Genome-wide conservation and degeneration of SSRs, Simple Sequence Repeats, across long evolutionary time scale

Lothar Wissler[1], Lars Godmann[1], and Erich Bornberg-Bauer[*,1]

Supplementary Material

Affiliations: [1] Institute for Evolution and Biodiversity, University of Muenster, Hufferstrasse 1, D48149 Muenster, Germany

Detailed settings for SSR identification with SciRoko

score	12
mismatch penalty	4
SSR seed minimum length	8
SSR seed minimum repeats	3
maximum mismatches at once	3

Detailed settings for synteny identification with OrthoCluster

Max group size	1000
Min group size	2
Max out-of-map-mismatches	0
Max out-of-map-mismatch percentage	0%
Max in-map-mismatches	0
Max in-map-mismatch percentage	0%
Find order preserving blocks(-r)	Yes
Find strand preserving blocks(-s)	Yes
Find order and strandedness preserving blocks(-rs)	Yes
Find non-overlapping blocks	Yes

Full-genome SSR densities in Drosophila and outgroup species

To infer the direction of changes in genomic SSR density, we determined the number of SSRs per Mb for the twelve Drosophila genomes and five outgroup species (Table 1). SSRs were identified with the method described in the Methods section of the main article. Two major trends could be observed:

1. While in Drosophila genomes we observed SSR densities between 515 and 1,750 SSRs/Mb, the outgroup species genomes showed significantly reduced densities, between 105 and 405 SSRs/Mb (Figure 1A). This suggests that SSR densities have increased throughout the Drosophila clade, with lower SSR densities being the ancestral state.

2. Between the two subgenera Sophophora and Drosophila, we found a strong difference in SSR density (Figure 1B). Together with the previous finding, these data suggest that SSR densities have further increased within the Drosophila subgenus.

Comparison of SSR length between Drosophilidae and mammals

We compared length of SSRs between Drosophilidae (represented by the twelve Drosophila genomes) and mammals (represented by human, chimpanze, mouse, rat, pig). SSRs were identified with the same settings for both data sets (see Methods: Microsatellite identification). We show histograms and a boxplot indicating that mammalian SSRs are longer than those found in Drosophilidae (Figure 2). Furthermore, we statistically tested whether the length of Drosophila SSRs are a random sample from the length distribution of the mammalian SSRs; this hypothesis could be rejected (two-sample Kolmogorov-Smirnov test: $D = 0.1698$, $p < 2.2e^{-16}$).

Null Model for SSR locus conservation

With a limited set of different SSR motives and conservation measures based on global alignments, it is virtually impossible to obtain 0% matched SSRs, because SSRs with a frequent motif can be matched most of the time between two species given the introduction of enough gaps in the alignment. Here, we compute the "null model" or the expected conservation rate given randomly sampled (non-homologous) SSRs. For each syntenic region, the true SSRs are replaced by the same number of randomly sampled SSRs that occur somewhere in other syntenic regions for each of the species. As with real data, we compute C_{sim} and C_{pro} for these randomized data which we refer to as C_{sim_rand} and C_{pro_rand}, respectively. C_{sim_rand} is relatively stable at 12% regardless of the divergence time, whereas C_{pro_rand} decreases from an initial 12% to 0.15% for species pairs with increasing divergence time (Figure 3). In comparison, the real rates C_{sim} and C_{pro} are much higher, but approach C_{sim_rand} and C_{pro_rand} with increasing divergence time.

Similarity of genomic SSR motives and their correlation with divergence time

For each species pair, we compute similarity values as follows: Across all syntenic regions, we count the SSR motives, i.e. the SSR standardized motif plus the gene feature (exon/intron/intergenic), as aligned with Needleman-Wunsch. These counts are transformed into relative genomic frequencies. Then, for each species pair, the the similarity value is calculated as the Pearson correlation coefficient between the frequencies of all SSR motives.

Overall, we find that the similarity in SSR motif frequency is negatively correlated with divergence time (Figure 4, Pearson correlation -0.33, $p = 0.006246$). However, some species pairs deviate substantially from this global trend and show much higher (the three species *Dmoj*, *Dvir*, and *Dwil* to each other) or much lower similarities than expected from their divergence time (*Dmel* to all other Drosophila species).

Although the SSR locus conservation rates seem to be influenced by genomic SSR similarities for the three species *Dmoj*, *Dvir*, and *Dwil*, the genomic SSR similarities do not correlate with the SSR locus conservation rates and thus indicate that our measures of SSR locus conservation are not susceptible to highly similar SSR compositions:

- C_{sim}: Pearson cor = -0.2020841, $p = 0.1065$
- C_{pro}: Pearson cor = -0.2410394, $p = 0.05308$

TABLES AND FIGURES

Table 1: Summary of the abundance of SSRs in the twelve Drosophila and five outgroup genomes. For each genome, the number of SSRs, the genome size, and the SSR density (SSRs/Mb) are given.

Species	Version	Group	SSRs	genome size	SSRs/Mb
Anopheles gambiae	P3	Outgroup	113,068	278.2 Mb	406.4
Aedes aegypti	L1	Outgroup	144,538	1,384.0 Mb	104.4
Bombyx mori	2.0	Outgroup	80,641	480.8 Mb	167.7
Culex quinquefasciatus	J1	Outgroup	109,885	579.0 Mb	189.8
Drosophila ananassae	1.3	Drosophila; Sophophora	128,768	231.0 Mb	557.5
Drosophila erecta	1.3	Drosophila; Sophophora	78,883	152.7 Mb	516.5
Drosophila grimshavi	1.3	Drosophila; Drosophila	281,313	200.5 Mb	1,403.3
Drosophila melanogaster	5.25	Drosophila; Sophophora	110,465	168.7 Mb	654.7
Drosophila mojavensis	1.3	Drosophila; Drosophila	340,579	193.8 Mb	1,757.1
Drosophila persimilis	1.3	Drosophila; Sophophora	191,682	188.4 Mb	1,017.6
Drosophila pseudoobscura	2.8	Drosophila; Sophophora	179,837	152.7 Mb	1,177.4
Drosophila sechellia	1.3	Drosophila; Sophophora	91,054	166.6 Mb	546.6
Drosophila simulans	1.3	Drosophila; Sophophora	81,802	137.8 Mb	593.5
Drosophila virilis	1.2	Drosophila; Drosophila	246,246	206.0 Mb	1,195.2
Drosophila willistoni	1.3	Drosophila; Sophophora	306,760	235.5 Mb	1,302.5
Drosophila yakuba	1.3	Drosophila; Sophophora	101,208	165.7 Mb	610.8
Tribolium castaneum	3.0	Outgroup	21,344	170.4 Mb	125.2

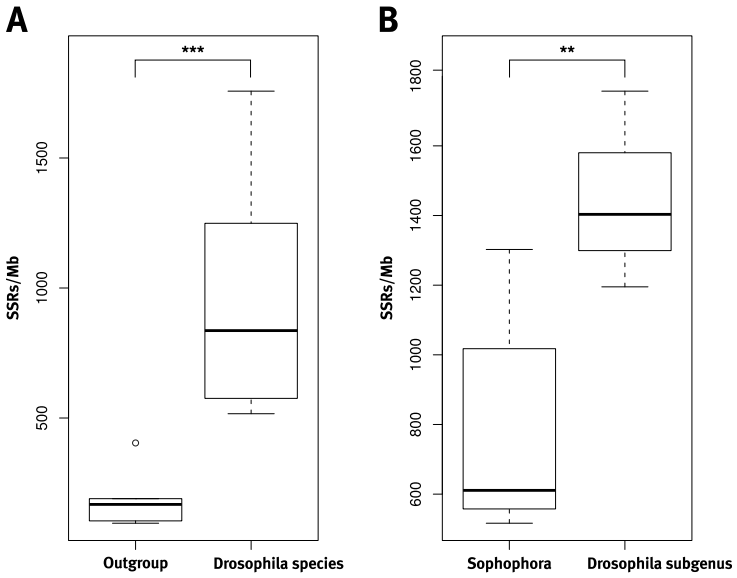

Figure 1: Comparison of full-genome SSR densities, i.e. number of SSRs per Mb, across 12 Drosophila species and 5 outgroup species (see Table 1). **A:** Boxplot comparing outgroup and Drosophila species. **B:** Boxplot comparing between Sophophora and Drosophila subgenus species. Wilcoxon rank sum test was used to determine whether observed differences are significant.

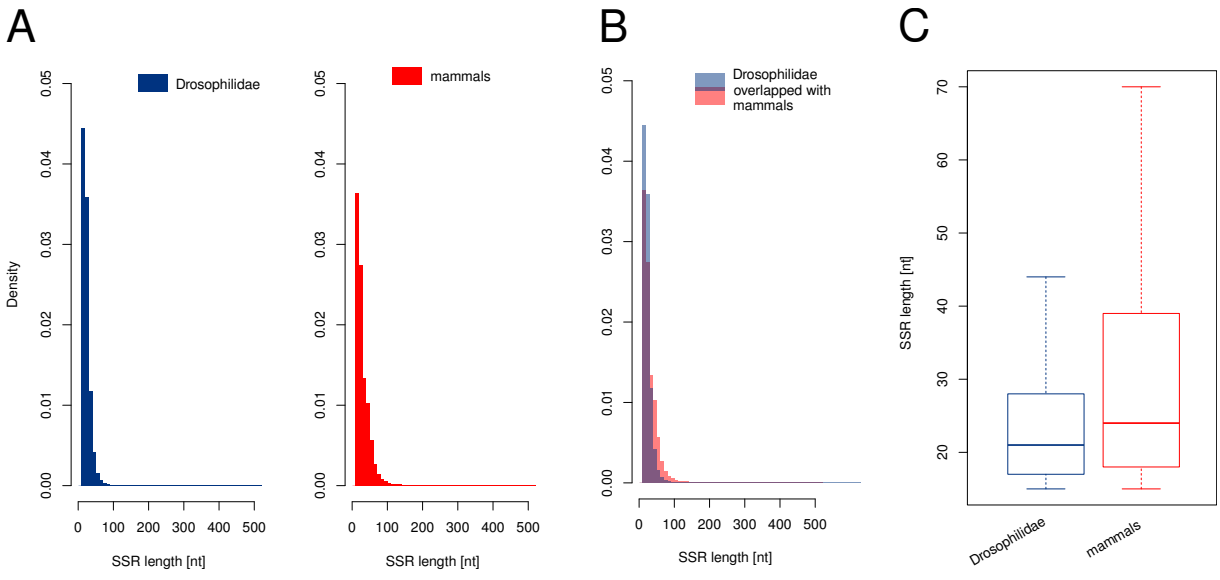

Figure 2: Comparison of SSR length distribution between genomes of Drosophilidae and mammals. **A:** Histogram of the relative frequency of SSRs with a given length in Drosophila (blue) and mammals (red). **B:** Overlapping the two histograms from **A** between Drosophilidae (semi-transparent blue) and mammals (semi-transparent red). **C:** Boxplot of SSR length distribution.

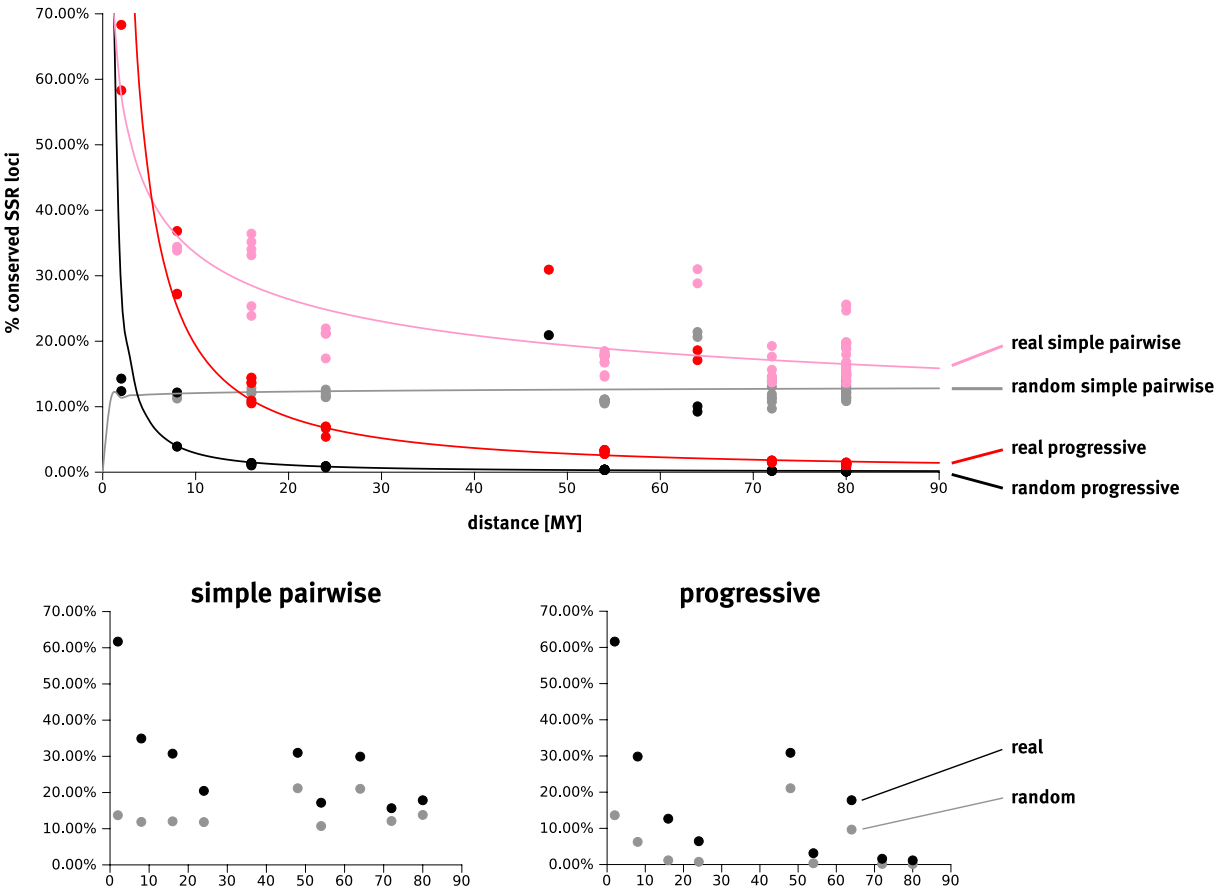

Figure 3: Benchmark of the simple pairwise and the progressive method and comparison between the real and a randomized dataset.

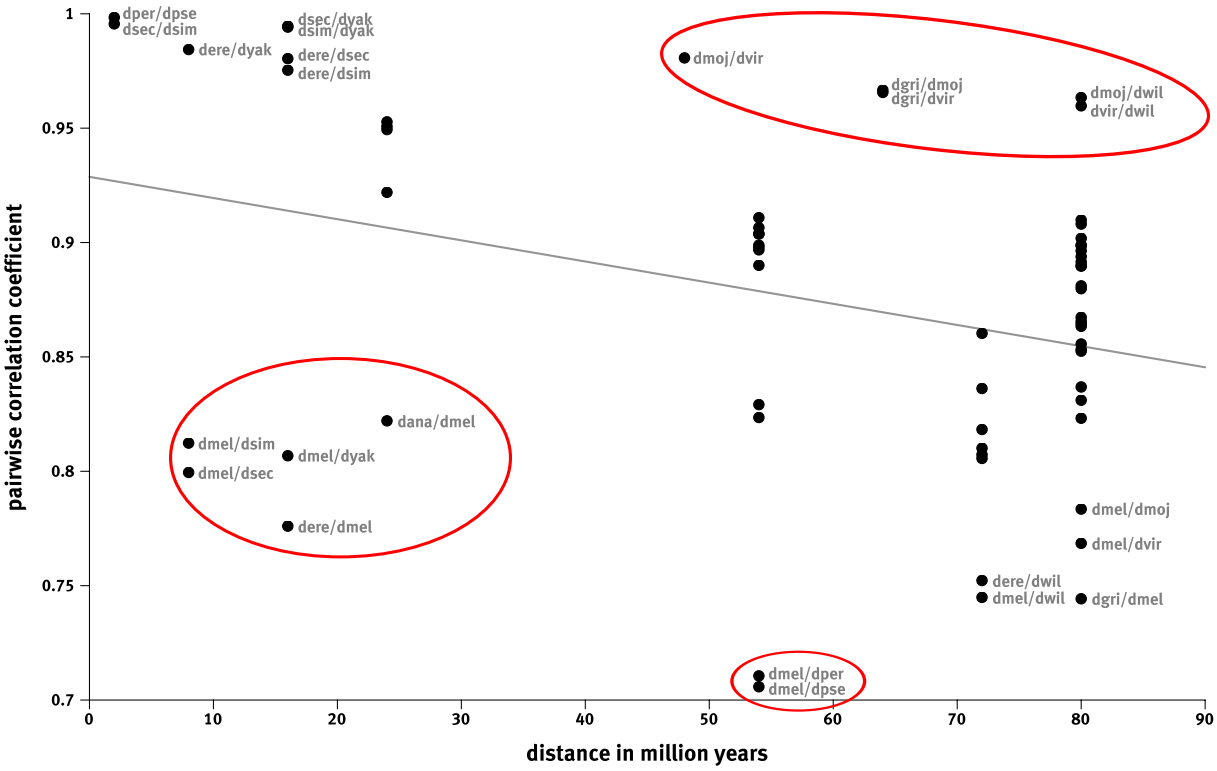

Figure 4: Pairwise similarities in terms of SSR frequencies across all syntenic regions in relation to divergence time. Similarity values are Pearson correlation coefficients obtained from correlating all SSR motif frequencies between any pair of Drosophila species.

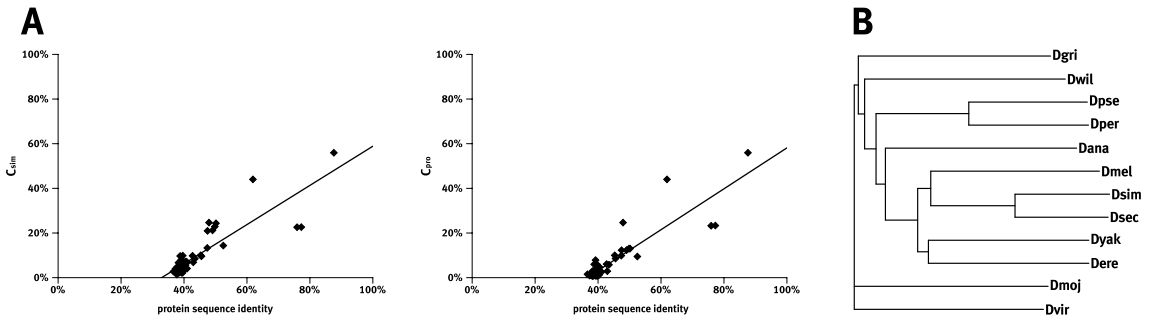

Figure 5: **A:** Pairwise protein sequence identity against SSR locus conservations rates C_{sim_true} and C_{pro_true}. Both C_{sim_true} (Pearson cor $= 0.864$, $p < 2.2e-16$) and C_{pro_true} (Pearson cor $= 0.892$, $p < 2.2e-16$) were found strongly correlated with protein sequence similarities. **B:** Neighbor Joining tree derived from the differences in pairwise C_{sim} values (after correction with the null model).