pagepress

Evolutionary dynamics of simple sequence repeats across long evolutionary time scale in genus *Drosophila*

Lothar Wissler, Lars Godmann, Erich Bornberg-Bauer

Institute for Evolution and Biodiversity, University of Muenster, Germany

Abstract

Repetitive DNA is among the fastest evolving types of genomic DNA, which includes simple sequence repeats (SSRs), short regions of tandemly repeated one to six nucleotides long motifs. SSRs are found most frequently in non-coding regions. Repeat number variation occurs rapidly and is presumably neutral such that polymorphic SSRs are frequently used as genetic markers to characterize and classify populations. Despite their rapid evolution, recent reports suggested that SSR *loci* can be retained over hundreds of millions of years. We here investigate the dynamics and genomic features of SSR evolution in syntenic regions conserved across twelve *Drosophila* species and within a *D. melanogaster* population dataset. We find that SSR *loci* decay exponentially with time, the percentage of retained SSRs mostly reflects species relationships and correlates well with the sequence similarity of neighboring genes. About 47% of repeat *loci* within syntenic regions may share common ancestry due to predicted conservation in at least two species from the *Drosophila* subgenera *Sophophora* and *Drosophila* respectively, *i.e.* after 80 million years of divergence time. Since *loci* which are highly polymorphic at the population level also decay faster across species, SSR evolution appears to be a gradual process in which conservation pressure may act at relatively constant rates across time scales. A higher proportion of SSR *loci* are retained among *Drosophila* subgenus species considering their evolutionary distance and the expected decay rate estimated across all *Drosophila* species. This prolonged SSR retention might be caused by a higher SSR mutation rate and a lower nucleotide substitution rate in the *Drosophila* subgenus compared to *Sophophora* species. SSRs in exons and on autosomes evolve more slowly than SSRs located outside of exons or on the sex chromosome, respectively, both within and across species. SSR variability and phylogenetic conservation thus varies depending on the genomic location. These findings provide new insights into the dynamics of SSRs at both micro- and macro-evolutionary scales. The development of robust models of SSR long-term evolution will facilitate more in-depth analyses in general and the prediction of neutrally evolving SSRs and SSRs evolving under purifying selection, extending our knowledge of the functional impact of SSRs in genome evolution.

Introduction

Simple sequence repeats (SSRs, also known as microsatellites) are a class of tandemly repeated DNA sequences which occur in all known taxa and typically represent significant fractions of the overall genomes.[1] For example, 3% of the human genome is made up of SSRs,[2] and between 1.1% and 4.3% were observed in *Drosophila* species. In the most widely used definition, the number of repeat units within an SSR can range from a few to several hundred repeats, with repeating units of 1 to 6 nucleotides of length.[3,4] The major mechanism creating SSR length variation is assumed to be through replication slippage.[5-7] Observed mutation rates of SSRs per locus and generation range between 10^{-6} and 10^{-2} which is much higher than the substitution rates observed in non-repetitive eukaryotic DNA (between 10^{-10} and 10^{-8}).[8-11] Many studies have investigated mechanistic properties of SSR evolution, but despite these efforts, some processes such as SSR birth and death,[7,12-14] as well as the influence of point mutations, recombination, and transposition on SSR length and the genomic SSR content are still not fully understood.[4,7,15,16]

SSRs have become one of the most popular types of genetic markers used in a variety of genetic analysis techniques. Many of their applications deal with intra-species analyses in population genetics,[17] for example testing geographic origins of invasive species.[18] Further applications include gene mapping, genetic maps and association studies, conservation biology, molecular anthropology, and paternal investigation.[1,14,19] Common to many SSR applications is the underlying assumption of neutrality, *i.e.* that repeat number variation and interrupting mutations in an SSR haves no fitness effect and areis a purely stochastic processes.[20-22] The assumed lack of selection pressure on SSR retention suggests that SSRs are not retained over long evolutionary time scales and thus not across different species. Accordingly, the transfer of SSR markers is particularly difficult between distant species.[23,24]

In the last 10 years, SSRs received increased attention in medical research due to accumulating evidence associating them with various cancers,[25,26] and the implication in several dozen human hereditary disorders.[27-29] While the majority of tandem repeats is located in the non-coding portions of the genome, presumably without any function, many SSRs are located within genes and regulatory regions.[30] SSRs in coding sequences, mostly triplet repeats with repeating motifs of 3 or 6 nucleotides, typically translate into repeats of amino acids, most prominently glutamine (*Q*) repeats. Such amino acid repeats are assumed to form intrinsically disordered structures and are frequently found in highly connected proteins (such as transcription factors and protein kinases) where repeats might serve as binding interfaces.[31-34] Some of these gene-associated repeats may even be selected for as they do not show signs of selection pressure against unstable repeats and can evolve independently in homologous genes.[30,35,36] Numerous case studies have revealed that repeat variation can influence phenotypes, *e.g.* by modulating gene expression, binding interfaces, or chromatin, DNA, and RNA structure.[30] All these findings, many of which are not confined to non-coding regions, suggest that not all SSRs fit the assumption of neutrality.

Furthermore, recent reports have suggested that SSR *loci* can be conserved across

Correspondence: Erich Bornberg-Bauer, Institute for Evolution and Biodiversity, University of Muenster, Huefferstrasse 1, D48149 Muenster, Germany.
Tel. +49.251.8321630 - Fax: +49.251.8324668.
E-mail: ebb@uni-muenster.de

Key words: genome evolution, simple sequence repeats, SSRs, microsatellites, *Drosophila*.

Contributions: this study was conceived by EBB and designed by all authors. LG performed pilot studies, LW all final computations and figures. EBB and LW wrote the paper.

Acknowledgments: LW and EBB acknowledge support from the Volkswagen Foundation. We thank Philine Feulner and Jenny Gu for helpful discussion, Bartholomäus Wieczorek for an early pilot study on ESTs and Till Bald for help with analyzing population data. We also thank two anonymous reviewers for insightful comments.

Conflict of interests: the authors report no potential conflict of interests.

Received for publication: 3 May 2012.
Revision received: 20 June 2012.
Accepted for publication: 21 June 2012.

Trends in Evolutionary Biology 2012; 4:e7
doi:10.4081/eb.2012.e7

genomes over 450 Myr.[22] Considering SSR abundance, it is not necessarily surprising to find some SSR *loci* retained over a long time scale. Mechanistically, mutations in viable SSRs favor their preservation, *i.e.* imperfections in SSRs can get purged during DNA replication slippage.[8] It is, however, not known how purging of interrupting mutations and repeat number variation in SSRs shapes SSR retention rates over several million years. Accordingly, accurate models which can quantitatively describe the evolution of SSR *loci* and the expected results under neutral evolution over macroevolutionary times have not been established, yet. To date, only one study has systematically evaluated the full-genome conservation of SSRs: Buschiazzo and Gemmell studied the conservation of human SSR *loci* in 11 vertebrate species using full-genome alignments present in the UCSC Genome Browser.[22,37] They could show that conservation of human SSR *loci* declines exponentially with increasing divergence time and suggested that the majority of SSRs in genomes are maintained by chance. Accordingly, we here take advantage of the recently published wealth of genomes and address issues associated to the possible stochastic nature and retention of SSRs across species. Common to all our analyses is the identification of conserved SSR *loci*. With an SSR locus, we refer to a putatively homologous position in different genomes; if an SSR with a minimum length and the same repetitive motif can be found in multiple genomes at the same locus, we infer that this SSR locus has been conserved. For this evolutionary analysis, we have developed a novel method that predicts homologous SSRs between pairs of species.

The presented study addresses three key issues: first, we ask if and how frequently SSRs are conserved over evolutionary long time scales. Since SSRs are highly polymorphic, their state in any given genome, such as the *model genome* which represents a species, are only snapshots of the current population. It is not clear if such a state has already been fixed in the population, or if similarities reflect common ancestry and conservation or different ancestry and convergent evolution (*homoplasy*). Therefore, we investigate SSR retention in well defined *loci* between pairs of orthologous genes. To account for the long evolutionary distances among *Drosophila* species, we focus largely on SSR locus retention as a less sensitive but more robust measure of conservation relative to the change in the repetitive sequence. The *Drosophila* genus with 12 species that split between 1 and 40 million years ago represents an ideal data set to study gradual changes of *loci* over a wide range of evolutionary time scales.[38,39] It also allows to test whether previous findings from vertebrates are valid in *Drosophila* because in *Drosophila*, SSR evolution is assumed to be much slower.[19] Second, we compare the retention of specific SSRs to the degree of polymorphism from 37 individual genomes of the well-studied model organism *D. melanogaster* (http://dpgp.org/) in order to determine if population based volatility correlates with cross-species retention. Third, we compare retention rates between different genomic features (exons, introns, and intergenic regions) and between sex chromosome and autosomes. These comparisons address the question whether SSR conservation depends on the genomic localization of the repeat locus.

Materials and Methods

Dataset

For cross-species analyses, sequence data and gene feature files (GFFs) for each of the twelve *Drosophila* species were obtained from the FlyBase ftp server (ftp://ftp.flybase.net, release FB2010_02).[40,41] Species and abbreviations are: *D. ananassae* (Dana), *D. erecta* (Dere), *D. grimshawi* (Dgri), *D. melanogaster* (Dmel), *D. mojavensis* (Dmoj), *D. persimilis* (Dper), *D. pseudoobscura* (Dpse), *D. sechellia* (Dsec), *D. simulans* (Dsim), *D. virilis* (Dvir), *D. willistoni* (Dwil), and *D. yakuba* (Dyak). For each species, whole chromosome sequences were scanned for SSRs. Genomic features were derived from GFFs, classifying SSR localization into either *exonic, intronic*, or *intergenic*. For intra-species analyses, the population dataset of *D. melanogaster* was obtained from http://dpgp.org/ using the DPGP assemblies (release 1.0). Out of the 50 genomes, only Trudy Mackay's set of 37 inbred lines sampled in Raleigh, NC were used as they all cover the 2L, 2R, 3L, 3R, and X chromosomes of *Dmel*.[42] Genomic features were transferred from the FlyBase *Dmel* GFF since these population genomes are aligned to the *Dmel* reference genome.

For a comparison of SSR lengths between Drosophilidae and mammals, unmasked genome DNA sequences were obtained from the Ensembl ftp server (release 64, ftp://ftp.ensembl.org) for the following mammalian species:[43] *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Sus scrofa* (pig).

Simple sequence repeat identification

SciRoko version 3.4 was used to identify perfect and imperfect SSRs in DNA sequences using its mismatched fixed penalty mode with default settings (see Supplementary File 1 for detailed settings).[44] Accordingly, identified SSRs have a minimum length of 12 nt given a perfect repeat structure, while interrupting mismatch mutations are penalized and require repeats to be longer (*e.g.* 16 nt in case of one mismatch mutation).

Ortholog identification

Orthologous relationships between genes were obtained from FlyBase precomputed files resource (ftp://ftp.flybase.net/releases/FB2010_02/precomputed_files/). The *FlyBase melanogaster gene ortholog report* lists *Dmel* genes and their orthologs in other sequenced strains. When a *Dmel* gene maps to one gene in all other species, we assume that all these genes are orthologous to each other. A total of 7655 one-to-one orthologous gene clusters were thus determined.

Syntenic regions

Syntenic regions that are shared by all 12 species were identified with the tool OrthoCluster using the build from December 2007.[45,46] Using the 7655 orthologous gene clusters (see above) and the gene feature files of the annotated chromosomes, these regions were identified where at least two orthologous genes are direct neighbors to each other and where the order and their strandedness is conserved across all twelve genomes (see Supplementary File 1 for detailed settings). These settings yielded 1139 clusters of syntenic regions that included 2 to 7 genes.

Calculating simple sequence repeat locus conservation

To determine whether or not a homologous SSR locus exists in a pair of species, the following requirements have to be met: they share the same repeating motif and occur at the same relative position in the respective genome. In a novel approach, we predict homologous SSRs between two species based on four criteria: i) they occur in the same microsynteny block, ii) they have exactly the same repeating standardized motif (as predicted by SciRoko), iii) they occur in the same gene feature (*i.e.* exon, intron, or intergenic), and iv) they occur at the same position relative to other SSRs in the microsynteny block (see Supplementary File 1 for more details). If two or more SSRs from different organisms fulfill these criteria, we predict that these are homologous SSRs and represent a *conserved SSR locus*. Between two genomes, we will determine *C*, the rate of SSR locus conservation, as the percentage of conserved *loci* among all repeat *loci* of the two species. Note that two SSRs at the same locus are considered to be homologs independently of their *states*, *i.e.* their number of repeat units, thus allowing for variation in length and mismatch mutations. Two methods that differ in sensitivity and specificity were developed to identify such conserved SSR *loci* between two species: *simple pairwise* and *progressive*. In both methods, all SSRs within a syntenic region of a species are translated into an ordered list of strings, where for each SSR the string is composed of the repeating motif and the type of genomic feature it is localized in. SSR locus conservation in the *simple pairwise* method is then based only on a global pairwise alignment of two species' lists using an implementation of the Needleman-Wunsch algorithm.[47] For each match, mismatch, or gap, we used the scores +2, -10, and -1, respectively. In the resulting global alignment, identical strings are aligned and represent a conserved repeat locus, whereas SSRs that do not meet the conservation requirements as outlined above are matched with a gap and represent repeat *loci* that were gained or lost in either of the two species. Note that due to a higher penalty for mismatches relative to gaps, we prevent non-homologous SSRs being aligned to each other. C_{sim}, the rate of SSR locus conservation measured by the *simple pairwise* method, is the fraction of the number of conserved SSRs to the number of all SSRs of the two species.

In the *progressive* method, the conservation of SSR *loci* is evaluated using the phylogenetic position of the species pair in the species tree and by performing progressive alignments as follows. Given three species A, B, and C and a species tree of ((A, B), C), repeat *loci* between A and C can only be conserved if the same locus has been found conserved between the more closely related species A and B. Technically, we start at the leaves A and B of the tree and visit internal ancestral nodes until we arrive at the node of their most recent common ancestor. At each visit to an ancestral node, we define its SSRs as all SSR *loci* which have been conserved between the two child nodes as described above. Accordingly, C_{pro}, the rate of SSR locus conservation determined by the *progressive* method, is the fraction of conserved SSRs at the most recent ancestral node to the sum of all SSRs of the two species.

The *simple pairwise* and the *progressive* methods will produce the same results for sister species, *i.e.* species A and B given the tree ((A, B), C). However, in all other cases, the *progressive* method requires that an SSR is found conserved in *all* child nodes of their most recent ancestral node for an SSR to be considered conserved.

For quality control, we have additionally measured sequence conservation of flanking regions (50 bp up- and downstream) of pairs of conserved SSRs (determined by simple pairwise). On average, for any pair of matched SSR *loci*, 51% sequence identity (±17%) was found in between the flanking regions, indicating that even our simple pairwise method, which is prone to homoplasy, has a high success rate for identifying conserved homologous SSR *loci*. A list of 26,686 pairs of conserved SSR *loci* with higher than 50% sequence identity in the flanking region is provided as Supplementary File 2.

Taring C_{sim} and C_{pro}

Both SSR locus conservation rates, C_{sim} and C_{pro}, were tared to account for the chance of aligning non-homologous SSRs. For each species pair, a null model for SSR locus conservation was created in which SSRs were randomly sampled from the genomes (see Supplementary File 1). The simple pairwise and progressive method used on this dataset produced C_{sim_rand} and C_{pro_rand} which represent the rate of matching non-homologous SSRs and thus the fraction of expected false positive SSR *loci* considered as conserved. With the rate of expected false positives known, rates of true SSR locus conservation can be derived as follows: $C_{sim\_true}=C_{sim}$ - C_{sim_rand}, and $C_{pro\_true}=C_{pro}$ - C_{pro_rand}.

Relating simple sequence repeat conservation rates to divergence times

For all tested species pairs, [*i.e.* (*A,B*), (*A,C*), (*B,C*)...] rates of SSR locus conservation (C_{sim} and C_{pro}) were mapped to their divergence time *d*. *d* is twice the distance from one child node to the most recent common ancestor node and thus represents the time span each of the two genomes could diverge from each other. Adopted time estimates of species splits provide a resolution of 2 to 80 million years.[48]

Comparing states of conserved simple sequence repeat *loci*

The *states* of conserved SSR *loci*, *i.e.* the exact sequence at pairs of homologous SSRs, were classified into four categories as follows: i) *perfect:* both SSRs have the same state, *i.e.* identical repetitive sequences (including the same number of repeat units and possibly interrupting point mutations); ii) *interrupted:* both SSRs have the same length but differ in the number of interrupting mutations, *i.e.* single point mutations that interrupt perfect tandem repeats; iii) *variable:* both SSRs are not interrupted but vary only in the number of repeats; iv) *similar:* any pair of conserved SSR *loci* that did not match any of the three previous categories is categorized as *similar*. Thus, *similar* SSRs at conserved *loci* are characterized by both differences in interrupting mutations and number of repeats.

Assessing sequence conservation of orthologous protein sequences

All 1:1 orthologous genes with conserved synteny, *i.e.* all proteins that were used to construct the syntenic regions conserved across all 12 *Drosophila* species, were extracted from the OrthoCluster output. For each pair of species, the protein sequences of these 1:1 orthologs were globally aligned with MUSCLE and all alignments were concatenated.[49] The overall conservation rate of a species pair was then computed as the percentage of identical residues in the concatenated alignment.

Construction of the simple sequence repeat neighbor joining tree

The similarity matrix based on all pairwise C_{sim} values was converted into a distance matrix using R's dist function. A Neighbor Joining tree was then generated with NJ (Neighbor Joining Tree Estimation) as implemented in the R package ape.[50,51] The NJ tree will be used to test whether the accepted species tree (Figure 1A) can be recovered using SSR locus conservation rates between species.

Simple sequence repeats within a *D. melanogaster* population

SSRs in the DPGP dataset were identified identically to the cross-species analysis using SciRoko with the afore mentioned settings. On the DPGP dataset, the following three analyses were performed: i) SSR locus and state conservation rates were determined for all repeat *loci* that are located within the same syntenic regions which have been found conserved across all 12 *Drosophila* species. Since the DPGP genomes are already aligned, conserved SSR *loci* were identified as significant SSRs with the same standardized motif at overlapping positions. Otherwise, the exact same method was applied as used in the cross-species analysis; ii) those SSR *loci* that were found conserved at least among *Dmel*, *Dsim*, and *Dsec*, were sampled from the DPGP dataset. The distribution of two characteristics were compared between all repeat *loci* and the subset of *loci* conserved across *Drosophila* species: a) the number of individuals in which a significant SSR was present at each of these *loci*, and b) the number of different states at these *loci* within the population. Two-sample Kolmogorov-Smirnov tests were used to test whether the distribution of cross-species conserved *loci* is significantly different from the distribution of all SSR *loci*; iii) for the comparison of SSR conservation between sex chromosome and the autosomes, all genomic SSR *loci* were used. Conserved SSR *loci* were identified throughout the X chromosome (representing the sex chromosome in *D. melanogaster*) and 2L, 2R, 3L, and 3L chromosomes representing the autosomes. *Loci* in these two sets were then tested for differences in the population spread (see ii a) and allelic diversity (see ii b).

Results and Discussion

Genome-wide distribution of simple sequence repeat *loci*

Across the twelve *Drosophila* genomes (Figure 1A), a total of 2,138,597 SSRs were identified. Despite very recent species splits, some *Drosophila* genomes differ quite strongly in their total number of SSR *loci* (Figure 1B), and this trend persists even after transforming genomic SSR counts into *density* values (number of SSRs per Mb) to account for differences in genome size (Figure 1C). Hierarchical clustering based on genome-wide SSR densities indicates that the twelve *Drosophila* species can be separated into three groups (see Supplementary File 1 for details): the first group consists of all six species from the *melanogaster* group, all having relatively low SSR densities. The second and third groups show much higher SSR densities and are more similar to each other than to the *melanogaster* group. The highest density is observed in *D. mojavensis*, which is the only representative of the third group. The second group is categorized by intermediate SSR densities and comprises the remaining 5 species, including both *obscura* group species, *D. willistoni*, and the two Drosophila subgenus species *D. grimshawi* and *D. virilis*. These inferred groups are only partially congruent with the established phylogeny among the *Drosophila* species, and we predict that genomic SSR densities and thus the simple sequence repeat content in the genomes of different species can change rapidly. If such a change can also occur at variable rates, it may explain the contradictions between the clustering and the phylogenetic tree. The groupings of *Drosophila* species based on their genome-wide SSR densities agrees with an earlier study investigating the abundance of amino acid repeats, which are typically encoded by tri-nucleotide SSRs. This study reported the smallest repeat content in orthologous protein-coding genes from species of the *melanogaster* subgroup relative to the other *Drosophila* species.[52] With increasing distance to the melanogaster subgroup, repeat content becomes more variable, but generally increases and is highest in three Drosophila subgenus species and *D. willistoni*.[52] These findings suggest that genomic SSR content is variable and not necessarily limited to non-functional parts (*junk DNA*) of the genome. Therefore, a fraction of SSRs may play a functional role in genome evolution.

In order to address the direction of SSR density change, SSR density was evaluated in five outgroup species to the 12 *Drosophila* species (*Anopheles gambiae*, *Aedes aegypti*, *Culex quinquefasciatus*, *Bombyx mori*, *Tribolium castaneum*). These outgroup species show much lower full-genome SSR densities (see Supplementary Material), and it is thus likely that the ancestral genome had relatively low SSR densities. Although we cannot rule out that the tested outgroup species all have undergone genome changes that led to a decrease in SSR density, these data probably indicate a significant increase in SSR density throughout the *Drosophila* clade. Moreover, our data provide systematic evidence for the evolution towards higher SSR densities in the *Drosophila* subgenus compared to the *Sophophora* subgenus as already suggested by previous studies.[53-58]

Although more closely related *Drosophila* species are more similar in SSR density to each other than to more distant species, differences in SSR density are not always proportional to the divergence time. For instance, both species pairs *Dsim-Dsec* and *Dere-Dyak* have roughly the same estimated divergence time of 2 Myr, but while *Dsim* and *Dsec* are highly similar in their SSR densities (546.6 to 593.5 SSRs/Mb), *Dere* and *Dyak* are more strongly divergent (516.5 to 610.8 SSRs/Mb). Finally, observed genome-wide SSR densities suggest that the number and density of repeat *loci* is variable even between recently split species (<5 Myr).

All these global comparisons based on genome-wide SSR abundance and density indicate that SSR gains and losses occur frequently in *Drosophila* genome evolution. The following analyses address the evolutionary dynamics of SSRs in more detail and investigate within- and between-species conservation of orthologous SSR *loci*. Since the genome-wide identification of an orthologous SSR locus among multiple species is not trivial, all of these in-depth analyses are restricted to syntenically conserved regions of the *Drosophila* genomes.

Between pairs of syntenic regions, the conservation of SSR *loci* was evaluated among all *Drosophila* species using a novel approach that relies on globally aligning string representations of SSRs found in syntenic regions (see Methods and Supplementary Material for details). Using this approach, two measures are derived using slightly different mehods: In the *simple pairwise* method, the percentage of conserved SSR *loci* (C_{sim}) is determined by directly aligning the SSRs in microsynteny blocks between a species pairs regardless of their position in the phylogenetic tree. In contrast, the percentage of conserved SSR *loci* inferred with the *progressive* method (C_{pro}) is more stringent for non-sister species: at all ancestral nodes until the most recent common ancestor of the two species, it iteratively determines the conserved SSR *loci*, thus requiring that any conserved SSR locus between the two species under investigation has also been conserved in all other species that have split after the most recent common ancestor node. The *simple pairwise* method is likely to over-estimate the rate of conservation for distantly related species because it ignores the phyloge-

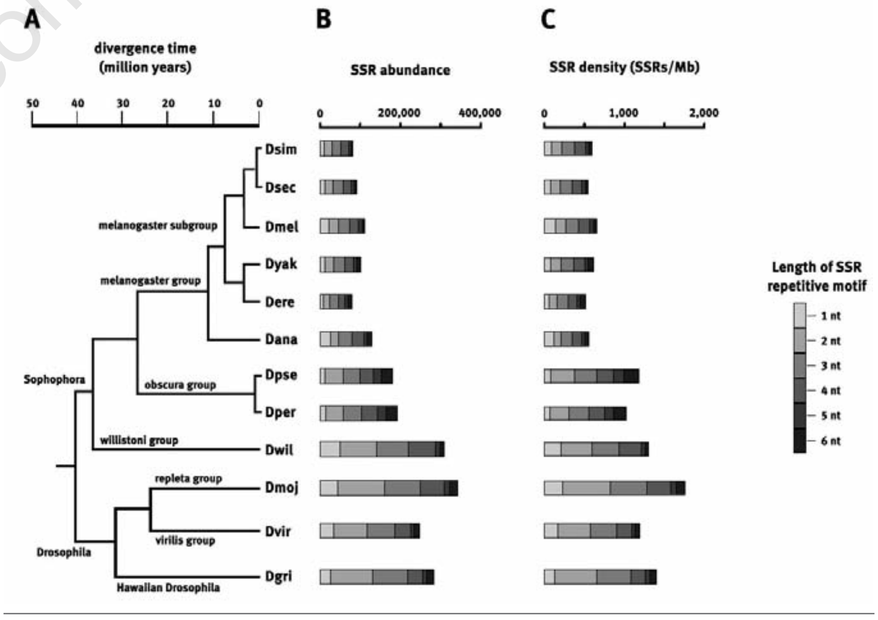

Figure 1. Full genome simple sequence repeat (SSR) composition of the 12 *Drosophila* species. SSR composition was evaluated distinguishing SSRs with mono-to hexa-nucleotide repeats. A) species tree with branch lengths indicating divergence time in million years; B) absolute numbers of SSRs per species; C) SSR densities, *i.e.* SSR counts per species normalized by the genome size in megabases.

netic context and the divergence time and can thus align non-homologous SSRs that have emerged at similar positions. In contrast, the *progressive* method is very stringent, in particular for species pairs whose ancestral node is deeply rooted because it requires that the homologous SSR is conserved in *all* child nodes of the ancestral node. Thus, the *progressive* method under-estimates SSR conservation because lineage-specific loss will cause homologous SSR *loci* in all other lineages of a clade to be ignored.

Overall macroevolutionary trend of simple sequence repeats

As mentioned above, the assessment of conservation of SSRs is restricted to microsynteny blocks, *i.e*. genomic regions in which the neighborhood of two or more orthologous genes is perfectly conserved across all 12 genomes, in order to facilitate correct identification of orthologous SSRs. Syntenic regions were identified with OrthoCluster.[45,46] Parameters were chosen so that microsynteny blocks consist of two ore more orthologous genes that are direct neighbors and have conserved strandedness and orientation across all 12 genomes (see Methods for details). Across all 12 *Drosophila* species, 1139 microsyntenic blocks were identified, containing 2709 orthologous genes that are positionally conserved among all of the species. These syntenic regions cover on average 5.25% of the genome and contain 95,801 SSRs across all 12 species, *i.e*. 4.48% of all identified SSRs. For all possible pairs among the twelve *Drosophila* species, the percentage of conserved SSR *loci* in these syntenic regions was determined; with reference to the divergence times between these pairs, this analysis yielded 66 data points within 40 million years of evolution.

We found that the rate of SSR locus conservation (C_{sim} and C_{pro}) decays exponentially with increasing divergence time for both the *simple pairwise* and the *progressive* method (Figure 2; see Methods for details). Such an exponential decay has been observed as well for the conservation of human SSR *loci* in vertebrate genomes and may be a general trend in SSR macroevolution.[22] As mentioned before, SSR mutation rates are much lower in *Drosophila* than in vertebrates, and reduced mutation rates might partly be attributed to the facts that i) SSRs in *D. melanogaster* are shorter,[58] and ii) shorter SSRs are less mutable.[2,59] By applying the same SSR identification method used in this study for Drosophilidae to five mammal genomes (human, chimpanzee, mouse, rat, pig), we indeed found that *Drosophila* SSRs are shorter than mammalian SSRs (on average 26 nt *vs.* 32 nt; see Supplementary File 1 for details).

Simple sequence repeat *locus* conservation within and between *Drosophila* species

We sampled SSR *loci* which, according to our methodology, were conserved among at least three *Drosophila* species. All conserved *loci* were required to occur in *D. melanogaster* (*Dmel*) which allowed to test if repeat *loci* that are conserved beyond species boundaries are already more strongly conserved in the *Dmel* population compared to non-conserved *loci*. Accordingly, at SSR *loci* which are conserved across species, we found that a homologous SSR is present in almost all 37 individuals of the *Dmel* population data set (Figure 3A). Moreover, cross-species conserved repeat *loci* were found to be monomorphic, *i.e.* only one state exists in all individuals in the *Dmel* population, for most of these *loci* (Figure 3B). We find that, independently from the degree of cross-species conservation, most SSR *loci* in the population have a strong locus conservation and have a small number of different states. However, we found that cross-species conserved SSR *loci* have, on average, a stronger locus conservation in the population (two sample Kolmogorov-Smirnov test: D=0.227, P<2.2e^{-16}) and have a lower number of different states (two sample Kolmogorov-Smirnov test: D=0.121, P=2.7e^{-12}). Our findings suggest that a number of repeat *loci* may evolve more slowly than others which facilitates their long-term conservation across several species. Such decreased SSR mutability could be caused by both neutral evolution and natural selection. For instance, it has been shown that mutation rates depend on many inherent characteristics of SSRs such as motif type, length of the motif and the repeat, GC content etc.[60] Therefore, more stable SSRs could have a higher chance of long-term conservation. Alternatively, increased selective constraints might be responsible for the slower evolution of some *loci*. Our microsynteny blocks contain SSRs within or in close proximity to genes where some of these

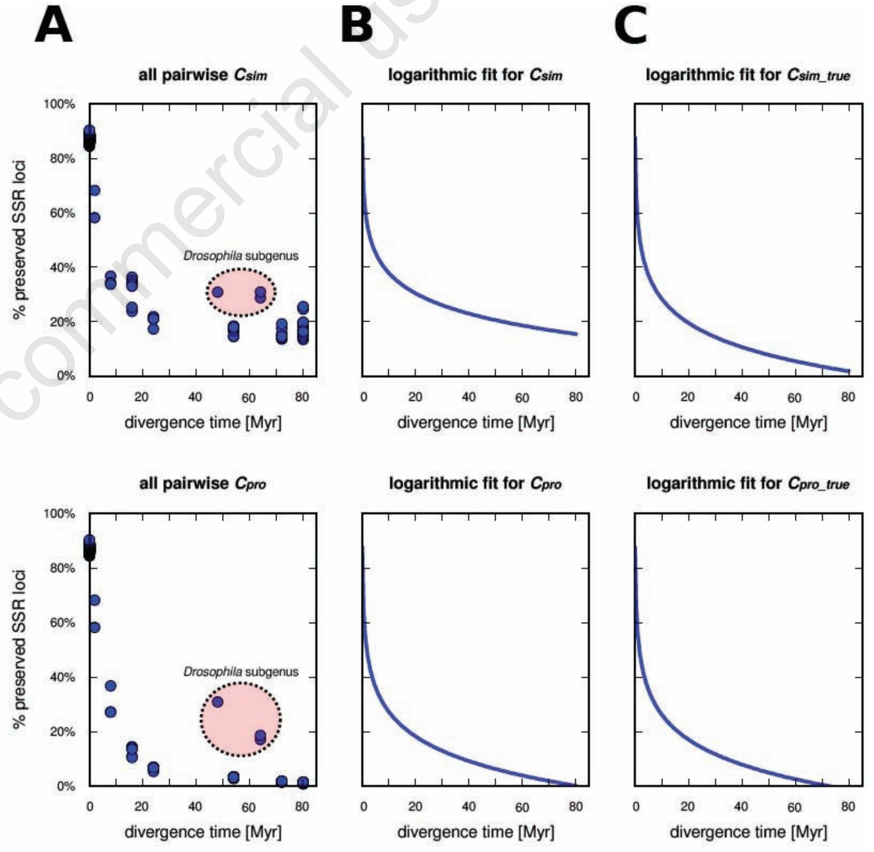

Figure 2. Rates of simple sequence repeat (SSR) *locus* conservation within 37 individuals of a *D. melanogaster* population and between 12 *Drosophila* species relative to their divergence time. A) Between all individual/species pairs, two measures for SSR locus conservation, C_{sim} and C_{pro}, were obtained by applying two different methods, simple pairwise and progressive (see Materials and Methods); B) both rates can be well approximated by a logarithmic fit; C) taring the conservation rates using randomly sampled SSRs as a null model removes potential false positive conserved loci so that C_{sim_true} and C_{pro_true} represent the most conservative and stringent estimates for rates of SSR locus conservation.

repeat *loci* might in fact overlap with functionally relevant genomic regions such as regulatory units (*e.g.* promoters of genes) or encode for functional parts of the protein (*e.g.* amino acid repeats).[30] We discuss some of these issues further below in the context of a decreased evolutionary rate of exonic SSRs.

Slower simple sequence repeats decay in the *Drosophila* subgenus

Among all pairwise *Drosophila* species comparisons, we could find slight deviations from the overall steady exponential decay. Both C_{sim} and C_{pro}, even after normalization for any potential biases, suggest that three species pairs have increased repeat locus conservation rates: *Dmoj-Dvir*, *Dmoj-Dgri*, and *Dvir-Dgri*. All these pairs belong to the *Drosophila* subgenus and show rates of SSR locus conservation that are substantially higher than expected based on the divergence times (highlighted in Figure 2). We hypothesize that these differences may be caused by SSRs getting less frequently lost in the *Drosophila* subgenus than in *Sophophora* species. The loss of an SSR can be caused by repeat variation below a minimum threshold or by one or more point mutations interrupting the perfect repeat structure. If the SSR has become too short or interruptions too frequent, DNA replication slippage will no longer occur and the SSR is *dead*.[13,61] However, it has been suggested that in viable SSRs, interruptions get removed during DNA replication slippage.[8] In fact, these two mechanisms in combination, repeat variation through replication slippage and single nucleotide substitutions, could explain the increased SSR locus retention rate in the *Drosophila* subgenus: evolutionary rates (synonymous mutation rates, *dS*) were found to be lower for the three *Drosophila* subgenus species compared to *Sophophora* species which translates into a decreased rate of interrupting mutations.[62] On the other hand, for at least one of the three *Drosophila* subgenus species, *D. virilis*, an increased SSR mutation rate in comparison to *D. melanogaster* (*Sophophora*) has been reported.[56] Thus, a higher rate of DNA replication slippage leads to an increased chance for interruptions to be purged. Therefore, such an increased ratio of SSR mutation rate to nucleotide substitution rate may lead to a prolonged SSR locus retention, as observed in the *Drosophila* subgenus, under neutral evolution. In turn, we do not expect natural selection to play a significant role in the observed increase of SSR locus conservation.

Simple sequence repeats *loci* in exons evolve more slowly

As a unifying trend, we found that repeat *loci* which are located in exons are more often conserved and retained between species over longer evolutionary time scales than *loci* in introns and intergenic regions (Figure 4). The rates of SSR *locus* conservation between intronic and intergenic regions are highly similar, suggesting that overall, SSR *loci* outside of coding regions evolve neutrally and that selection pressure, if any, acts equally strongly on intronic and intergenic SSR *loci*. As suggested before, elevated conservation rates for exonic SSR *loci* could be a consequence of purifying selection acting on coding regions with SSRs frequently encoding functionally or structurally important amino acid repeats.[63]

While a significant fraction of repeat *loci* are frequently conserved between species - especially between closely related ones - the states at such *loci*, *i.e.* the exact repetitive sequences are much more volatile, showing variation in numbers of repeats and interrupting mutations. Overall, across all twelve *Drosophila* genomes, *perfect* conservation is relatively rare and was found for only 15% of all conserved repeat *loci* (Figure 5). The most frequent state of repetitive sequences at conserved *loci* is *variable*, *i.e.* the sequences at a conserved *locus* show repeat number variation. However, similar to the rate of repeat *locus* conservation, the conservation of *states* at homologous *loci* strongly depends on the divergence time: between the most closely related species, the largest fraction of all conserved SSR *loci* have *perfectly* conserved states. With increasing divergence time, most repeat *loci* show either a *variable* or an *interrupted* state (Figure 5).

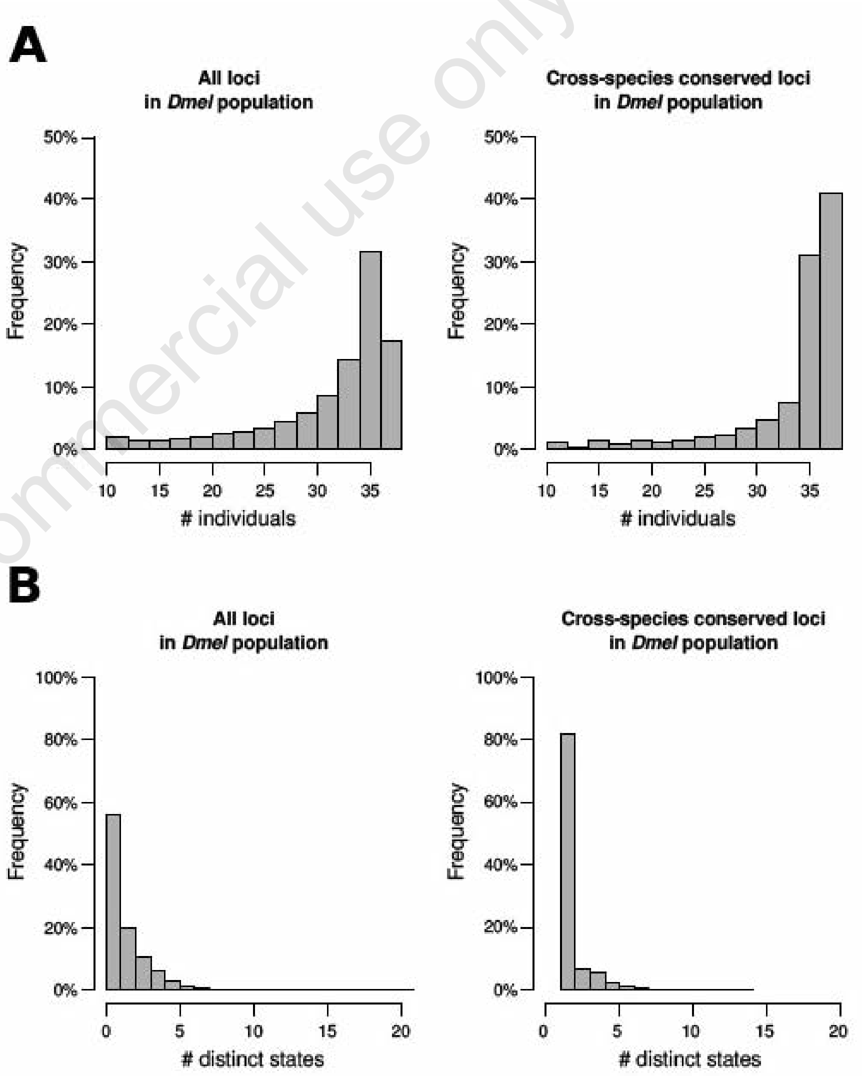

Figure 3. Comparison of the abundance and variability between all simple sequence repeats (SSR) loci and those found conserved between multiple *Drosophila* species across 37 *D. melanogaster* (*Dmel*) individuals. A) Number of individuals sharing a significant SSR at the same locus; B) number of distinct states (alleles) at the same SSR locus. For these analyses, only repeat loci shared by at least 10 *Dmel* individuals were considered; unrestricted results are given in the Supplementary Material.

Since exonic SSRs were found to be more strongly conserved than SSRs in non-coding regions, we specifically tested whether perfect state conservation was also more frequent for conserved SSR *loci* located in exons. Statistical evaluation for cross-species and within-species conserved *loci* all support a significant bias towards exonic SSRs being more frequently perfectly conserved than the states of conserved SSR *loci* outside of exons ($P<1e^{-10}$; see Supplementary File 1). In fact, the inferred odds ratios for cross-species and among *D. melanogaster* individuals conserved repeat *loci* suggest that exonic SSRs are 1.4 to 1.8 times more often perfectly conserved than non-exonic SSRs. Our analyses also indicate that this increased perfect state conservation is independent from the divergence time, with a significantly increased perfect conservation of exonic SSRs found within the *D. melanogaster* population, across *Drosophila* species with less than 10 Myr divergence times, and across *Drosophila* with the maximum divergence time of 80 Myr (see Supplementary File 1). These findings suggest that substitutions in the repetitive sequences occur at fairly constant rates over time. Our findings of exonic SSRs being more constrained than non-coding SSRs is largely in line with an earlier study investigating repeat expansion, *i.e.* the fixation of SSR mutations leading to a gain of a repeat unit, in coding and non-coding regions in a variety of eukaryotes.[64] However, this study reported that such repeat expansion, due to their potential of inducing frameshift mutations, is only significantly constrained for non-triplet repeats, *i.e.* for SSRs that consist of repeat units whose length is not divisible by three.[64] Since most of exonic SSRs (93%) in this study are triplet repeats and were found to be more perfectly conserved in exons than on non-coding regions, our study suggests that evolutionary constraints also affect triplet repeats in exons.

We note that these trends reflect the evolution of a large group of SSRs, and exceptions at single *loci* will likely exist. Moreover, the degree of conservation does not necessarily predict whether or not the SSR plays a functional role, *e.g.* as a regulatory element or encoding an amino acid repeat. Over macroevolutionary timescales, conserved homologous repeat *loci* more frequently contain slightly different instead of perfectly conserved sequences (Figure 5). Even for SSRs that are functionally relevant,[4] it might be sufficient to be maintained as an imperfect SSR, *i.e.* as a sequence that is conserved only at critical positions while other sites are allowed to change to still preserve their function. Moreover, the accumulation of interrupting mutations might facilitate the retention of functionally relevant sequences as imperfect repeats; interrupting the perfect repeat structure of SSRs has been shown to reduce the chance of replication slippage and thus increase SSR stability.[65] Similarly, repeat number variation might exploit the potential of SSRs to function as *tuning knobs*, typically acting as regulatory elements.[66-68] Thus, the degree of conservation of an SSR is not necessarily linked to its functional relevance, and variation can both be caused by neutral evolution and natural selection.

Simple sequence repeat *loci* on the sex chromosome evolve faster

Many studies have reported accelerated evolutionary rates of *loci* on the sex chromosome compared to autosomes.[4,69-72] Several factors are thought to be responsible for this accelerated evolution, including a higher number of cell divisions per generation in the male germ line than in females leading more frequently to replication errors in males, hemizygosity and the immediate exposure of mutations to selection, and differences in effective and relative population size. We tested whether or not SSR *locus* conservation is affected by their localization on autosomes (*A-linked*) *vs*. the sex chromosome (*X-linked*). *D. melanogaster* was used as the focal species since we can exploit both within-species comparison using the DPGP dataset and cross-species comparisons in syntenic regions. Since the *D. melanogaster*

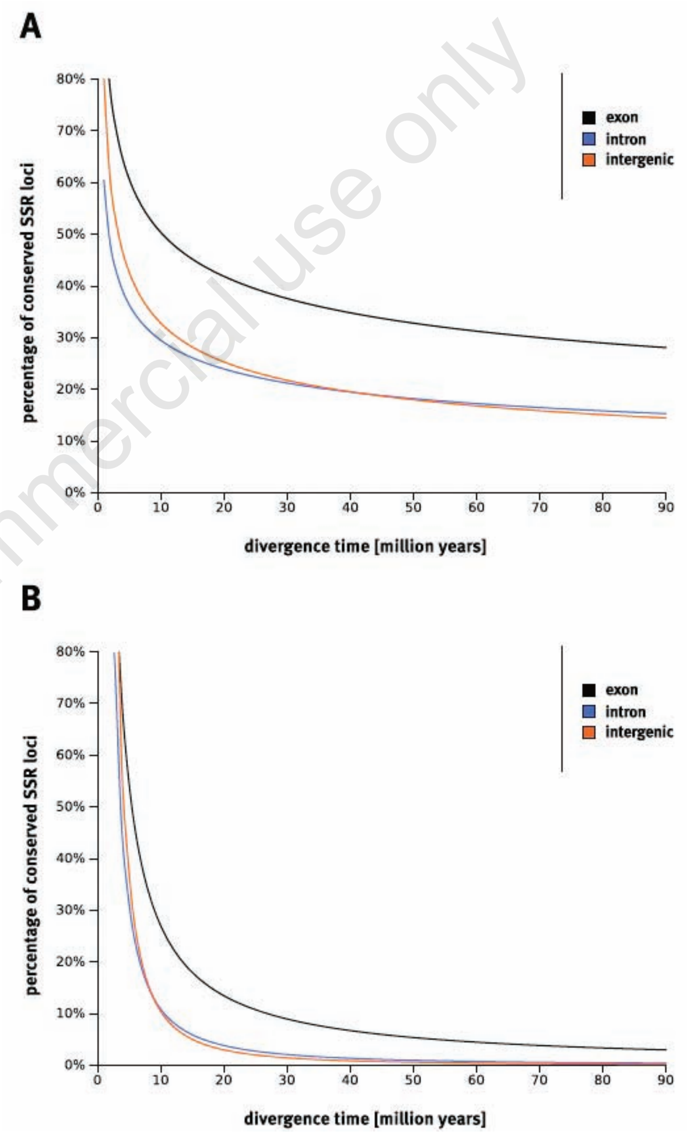

Figure 4. Rate of simple sequence repeat (SSR) *locus* conservation across 12 *Drosophila* species relative to the divergence time between the species distinguishing their genomic features between exons, introns, and intergenic. A) Logarithmic fit for the rate of conservation derived from the simple pairwise method; B) logarithmic fit for the rate of conservation derived from the progressive method. Raw data on which the logarithmic fits were constructed can be found in the Supplementary Material.

genome sequence is perfectly resolved into chromosomes, we can infer which SSRs in syntenic regions are linked to autosomes or to the sex chromosome. For the cross-species comparison, all other *Drosophila* species except *D. pseudoobscura* and *D. willistoni* were used. These two species have independently gained the *neo-X* chromosome by fusing part of the X chromosome with an autosome.[73] Across species, we found that conservation of X-linked SSR *loci* is 2-fold lower, and X-linked *loci* are conserved among fewer species than A-linked *loci* (see Supplementary File 1). These findings fit very well the reported 2-fold higher nucleotide divergence in *Drosophila miranda* and a two-fold increased mutability of primate mono-nucleotide SSRs when contrasting sex chromosome with autosomes.[60,71] Within the *Dmel* population, an SSR *locus* is less frequently shared among all individuals and a higher number of distinct states (alleles) can be found for X-linked relative to A-linked repeat *loci* (see Supplementary File 1). These findings indicate that the faster evolutionary rate of X-linked *Drosophila* SSRs can be seen at both micro- and macro-evolutionary time scales.

Repeat *locus* conservation *vs.* conservation of genes

Different genomic conservation and divergence measures such as protein sequence identity among orthologs, the frequency of positionally conserved introns, and the conservation of synteny seem to be highly correlated with each other.[74] We tested whether over a broad time scale, SSRs follow an evolutionary trajectory that is distinct from other, selectively non-neutral genomic regions. Sequence divergence between the orthologous genes used to identify syntenic regions was found highly correlated to the conservation rates of the SSR *loci* located in these syntenic regions (see Supplementary File 1). While the exponential decay of orthologous proteins is much slower, protein sequences and repeat *loci* among species follow roughly the same dynamics. The observed exponential decay of both these genomic features might be due to the time-dependent rate of molecular evolution, *i.e.* stem from the fact that micro- and macro-evolutionary derived measures reflect mutation and substitution rates, respectively.[75] We show that the species tree can be correctly reconstructed almost to the full extent only based on pairwise SSR *locus* conservation rates (see Supplementary File 1). All *Sophophora* species were positioned correctly, whereas the three *Drosophila* subgenus species *Dgri*, *Dmoj*, and *Dvir* could not be placed correctly. The difficulty to position the *Drosophila* subgenus species correctly might be a consequence of the afore mentioned high SSR conservation rates among these three species. Nonetheless, this analysis indicates that SSR *locus* conservation is a more robust measure than SSR sequence conservation and has the potential to resolve species relationships much further.[21] Note that we do not report on tree reconstruction based on C_{pro} since the progressive method implicitly uses the species tree so that it is a prerequisite of this method to know the species tree to be applicable.

Common ancestry and the birth and death of simple sequence repeats in *Drosophilidae*

We predict homologous SSRs which are located in syntenic regions between pairs of species, but do not discriminate between putatively gained and lost repeats. Since the SSR *locus* conservation rates C_{sim} and C_{pro} reflect the percentage of homologous SSR *loci* among all SSRs of two species, the fraction of non-conserved *loci* could be the result of either of the two processes: An SSR which was present in the last common ancestor but has not been conserved in both species, or an SSR which was not present in the last common ancestor but was gained lineage-specifically. As an approximation for common ancestry, we determined the percentage of SSR *loci* that our *simple pairwise* method predicts to be homologous between any pair consisting of one *Drosophila* and one *Sophophora* subgenus species. The results of these pairwise analyses indicate that 47% of all SSRs which are present in any of the twelve *Drosophila* species may have been present at the root of the Drosophilidae. Common ancestry of only half of the SSRs suggests that the other half of the extant SSRs in Drosophila *species* has been gained after the split from the most recent common ancestor of all twelve *Drosophila* species. In contrast, we consider SSRs as not present or *dead* if the repeat sequence mutated below the minimum length threshold of 12 nt, or if interrupting mutations disrupted the perfect repeat pattern too heavily (see Methods). Testing for the phylogenetic spread of SSR *loci* (*i.e.* the number of species in which a homologous SSR *locus* can be found) revealed that 2867 of the 95,801 SSRs in syntenic regions - *i.e.* only about 6% of repeat *loci* present in the last common ancestor - are predicted to be conserved across at least 10 of the 12 extant species. These data support a high rate of birth and death of repeat *loci* which is likely associated with the full-genome variability of SSR content observed among the *Drosophila* species (Figure 1).

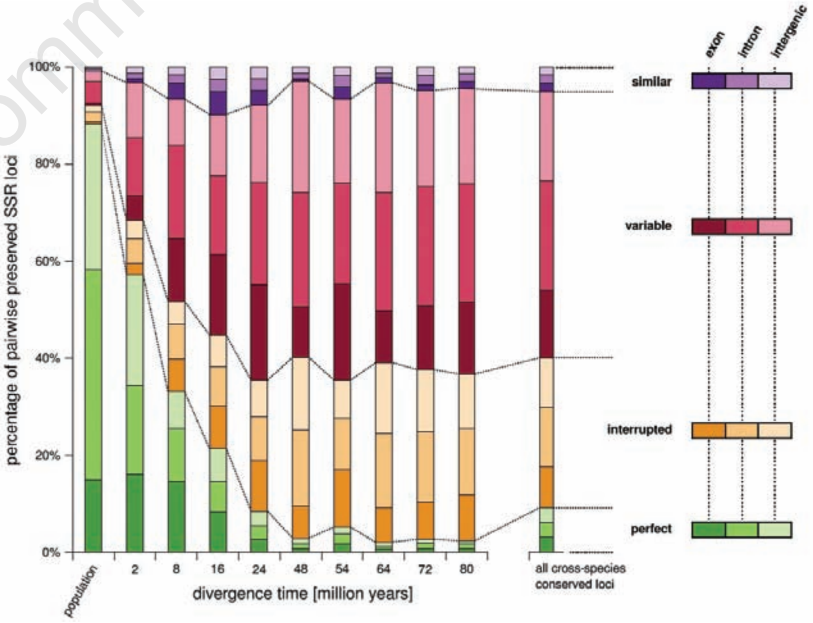

Figure 5. Similarity of the state of a conserved simple sequence repeat (SSR) *locus* depending on the divergence time of the species pair under investigation. Pairs of homologous SSRs are classified into one of four categories: perfect refers to 100% sequence identity of the repetitive sequence of the two SSRs; interrupted describes that the two SSRs have the same length and number of repeats but differ in the number of mismatches; variable refers to two SSRs that are not interrupted but whose repetitive sequences differ only in the number of repeat units; similar refers to differences between two SSRs due to a combination of repeat number variation and interrupting bases.

Conclusions

Our analyses extend Buschiazzo and Gemmell's study which was limited to pairwise comparisons of human against 10 other species and included only 2 data points with species that split less than 90 Ma.[22] In contrast, we exploit the *Drosophila melanogaster* population data set to bridge the gap between SSR micro- and macro-evolution. This gap will shrink further with additional genome sequences being released in the near future, either from very closely related sister species or from separate populations of the same species. The increased resolution of the presented cross-species analysis, using all against all comparisons, facilitated identifying outliers from the overall trend, *i.e.* the increased SSR *locus* retention rate within the *Drosophila* subgenus (Figure 2), which might have remained hidden with a one against all approach. The prolonged retention of SSR *loci* in the *Drosophila* subgenus is not a methodological artifact but is predicted even after correcting for the high similarity in genomic SSR frequencies (see Supplementary File 1). It is possible that an increased ratio of SSR mutation rate to nucleotide substitution rate is responsible for purging interrupting mutations from SSR sequences, thereby prolonging their retention time in *Drosophila* subgenus genomes.

Throughout this study, we employed two different methods (*simple pairwise* and *progressive*) to identify conserved SSR *loci* between species pairs. In conjunction with a null model, drawn from random sampling of SSRs, we can set a lower boundary for false positives among conserved *loci*. Overall, C_{sim} and C_{pro} provide upper and lower bounds respectively for the rates of SSR *locus* conservation and thus define a zone wherein homoplasy provides an alternative explanation for the observed SSRs at one *locus* (Figure 6). Our study extends beyond presence/absence of SSRs and for the first time provides data on how repetitive sequences at conserved *loci* (which we referred to as *states*) evolve within and between species. In general, comparative studies of SSRs suffer from difficulties in defining and identifying an SSR since an SSR slightly below a certain length will not be detected as such but may be easily resurrected by expansion in a closely related species or even within a population.[7,76] Ultimately, we suggest that, according to the decay rate given in Figure 6, reliable inferences about the conservation of an SSR *locus* beyond a few million years require the proven existence of homologous states in intermediate species. Otherwise, homoplasy is at least a likely alternative explanation.

For technical reasons, the conservation rates obtained in our study originate only from SSR *loci* located in syntenic regions, *i.e.* genomic blocks that are conserved across all 12 *Drosophila* species. In general, these genomic regions might be more strongly conserved and constrained than other regions in the genome so that the rates reported here would over-estimate the full-genome conservation rates.[77] On the other hand, especially in the population data set, we have frequently observed cases where one interrupting mutation (*i.e.* a single nucleotide substitution) in a short SSR is sufficient to put the sequence below the minimum threshold for detection. This may falsely suggest that an SSR has been lost altogether. However, it has been suggested that such interruptions are only transition states in microsatellite SSR evolution.[8] The described phenomenon is again influenced by the definition of an SSR and will lead to an over-estimation of loss. It may also explain why SSR *locus* conservation within one population of the same species is as low as around 90%. Therefore, the conservation rates reported here are quite conservative as they hinge upon a strict criterion of definition. The results presented here will also help disentangling the involved matter of causes and consequences of repeat evolution. Further research will need to look deeper into how general the trends of rate conservation across time scales and different genomic features are and if the causes of rate divergence between species can be further pinned down.

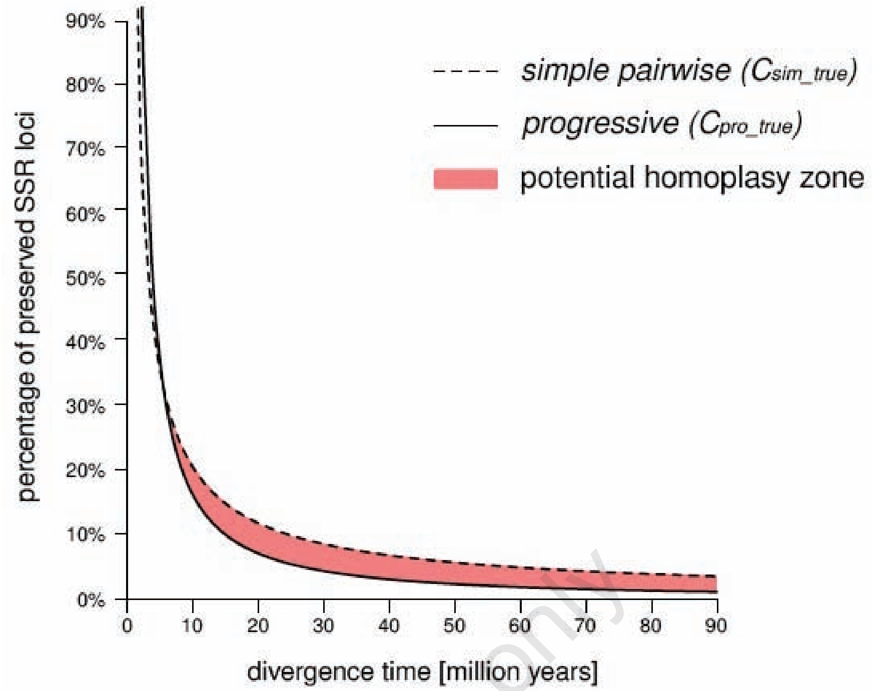

Figure 6. Fitted rates of C_{sim_true} (dashed line) and C_{pro_true} (solid line). The two measures are the observed rates (C_{sim} and C_{pro}) corrected for the potential of aligning nonhomologous loci based on a randomization test (see Supplementary Material). C_{sim_true} and C_{pro_true} represent the most conservative upper and lower boundary of the rate of simple sequence repeat (SSR) *locus* conservation between Drosophila species pairs in relation to their divergence times. While the progressive method is conservative and likely to overestimate SSR loss, the simple pairwise method is susceptible to matching SSRs that are not of common ancestry but a result of convergent evolution (homoplasy). In combination, these two measures allow for an estimate of potential homoplasy (homoplasy zone).

References

1. Schlötterer C. The evolution of molecular markers-just a matter of fashion? Nat Rev Genet 2004;5:63-9.
2. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 2004;5:435-45.
3. Balaresque P. Microsatellites or the eukaryotic genome: life cycle concept and neutrality issues. Med Sci (Paris) 2007;23:729-34.
4. Li YC, Korol AB, Fahima T, et al. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 2002;11:2453-65.
5. Levinson G, Gutman GA. Slipped-strand

mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 1987;4: 203-21.

6. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res 1992;20:211-5.
7. Buschiazzo E, Gemmell NJ. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 2006;28: 1040-50.
8. Harr B, Zangerl B, Schlötterer C. Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila. Mol Biol Evol 2000; 17:1001-9.
9. Schlötterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma 2000; 109:365-71.
10. Schug MD, Mackay TF, Aquadro CF. Low mutation rates of microsatellite loci in Drosophila melanogaster. Nat Genet 1997;15:99-102.
11. Baer CF, Miyamoto MM, Denver DR. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet 2007;8:619-31.
12. Messier W, Li SH, Stewart CB. The birth of microsatellites. Nature 1996;381:483.
13. Taylor JS, Durkin JM, Breden F. The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol Biol Evol 1999;16:567-72.
14. Chambers GK, MacAvoy ES. Micro-satellites: consensus and controversy. Comp Biochem Physiol B Biochem Mol Biol 2000;126:455-76.
15. Meglécz E, Anderson SJ, Bourguet D, et al. Microsatellite flanking region similarities among different loci within insect species. Insect Mol Biol 2007;16:175-85.
16. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 2000;10:967-81.
17. Barker GC. Microsatellite DNA: a tool for population genetic analysis. Trans R Soc Trop Med Hyg 2002;96:S21-4.
18. Reusch TBH, Bolte S, Sparwel M, et al. Microsatellites reveal origin and genetic diversity of Eurasian invasions by one of the world's most notorious marine invader, Mnemiopsis leidyi (Ctenophora). Mol Ecol 2010;19:2690-9.
19. Bhargava A, Fuentes FF. Mutational dynamics of microsatellites. Mol Biotechnol 2010;44:250-66.
20. Selkoe KA, Toonen RJ. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 2006;9:615-29.
21. Sun JX, Mullikin JC, Patterson N, Reich DE. Microsatellites are molecular clocks that support accurate inferences about history. Mol Biol Evol 2009;26:1017-27.
22. Buschiazzo E, Gemmell NJ. Conservation of human microsatellites across 450 million years of evolution. Genome Biol Evol 2010;2010:153-65.
23. Jarne P, Lagoda PJ. Microsatellites, from molecules to populations and back. Trends Ecol Evol 1996;11:424-9.
24. Barbará T, Palma-Silva C, Paggi GM, et al. Cross-species transfer of nuclear microsatellite markers: potential and limitations. Mol Ecol 2007;16:3759-67.
25. Shah SN, Hile SE, Eckert KA. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 2010;70:431-5.
26. Woerner SM, Kloor M, von Knebel Doeberitz M, Gebert JF. Microsatellite instability in the development of DNA mismatch repair deficient tumors. Cancer Biomark 2006;2:69-86.
27. Mirkin SM. Expandable DNA repeats and human disease. Nature 2007;447:932-40.
28. Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 2005;6:729-42.
29. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 2008;18: 1011-9.
30. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 2010;44: 445-77.
31. Simon M, Hancock JM. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol 2009;10:R59.
32. Dosztányi Z, Chen J, Dunker AK, et al. Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res 2006;5:2985-95.
33. Hancock JM, Simon M. Simple sequence repeats in proteins and their significance for network evolution. Gene 2005;345:113-18.
34. Faux NG, Bottomley SP, Lesk AM, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 2005;15:537-51.

35 Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nat Genet 2005;37: 986-90.

36. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A 2007;104:8472-7.
37. Fujita PA, Rhead B, Zweig AS, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res 2011;39:D876-82.
38. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007;450: 203-18.
39. Singh ND, Larracuente AM, Sackton TB, Clark AG. Comparative genomics on the drosophila phylogenetic tree. Annu Rev Eco Syst 2009;40:459-480.
40. Crosby MA, Goodman JL, Strelets VB, et al. FlyBase: genomes by the dozen. Nucleic Acids Res 2007;35:D486-91.
41. Drysdale R, FlyBase Consortium. FlyBase : a database for the Drosophila research community. Methods Mol Biol 2008;420: 45-59.
42. Jordan KW, Carbone MA, Yamamoto A, et al. Quantitative genomics of locomotor behavior in Drosophila melanogaster. Genome Biol 2007;8:R172.
43. Flicek P, Amode MR, Barrell D, et al. Ensembl 2011. Nucleic Acids Res 2011;39: D800-6.
44. Kofler R, Schlötterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 2007;23:1683-5.
45. Zeng X, Pei J, Vergara IA, et al. OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. In: In 11th International Conference on Extending Database Technology (EDBT'08), Nantes, France. 2008.
46. Vergara IA, Chen N. Using OrthoCluster for the detection of synteny blocks among multiple genomes. Curr Protoc Bioinformatics 2009;27:6101-18.
47. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443-53.
48. Markow TA, O'Grady PM. Drosophila biology in the genomic age. Genetics 2007;177: 1269-76.
49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32: 1792-7.
50. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4: 406-25.
51. Gascuel O, Steel M. Neighbor-joining revealed. Molecular biology and evolution 2006;23:1997-2000.
52. Huntley MA, Clark AG. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. Mol Biol Evol 2007;24:2598-609.
53. Lowenhaupt K, Rich A, Pardue ML. Nonrandom distribution of long mono- and dinucleotide repeats in Drosophila chromosomes: correlations with dosage compensation, heterochromatin, and recombination. Mol Cell Biol 1989;9:1173-82.
54. Marin I, Labrador M, Fontdevila A. The evolutionary history of Drosophila buzzatii. XXVI. High content of non-satellite

repetitive DNA in D. Buzzatii and in its sibling D. koepferae. Genome 1992;35:967-74.

55. Marín I, Fontdevila A. Evolutionary conservation and molecular characteristics of repetitive sequences of Drosophila koepferae. Heredity 1996;76:355-66.
56. Schlötterer C, Harr B. Drosophila virilis has long and highly polymorphic microsatellites. Mol Biol Evol 2000;17:1641-6.
57. Ross CL, Dyer KA, Erez T, et al. Rapid divergence of microsatellite abundance among species of Drosophila. Mol Biol Evol 2003;20:1143-57.
58. Schug MD, Regulski EE, Pearce A, Smith SG. Isolation and characterization of dinucleotide repeat microsatellites in Drosophila ananassae. Genet Res 2004;83:19-29.
59. Webster MT, Smith NGC, Ellegren H. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci U S A 2002;99:8748-53.
60. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome research 2008;18:30-8.
61. Boyer JC, Hawk JD, Stefanovic L, Farber RA. Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. Mutat Res 2008;640:89-96.
62. Larracuente AM, Sackton TB, Greenberg AJ, et al. Evolution of protein-coding genes in Drosophila. Trends Genet 2008;24:114-23.
63. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 2004;21:991-1007.
64. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 2000;10:72-80.
65. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 1998;95:10774-8.
66. DG K, Soller M, Kashi Y. Evolutionary Tuning Knobs. Endeavour 1997;21:36-40.
67. Trifonov EN. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. in: S. P. Wasser ed., Evolutionary theory and processes: nodern horizons, papers in honor of Eviatar Nevo. Amsterdam, The Netherlands: Kluwer Academic Publishers; 2003.
68. Vinces MD, Legendre M, Caldara M, et al. Unstable tandem repeats in promoters confer transcriptional evolvability. Science 2009;324:1213-6.
69. Charlesworth B, Coyne J. The relative rates of evolution of sex chromosomes and autosomes. American Naturalist 1987;130:113-46.
70. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nature reviews Genetics 2006;7:645-53.
71. Bachtrog D. Evidence for male-driven evolution in Drosophila. Molecular biology and evolution 2008;25:617-9.
72. Grath S, Parsch J. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of Drosophila evolution. Genome Biol Evol 2012;4:346-59.
73. Meisel RP, Han MV, Hahn MW. A complex suite of forces drives gene traffic from Drosophila X chromosomes. Genome biology and evolution 2009;1:176-88.
74. Zdobnov EM, von Mering C, Letunic I, Bork P. Consistency of genome-based methods in measuring Metazoan evolution. FEBS letters 2005;579:3355-61.
75. Ho SYW, Lanfear R, Bromham L, et al. Time-dependent rates of molecular evolution. Molecular ecology 2011;20:3087-101.
76. Merkel A, Gemmell NJ. Detecting microsatellites in genome data: variance in definitions and bioinformatic approaches cause systematic bias. Evol Bioinform Online 2008;4:1-6.
77. Lemoine F, Lespinet O, Labedan B. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evol Biol 2007;7:237.