* promoter have been identified immediately upstream of the putative *D. simulans vasa* ortholog *GD23992* (*GD23992* located on Scf_2L:14,574,667..14,587,306). However, the alignment to *D. simulans* is discontinuous, contains large gaps and decreases in analogy in proximity to the TSS of the *GD23992*. Screenshot obtained from the tool BLAST from FB2019_05. The genomic start- and end-coordinates are added and irrelevant sub-alignments are veiled via the open-source graphics editor Inkscape.

Figure 19

>gnl|dsim|Scf_2L type=golden_path_region; loc=Scf_2L:1..23539531; ID=Scf_2L; dbxref=GB:CM002910; MD5=4db334c02c86dfa856dc1a48c595acf1; length=23539531; release=r2.02; species=Dsim;

HSP # = 2 , **Score** = 165.028 bits (83) , **Expect** = 1.19969e-38
Identities = 167 / 194 (86.1%) , **Positives** = 167 / 194 (86.1%) , **Gaps** = 2 / 194 (1%)
Strand = Plus / Plus

| GBrowse | | Subject FASTA |

```
vasa  promoter & 5'UTR:  1866   ATCACTTAGGTTGCTTGAATATTACAATTTCATTAATAGCTAAATCACATTTGATGTGTT   1925
                                |||||| |||| | |||||| ||||  ||||| ||||||||||||||||||||||||
Dsim\Scf_2L:  14574517          ATCACTAAGGTGGTTTGAACATTAGTATTTCGTTAATAGCTAAATCACATTTGATGTGTC   14574576

vasa  promoter & 5'UTR:  1926   AGTGGAAAACGGCTATATATATAAATTATCGAAATTGTGAATATCGAATTGCGATAGCAC   1985
                                ||||||||||||||| |||||  | |||||||||||| ||||| |  | | |||||||||
Dsim\Scf_2L:  14574577          AGTGGAAAACGGCTGTATATGGATATTATCGAAATTTTGAAAA--GTAATGCGATAGCAC   14574634

vasa  promoter & 5'UTR:  1986   AATGGGAAATTCCACCACTAGATTTTTGGTACTTTTAACAGATCCTTTTCGGTTTTGCGT   2045
                                |||||| ||||||||||||| | |  ||||||||||||||||||||||||||||||||||
Dsim\Scf_2L:  14574635          CATGGGATATTCCACCACTAGCTCTGGATTACTTTTAACAGATCCTTTTCGGTTTTGCGT   14574694

vasa  promoter & 5'UTR:  2046   TGCGCGAAGTGATC    2059
                                ||||||| ||||||
Dsim\Scf_2L:  14574695          TGCGCGAGGTGATC    14574708
```

Figure 19 The crucial sequence region of the *D. melanogaster vasa* promoter is covered in the alignment: This Figure shows a part of the alignment of the *vasa* promoter & 5'UTR of the plasmid vas-φ-C31(3xP3-EGFP) (indicated on the left: *vasa* promoter & 5'UTR) to the *D. simulans* genome (indicated on the left: Dsim\Scf_2L; r2.02) on sequence level (same alignment as in Figure 18). The shown sequence Dsim\Scf_2L: 14574517...14574708 constitutes the first of the two regions (from the left) in the alignment (Figure 18) with a key of 80-200 (first one of the two purple fragments in Figure 18). The 40 bp sequence that is reported to be crucial for germ-line specific transgene expression (Sano *et al*. 2002) is indicated in red. This 40 bp sequence fragment is conserved (38 of 40 bases are analog) in close proximity to the TSS of the putative *vasa* ortholog *GD23992*. Screenshot from FB2019_05, BLAST, Thurmond *et al*. 2019, default settings, low complexity filter off. via the open-source graphics editor Inkscape.

Table 1

Table 1. Devices and technical material

Material or Device	Manufacturer
Universal hood II, Molecular Imager Gel Doc XR system	Bio-Rad Laboratories Ges.m.b.H.
ProFlex base	Applied Biosystems, Thermo Fisher Scientific
CFX Connect Optics Module	Bio-Rad Laboratories Ges.m.b.H.
2720 thermal cycler	Applied Biosystems, Thermo Fisher Scientific
Centrifuge 5425	Eppendorf AG
PIPETMAN classic p2	Gilson Incorporated
PIPETMAN classic p10	Gilson Incorporated
PIPETMAN classic p20	Gilson Incorporated
PIPETMAN classic p100	Gilson Incorporated
PIPETMAN classic p200	Gilson Incorporated
PIPETMAN classic p1000	Gilson Incorporated
PIPETBOY acu 2	INTEGRA Biosciences GmbH
GD 100-p12	Grant Instrument™
PSC-20	Grant Bio™
MSH basic Yellow-line	IKA®-Werke GmbH & Co. KG
L13701 Shaking Incubator	GFL Gesellschaft für Labortechnik mbH
Universal lab incubator inb50	Memmert GmbH + Co. KG
Test-tube-rotator 34528	Snijder Labs
vortex-genie touch mixer 1	Scientific Industrie, Inc.
MW 2235 CW	Bomann
MP-250V Power Supply	Major Science
Horizon 58 Agarose Gel Electrophoresis Apparatus	Apogee Electrophoresis
Horizon 11-14	Life Technologies Inc.
Dr-36vl	CLF Plant Climatics GmbH
M-152 Manipulator	NARISHIGE Group
Femtojet	Eppendorf AG
Stereomikroskop ms5	Leica Mikrosysteme Handelsges.m.b.H.
Trinokulartubus	Leica Mikrosysteme Handelsges.m.b.H.
IMC 521234 Schieber Integrated Modulation Contrast	Leica Mikrosysteme Handelsges.m.b.H.
Kl 200, Cold Lightsource	SCHOTT AG
LaboStar® Ultra pure water and reverse osmosis systems	Evoqua Water Technologies GmbH
VX-120, Autoclave	Systec GmbH

MDF-U7386S, HCFC-Free Ultra-Low Temperature Freezer	SANYO Electric Co., Ltd.
Centrifuge 5804	Eppendorf AG
Centrifuge 5430R	Eppendorf AG
Centrifuge 5424 R	Eppendorf AG
multiply® - µStrip 0.2 mL chain, qPCR strip	SARSTEDT AG & Co. KG
8 Lid chain flat	SARSTEDT AG & Co. KG
PCR 8er-capstrips, farblos, doomed	Biozym Scientific GmbH
Pipette tip 200 µL gelb	SARSTEDT AG & Co. KG
Pipette tip 20µL farblos	SARSTEDT AG & Co. KG
Pipette tip 1000µL blue	SARSTEDT AG & Co. KG
CELLSTAR® Serological Pipettes 2 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 5 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 10 mL	Greiner Bio-One International GmbH
CELLSTAR® Serological Pipettes 25 mL	Greiner Bio-One International GmbH
96 Well Microplatte	Greiner Bio-One International GmbH
Embryo Collection Cage-Large	Genesee Scientific Corporation
Micro Tubes 1.5 mL Neutral	SARSTEDT AG & Co. KG
Safeseal micro tube 2mL PP	SARSTEDT AG & Co. KG
CELLSTAR® tubes 50 mL, PP, graduated	Greiner Bio-One International GmbH
CELLSTAR® tubes 15 mL, PP, graduated	Greiner Bio-One International GmbH
Test Tubes PS Round 100 x 16mm	Greiner Bio-One International GmbH
DURAN® Laboratory glass bottles	DWK Life Sciences GmbH
Erlenmeyer Flasks	KAVALIERGLASS, a.s.
Measuring cylinders	VWR International, LLC.
Inoculation Loops	SARSTEDT AG & Co. KG
Microscope slide, cut edges, frosted end	Carl Roth GmbH + Co. KG
Colour-fixed indicator sticks	Carl Roth GmbH + Co. KG
pH 1000L, phenomenal	VWR International, LLC.
Microloader Pipette tips, 20 µL	Eppendorf AG

Table 2

Table 2. Chemicals and Reagents

Reagent	Manufacturer	Art. No.
Formaldehyd	Carl Roth GmbH + Co. KG	CP10.1
Propanol	Carl Roth GmbH + Co. KG	CP41.4
Ethanol Absolute	Scharlab,S.L.	et00051000
PUFFERAN® TRIS	Carl Roth GmbH + Co. KG	AE15.1
ROTIPURAN® ortho-Phosphoric 85 %	Carl Roth GmbH + Co. KG	6366.1
LB Broth (lennox)	Carl Roth GmbH + Co. KG	X964.2
LB broth base (Lennox L Broth Base)	Invitrogen	12780-052
Peptone/Trypsin aus casein	Carl Roth GmbH + Co. KG	6681.1
Peptone	Carl Roth GmbH + Co. KG	2366.1
Glycerol	Carl Roth GmbH + Co. KG	3783.1
Agarose for gel electrophoresis	Carl Roth GmbH + Co. KG	3810.4
Ethidium bromide (1 % lsg in H2O)	Carl Roth GmbH + Co. KG	221.2
Acetic Acid	Carl Roth GmbH + Co. KG	3738.5
EDTA disodium salt dihydrate	Carl Roth GmbH + Co. KG	8043.2
LB-Agar (lennox)	Carl Roth GmbH + Co. KG	X965.2
Agar-Agar, Kobe I	Carl Roth GmbH + Co. KG	5210.5
Saccharose	Carl Roth GmbH + Co. KG	4621.2
100 % apfelsaft aus apfelsaftkonzentrat	Spar	
tryptone	Carl Roth GmbH + Co. KG	8952.3
Sodium chloride	Carl Roth GmbH + Co. KG	3957.2
Yeast extract	Carl Roth GmbH + Co. KG	2363.1
Potassium chloride	Merck KGaA	1.049.361.000
Bleach	DanKlorix	
Silica gel orange	Carl Roth GmbH + Co. KG	T199.4
Oil 10S, Poly(chlortrifluorethylen) 800, VOLTALEF®	VWR International, LLC.	24.627.188
Dry yeast, fermipan red, *Saccharomyces cerevisiae*	Casteggio Lieviti SRL.	
2- Propanol, ROTISOLV®	Carl Roth GmbH + Co. KG	7590.1
Tris-HCl	Carl Roth GmbH + Co. KG	9090.1
Sodium dodecyl sulphate	Carl Roth GmbH + Co. KG	2326.1
Potassium acetat	Carl Roth GmbH + Co. KG	4986.1
Diethyl pyrocarbonate	Carl Roth GmbH + Co. KG	K028.1
Heptane	Carl Roth GmbH + Co. KG	7725.2

Magnesium chloride	Carl Roth GmbH + Co. KG	KK36.1
Nipagin	Thermo Fisher Scientific Inc.	126960025
XbaI	New England Biolabs GmbH	R0145S
NotI - HF®	New England Biolabs GmbH	R3189S
KpnI	New England Biolabs GmbH	R0142S
HindIII	New England Biolabs GmbH	R0104S
CutSmart® Buffer	Solis Biodyne	B7204S
dNTPs	Solis Biodyne	02-21-00400
2.1 Buffer	New England Biolabs GmbH	B7202S
FIREPol® DNA Polymerase	Solis BioDyne OÜ	01-01-0500
10 x Reaction Buffer B	Solis Biodyne	01-01-0500
25 mM MgCl2	Solis Biodyne	01-01-0500
10x Solution S	Solis Biodyne	01-01-0500
100 bp DNA ladder	New England Biolabs GmbH	N3231S
1 kb DNA ladder	New England Biolabs GmbH	N3232L
Gel Loading Dye, Purple (6X)	New England Biolabs GmbH	B7024S
T4 DNA Ligase	New England Biolabs GmbH	M0202S
Q5® High-Fidelity DNA Polymerase	New England Biolabs GmbH	M0491S

Table 3

Table 3. Kits

Kit	Manufacturer	Art. No.
QIAquick PCR purification kit	QIAGEN GmbH	28104
GeneJet Plasmid Miniprep Kit	Thermo Fisher Scientific Inc.	K0502
Plasmid midi kit (100)	QIAGEN GmbH	12145
GeneJET Gel Extraction Kit	Thermo Fisher Scientific Inc.	K0692
NEBuilder® HiFi DNA Assembly Cloning Kit	New England Biolabs GmbH	E5520S

Table 4

Table 4. Solutions

Solution	Preparation
SOB	950 mL milliQ + 20 g tryptone + 5 g yeast extract + 0.5 g NaCl
SOC	200 μL Glucose + 50 μL MgCl2 + 10mL SOB medium
Nipogin mix	2.25 g Nipagin + 15 mL Ethanol absolute
Apple Juice Agar	13 g Agar Agar + 8 g Saccharose + 250 mL milliQ + 83 mL apple juice + 3 mL Nipogin mix
TAE Stock Solution 50x	242 g Tris base + 15.5 mL Phosphoric Acid + 100 mL of 0.5 M EDTA (pH=8)
TAe Solution 1X	100 mL TAE Stock Solution 50x + 4900 mL deionized H2O

Table 5

Table 5. Biological and genetic material

Organism	Supplier/Obtained from	Art. No. / Reference
NEB® Stable Competent *E. coli* (High Efficiency)	New England Biolabs GmbH	C2987I
Invitrogen™One Shot™ TOP10 Chemically Competent *E. coli*	Thermo Fisher Scientific Inc.	C404003
Invitrogen™Subcloning Efficiency™ DH5α Competent *E.coli*	Thermo Fisher Scientific Inc.	18265017
Dsim #2176 (*y*[1] *w*[1]; pBac{3XP3::EYFP-,attP}) (*Drosophila simulans*)	David Stern	Stern *et al.* 2017
Bloomington #32218 (*w*[*]; P{y[+t7.7] *w*[+mC]=10XUAS-IVS-mCD8::RFP}attP2) (*Drosophila melanogaster)*	Bloomington *Drosophila* Stock Center	
wOR(P) gypsy-lacZ (*Drosophila melanogaster*)		Sarot *et al.* 2004
w^{1118} (*Drosophila melanogaster*)	Vienna *Drosophila* Resource Center	
Plasmids	**Obtained from**	**Reference**
pBlueRabbit (pBR)	Ben E. Housden	Housden *et al.* 2012
p.vas-FC31(3xP3-EGFP)	Frank Schnorrer	Zhang *et al.* 2014

Table 6

Table 6. Primers

Primer	Sequence (5' to 3')
LS1	AGCGGCCGCAGCAGGTGTCAACAGGTCGC
LS2	ATCTAGAGCCCATTCGGGCCACGATAA
LS3	ATCTAGA CCGCCAGTCTGCGAAACTCA
LS4	CCGGATCCCCCGGTACCCGCAGCAGGTGTCAACAGGTC
LS5	GCCCATTCGGGCCACGAT
LS6	TTATCGTGGCCCGAATGG
LS7	CTTGGCTGCAGGTCGCGACTCAGTGAGCCCATCAAGGC
EnhColonyFor	TCGCGCACGTTTCTTATTGCG
EnhColonyRev	AACGCTGGCGACTTCTTGGG
lacZ-For	GATACACTTGCTGATGCGGTGCTGATT
lacZ-Rev	CTGTAGCGGCTGATGTTGAACTGGAAG
lacZ - seq	GTTCAATGATGTCCAGTGCAG
attP-For	CCCAGGTCAGAAGCGGTTTTCG
attP-Rev	TACGTGTCCACCCCGGTCACAA

Table 7

Table 7. PCR from genomic DNA for haplotype amplification

	Iterations	Temperature	Duration
Initial Denaturation	1 time	98°C	5 min
Amplification	35 cycles	98°C	10 sec
		66°C / 70°C	30 sec
		72°C	30 sec per 1 kb + 15 sec
Final extension	1 time	72°C	2 min
Hold		24°C	infinite

Table 8

Table 8. Amplicons & PCR components

Amplicon	PCR Components (DNA template + primers)	Amplicon size [bp]
1S	DNA isolated from single fly from isofemale line 42 + LS1 +LS2	2667
2S	DNA isolated from single fly from isofemale line 42 + LS1 +LS3	4240
3S-1	DNA isolated from single fly from isofemale line 42 + LS4 +LS5	2667
3S-2	DNA isolated from single fly from isofemale line 42 + LS6 +LS7	3056
1A	DNA isolated from single fly from isofemale line 46 + LS1 +LS2	2667
2A	DNA isolated from single fly from isofemale line 46 + LS1 +LS3	4240
3A-1	DNA isolated from single fly from isofemale line 46 + LS4 +LS5	2667
3A-2	DNA isolated from single fly from isofemale line 46 + LS6 +LS7	3056
Colony PCR	Transformed *E.coli* + EnhColonyFor + EnhColonyRev	288
Screening for lacZ	Pooled DNA of Dsim#2176 F1 flies (parents microinjected) + lacZ-For + lacZ-Rev	407
Screening for attP	Pooled DNA of Dsim#2176 flies + attP-For + attP-Rev	120

Table 9

Table 9. Enhancer-reporter constructs ligation components and protocol

Plasmid	Ligation components	Protocol
p1S	1S + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p2S	2S + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p3S	3S-1 + 3S-2 + pBR	NEBuilder HiFi DNA Assembly Reaction Protocol
p1A	1A + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p2A	2A + pBR	Ligation protocol with T4 DNA Ligase (NEB,M0202)
p3A	3A-1 + 3A-2 + pBR	NEBuilder HiFi DNA Assembly Reaction Protocol

Table 10

Table 10. Colony PCR

Colony PCR		temperature	time
Initial Denaturation	1 time	95°C	10 min
Amplification	25 cycles	95°C	30 sec
		56°C	30 sec
		72°C	30 sec
Final extension	1 time	72°C	1 min
Hold		24°C	infinite

Table 11

Table 11. PCR from pooled genomic DNA

	Iterations	Temperature	Duration
Initial Denaturation	1 time	98° C	10 min
Amplification	38 cycles	95° C	30 sec
		62° C	30 sec
		72° C	45 sec
Final extension	1 time	72° C	7 min
Hold		24°C	infinite

Table 12

Table 12. Statistics of microinjections, from the top to the bottom: Record of donor and helper plasmid concentrations, total number of embryos injected, total number of larvae harvested & the percentage of injected embryos that reached larval stage, total number of flies that developed into adult flies & the percentage of larvae that developed into adults, the number of female adult flies, number of male adult flies, and the number of sterile male adult flies & percentage of sterility among males. Because female injected G0 flies were always pooled in groups of 5, it was not accessible if individual adult females were sterile.

	p1S	p1A	p2S	p2A	p3S	p3A	p1S repeat
c_{donor} [ng/µL]	250	250	250	250	250	250	250
c_{helper} [ng/µL]	190	211	200	233	240	242	190
embryos injected (#)	250	381	535	420	345	444	400
larvae yield (# / %)	118 / 47	127 / 33	134 / 25	53 / 13	65 / 19	75 / 17	111 / 28
adults (# / %)	58 / 49	71 / 56	83 / 62	25 / 47	19 / 29	41 / 55	69 / 62
♀ adults (#)	24	25	35	11	5	16	28
♂ adults (#)	34	46	48	14	14	25	41
♂ sterile (# / %)	19 / 56	30 / 65	30 / 63	6 / 43	8 / 57	13 / 52	26 / 63
eyes (white/red)	58/0	71/0	83/0	25/0	19/0	41/0	69/0