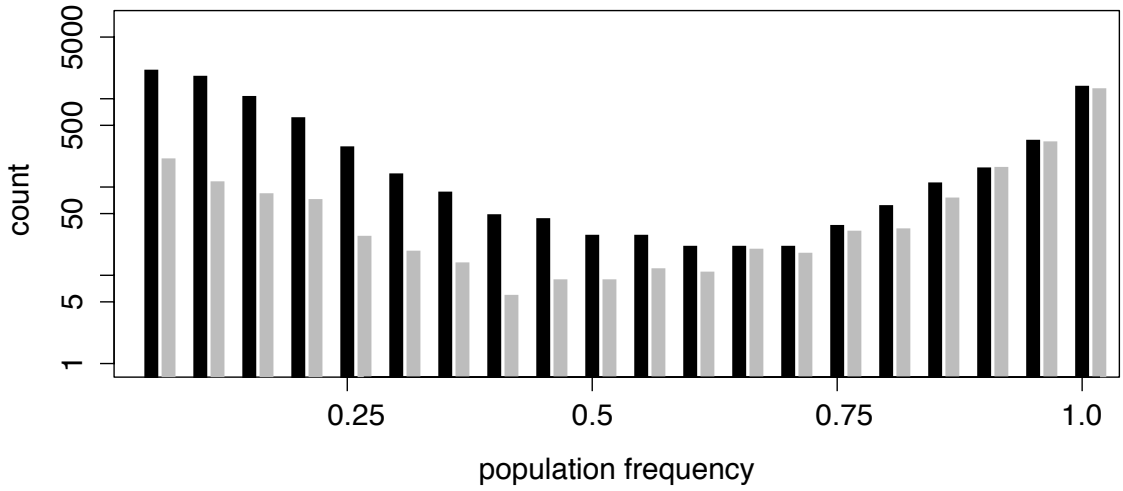
Supplement to 'Low activity of transposable elements in *Drosophila mauritiana*: causes and consequences'

Robert Kofler and Christian Schlötterer

April 17, 2015

Contents

1 Supplementary figures



Supplementary figure 1: Frequency distributions of TE insertions in *D. simulans* (black) and *D. mauritiana* (grey); Only TE insertions for which the population frequencies could be estimated are shown (not overlapping, minimum physical coverage of 10); *D. simulans*: $14,020$ insertions; *D. mauritiana*: $2,491$ insertions

2 Supplementary tables

Supplementary table 1: The abundance of TE insertion in *D. simulans* (Dsim) and *D. mauritiana* (Dmau). Only TE insertions in genomic regions being present in the assemblies of both species are considered. n number of TE insertions; n_e number of TE insertions for which population frequencies could be estimated (not overlapping, minimum physical coverage of 10); n_f number of fixed insertions; chr. arm: chromosome arm

chr. arm	species	length (Mb)	n	density ($\#/Mb$)	n_e	n_f	fixed (%)
genome	Dsim	111.0	8056	72.6	7097	1516	21.4
	Dmau	110.5	2764	25.0	2586	1710	66.1
X	Dsim	20.6	1617	78.3	1410	300	21.3
	Dmau	20.5	566	27.7	531	332	62.5
2L	Dsim	21.1	1489	70.6	1323	304	23.0
	Dmau	21.1	562	26.6	523	345	66.0
2R	Dsim	18.9	1328	70.3	1167	231	19.8
	Dmau	18.9	448	23.8	414	277	66.9
3L	Dsim	22.3	1619	72.7	1419	239	16.8
	Dmau	22.2	461	20.8	422	267	63.3
3R	Dsim	27.0	1617	60.0	1449	176	12.1
	Dmau	26.9	391	14.5	371	201	54.2
4	Dsim	1.1	386	351.4	329	266	80.9
	Dmau	1.1	336	311.9	325	288	88.6

3 Supplementary results

3.1 Local recombination rate difference on 3L

To investigate the influence of the recombination rate on the abundance of low frequency insertions we exploited a local recombination rate difference on chromosome 3L (True et al., 1996). Although the recombination rate is higher in *D. mauritiana*, on chromosome 3L, between polytene band 250 and 500 from the centromere, the recombination rate is actually higher in *D. simulans* (True et al., 1996).

If purifying selection due to higher recombination rate in *D. mauritiana* is the sole force responsible for the observed depletion of low frequency insertions in *D. mauritiana* than the number of low frequency insertions in this region on 3L should be lower in *D. simulans* than in *D. mauritiana*. Unfortunately, it is not trivial to translate the positions of these polytene bands into genomic coordinates of the *D. mauritiana* and *D. simulans* assemblies. We therefore investigated the abundance of low frequency insertions in all 5 Mbp windows on chromosome 3L using a step size 1 Mbp. The region of interest spans about 250 polytene bands. As the number of polytene bands is proportional to DNA content (True et al., 1996) and the total size of chromosome 3L is about 875 bands or 22.3 Mbp ($Dmau = 22.34$, $Dsim = 22.25$), we estimate that the region of interest has a size of about 6.4 Mbp. Therefore at least one window (size 5Mbp with 1Mbp steps) should capture the evolutionary forces acting on TE insertions in the region of interest without incurring noise by including TE insertions from outside the region. For all tested windows the number of low frequency insertions is significantly higher in *D. simulans* than in *D. mauritiana* (Chi-square test; $p < 2.2e-16$ for all windows; supplementary table 2). The recombination rate is therefore not responsible for the observed differences in the abundance of low frequency insertions between *D. simulans* and *D. mauritiana*.

Supplementary table 2: Abundance of low frequency (≤ 0.2) insertion for windows of 5 million base pairs (Mbp) along chromosome 3L for a population of *D. simulans* (Dsim) and *D. mauritiana* (Dmau); start: start position of window in Mbp; end: end position of window in Mbp; p significance of the difference of low frequency insertions between *D. mauritiana* and *D. simulans* for the given window; std.dev.: standard deviation

start	end	Dsim	Dmau	χ^2	p
0	5	187	23	128.1	2.20E-16
1	6	164	15	124.0	2.20E-16
2	7	189	14	150.9	2.20E-16
3	8	201	16	157.7	2.20E-16
4	9	188	24	126.9	2.20E-16
5	10	198	24	136.4	2.20E-16
6	11	198	30	123.8	2.20E-16
7	12	205	33	124.3	2.20E-16
8	13	195	33	115.1	2.20E-16
9	14	230	23	169.4	2.20E-16
10	15	226	24	163.2	2.20E-16
11	16	218	21	162.4	2.20E-16
12	17	216	18	167.5	2.20E-16
13	18	231	16	187.1	2.20E-16
14	19	237	15	195.6	2.20E-16
15	20	251	15	209.4	2.20E-16
16	21	252	18	202.8	2.20E-16
17	22	276	21	218.9	2.20E-16
average		214.5	21.2		
std.dev.		28.2	6.0		

Supplementary table 3: Overview of data sets used in this study. The geographic origin (origin) year of sampling, the study that published the data set, the pool-size, the number of reads (in million), the number of mapped reads (in million) and the average physical coverage (apc) of a TE insertion (in regions being present in all assemblies and having a minimum physical coverage of 10) are shown. Dmau *D. mauritiana*, Dsim *D. simulans*, Dmel *D. melanogaster*

species	origin	year	study	pool-size	reads	mapped	apc
Dmau	Mauritius	2006-2009	Nolte et al. (2012)	152	166.1	159.8	102.2
Dsim	South Africa	2013	Kofler et al. (2014)	793	169.0	146.1	60.3
Dsim	central Africa	2001-2009	Nolte et al. (2012)	50	64.9	58.4	26.6
Dmel	South Africa	2013	Kofler et al. (2014)	554	186.4	181.7	62.0

4 Supplementary material and methods

4.1 Previously published data sets

In this work we relied on publicly available Pool-seq data to estimate TE abundance in populations of *D. simulans*, *D. melanogaster* and *D. mauritiana* (Kofler et al., 2014; Nolte et al., 2012). An overview of the data used in this study can be found in supplementary table 3.

4.2 Estimating TE abundance

Estimating TE abundance with PoPoolation TE requires paired-end Pool-seq data from the population of interest, a TE annotation of the reference genome and a hierarchy of TE insertions, containing for every annotated TE insertion the family and the order (Kofler et al., 2012). We *de novo* annotated TE insertions in the genomes of *D. simulans* (r1.0; Palmieri et al., 2014), *D. mauritiana* (r1.0; Nolte et al., 2012) and *D. melanogaster* (v6.03; dos Santos et al., 2015) as described by Kofler et al. (2014). Briefly, we first obtained a library con-

taining the consensus sequences of *Drosophila* TEs (transposon_sequence_set.embl; v9.42; (Quesneville et al., 2005)) from FlyBase. To avoid identification of spurious TE insertions we excluded canonical TE sequences not derived from *D. melanogaster*, *D. simulans* and *D. mauritiana*. In contrast to our previous work (Kofler et al., 2014) we also included the canonical sequence of *Mariner*, which was first discovered in *D. mauritiana* (Hartl et al., 1997). We mapped the consensus TE sequences against the reference genomes with RepeatMasker open-4.0.3 (Smit et al., 2010) using the RMBlast (v2.2.28) search engine and sensitive search settings (-s). Finally we filtered TE insertions overlapping with microsatellites using SciRoKo 3.4 (Kofler et al., 2007) and bedtools (v2.17.0; Quinlan and Hall, 2010). Overlapping TE insertions of the same family were merged and disjoint TE insertions of the same family were linked. We resolved overlapping TE families by prioritizing the longest insert and retained only TE insertions with a minimum length of 100bp. We obtained the last requirement for PoPoolation TE, a hierarchy of TE sequences from the database of consensus TE sequences (v9.42 see above). We retrieved the sequences of annotated TE insertions from the reference genomes into a distinct file and subsequently masked stretches of TE sequences within reference genomes using the character 'N'. Finally we concatenated the fasta files of the (i) consensus sequences of TE insertions (ii) the TE sequences extracted from the reference genomes and (iii) the repeat masked reference genome into a single file, which we call 'TE-merged-reference'. Note that inclusion of TE sequences extracted from the reference genomes allows to identify diverged TE sequences with PoPoolation TE.

We mapped the short reads to the appropriate TE-merged-reference with bwa (v0.7.5a) (Li and Durbin, 2009) using the bwa-sw algorithm (Li and Durbin, 2010). Paired end information was restored with 'samro' (Kofler et al., 2012). The abundance of TE insertions was measured with PoPoolation TE similarly as described in (Kofler et al., 2012) using the following settings: `identify-te-insertions.pl` –te-hierarchy-level family, –min-count 1, –min-map-qual 15, –narrow-range 100; `crosslink-te-sites.pl` –min-dist 85, –max-dist 300 (400 for the data of Nolte et al. (2012), which have a slightly larger insert size) ; `estimate-polymorphism.pl` –te-hierarchy-level family, –min-map-qual 15; Subsequently we

filtered for TE insertions i.) located on the major chromosome arms (X, 2L, 2R, 3L, 3R, 4) ii) having a minimum physical coverage of 10 (physical coverage as defined here is the sum of paired end fragments that either confirm the presence or the absence of a TE insertion) and iii.) being supported by at least two paired end fragments. To allow for an unbiased comparison of TE abundance we randomly subsampled the physical coverage at each TE insertion to 60. Detailed statistics for every TE family at each step of our bioinformatics pipeline can be found in supplementary file 2 .

4.3 Estimating nucleotide polymorphism

We estimated genome-wide levels of nucleotide diversity in a natural population of *D. mauritiana* (Nolte et al., 2012) and a natural population of *D. simulans* from South Africa (Kofler et al., 2014) using Pool-Seq data (Schlötterer et al., 2014) and PoPoolation (Kofler et al., 2011). First, we aligned all reads to the appropriate reference genome (unmodified) with bwa aln (0.7.5a) (Li and Durbin, 2009) and the following parameters: -I -m 100000 -o 1 -n 0.01 -l 200 -e 12 -d 12; Duplicate reads were removed with Picard (v1.95; `http://picard.sourceforge.net/`). Reads with a mapping quality lower than 20 or reads not mapped as proper pair were removed with samtools (v0.1.19) (Li et al., 2009). We created a pileup file for each population with samtools (v0.1.19) (Li et al., 2009) and the following parameters: -B -Q 0; As alignments spanning indels are frequently unreliable and may lead to spurious SNP calls we removed regions flanking indels (5bp in each direction; minimum count of indel 3) from the pileup with PoPoolation (Kofler et al., 2011). Subsequently we subsampled the pileup to a uniform coverage of 70 with PoPoolation (Kofler et al., 2007) and the following parameters: –max-coverage 1400 –min-qual 20 –method withoutreplace; Finally we calculated π for windows of 100kb using PoPoolation and the following parameters: –min-count 2 –min-coverage 60 –max-coverage 80 –min-covered-fraction 0.6 –min-qual 20 –no-discard-deletions –pool-size 1300;

4.4 Orthologous regions

To allow for an unbiased comparison of TE abundance between species it is necessary to restrict the analysis of TE abundance to regions being present in the assemblies of all investigated species, i.e. orthologous regions. We masked all sequences derived from TEs in both reference genomes (see above) to avoid spurious alignments. We first identified such orthologous regions between *D. mauritiana* and *D. simulans* by aligning the respective reference genomes with with MUMmer (v3.23; nucmer) (Kurtz et al., 2004). Coordinates were extracted with the 'show-coords' tool (Kurtz et al., 2004) and only alignments of the major chromosome arms (X, 2L, 2R, 3L, 3R, 4) were considered. Due to the masking of TE sequences these raw alignments contain a plenitude of gaps where the TE insertions actually causing the gaps, are not present in the alignment. To mitigate this we linked gaps by merging alignments not separated by more than 20.000bp in both species. This threshold of 20.000bp has been arbitrary chosen because in a previous work using *D. melanogaster* (Kofler et al., 2014) we only found six TE regions with a size larger than 20.000bp. Second, we identified orthologous regions in the *D. melanogaster* genome (v6.03; dos Santos et al., 2015) that are also present in the assemblies of *D. simulans* and *D. mauritiana* with the following workflow: we masked regions in the *D. simulans* genome that are not present in the *D. mauritiana* assembly using the character 'N', than we aligned this masked *D. simulans* genome with the *D. melanogaster* reference genome using MUMer (see above) and finally we again filtered for alignments to major chromosomes and linked gaps smaller than 20.000bp.

4.5 TE insertions at similar positions in *D. simulans* and in *D. mauritiana*

TE insertions at similar genomic positions in *D. mauritiana* and *D. simulans* were identified as described previously (Kofler et al., 2014). We first generated a set of TE insertions that may potentially have similar insertion sites between these two species, by reciprocally aligning 1000 bp regions flanking each TE insertion, both at the 5' and the 3' end, to the

reference genomes of *D. simulans* and *D. mauritiana* using bwa-sw (v0.7.5a) (Li and Durbin, 2010). We retained only TE insertions (i) where the flanking regions of one species could be unambiguously mapped to the other species (mapping quality ≥ 15) and (ii) where the flanking regions from the other species could be mapped back to the initial positions. This procedure filters for insertions in non-repetitive regions and insertions that are present in the assemblies of both species. If a TE insertion of the same family was present within the boundaries of these flanking regions in both species, we mark these as insertions at similar sites. Note that this procedure allows for some uncertainty of the exact insertion position [as recommended when using PoPoolation TE (Kofler et al., 2012)].

4.6 Statistical analysis and visualization

For statistical analysis we used the R programming language (R Core Team, 2012). The abundance of TE insertions was visualized with Circos (v0.64) (Krzywinski et al., 2009).

References

dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., et al. (2015). Flybase: introduction of the drosophila melanogaster release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research*, 43(D1):D690–D697.

Hartl, D. L., Lohe, A. R., and Lozovskaya, E. R. (1997). Modern thoughts on an ancyent marinere: function, evolution, regulation. *Annual review of genetics*, 31(1):337–358.

Kofler, R., Betancourt, A. J., and Schlötterer, C. (2012). Sequencing of pooled dna samples (pool-seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS genetics*, 8(1):e1002487.

Kofler, R., Nolte, V., and Schlötterer, C. (2014). Massive bursts of transposable element activity in *Drosophila*. *bioRXiv*.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., and Schlötterer, C. (2011). Popoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS one*, 6(1):e15925.

Kofler, R., Schlötterer, C., and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics (Oxford, England)*, 23(13):1683–1685.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079.

Nolte, V., Pandey, R. V., Kofler, R., and Schlötterer, C. (2012). Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in Drosophila mauritiana. *Genome Research*, 23:99–110.

Palmieri, N., Nolte, V., Chen, J., and Schlötterer, C. (2014). Assembly and annotation of *Drosophila simulans* strains from Madagascar. *Genome resources*, xx:xx.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS computational biology*, 1(2):166–175.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, 15(11):749–763.

Smit, A. F. A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0.

True, J. R., Mercer, J. M., and Laurie, C. C. (1996). Differences in crossover frequency and distribution among three sibling species of drosophila. *Genetics*, 142(2):507–523.