Direct estimation of the spontaneous mutation rate by short-term mutation accumulation lines in *Chironomus riparius*

Ann-Marie Oppold[1,2] and Markus Pfenninger[1,2*]

[1] Senckenberg Biodiversity and Climate Research Centre, Molecular Ecology Group, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

[2] Faculty of Biological Science, Institute for Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Straße 9, 60438 Frankfurt am Main, Germany

Abstract

Mutations are the ultimate basis of evolution, yet their occurrence rate is known only for few species. We directly estimated the spontaneous mutation rate and the mutational spectrum in the non-biting midge *C. riparius* with a new approach. Individuals from ten mutation accumulation lines over five generations were deep genome sequenced to count *de novo* mutations (DNMs) that were not present in a pool of F1 individuals, representing parental genotypes. We identified 51 new single site mutations of which 25 were insertions or deletions and 26 single point mutations. We estimated a mutation rate of 4.2×10^{-9} (95% confidence interval: $2.8 \times 10^{-9} – 6.1 \times 10^{-9}$) which is in the range of recent estimates for other insects and supports the drift barrier hypothesis. We show that accurate mutation rate estimation from a high number of observed mutations is feasible with moderate effort even for non-model species.

Being the ultimate source of genetic variation for evolution to act upon, mutation is certainly among the most important genetic processes. The per generation rate at which spontaneous mutations occur in the genome is the central parameter to estimate the effective population size on recent time scales[1] or in the course of population history[2], equilibrium of genomic base composition[3] and divergence times[4]. Yet, the spontaneous mutation rate (μ) is so difficult to measure directly that it has been rarely estimated up to now. Consequently, only very few eukaryotic direct μ estimates are currently available[5-11], scarcely representing biodiversity. More estimates, in particular of non-model species would be highly desirable because they would shed light on the evolution of μ, its associated ecological and evolutionary circumstances[12] as for example the drift-barrier hypothesis[13] and Lewontin's paradox[14].

Currently, two approaches are applied to directly estimate μ: mutation-accumulation (MA) lines experiments and parent-offspring trios[6,8]. In the MA line approach, inbred lines are established and bred over many generations[15]. Due to the almost absent effect of selection, all except of the most deleterious mutations become eventually fixed and are thus readily identified and confirmed[5]. However, establishing inbred lines is not possible for all organisms, often require complex logistics to transfer generations, intensive care over long time spans and recessively deleterious mutations will be lost with their respective MA-line. In addition, mutator alleles may become fixed, altering the estimated mutation rate[16]. In the trio approach, parents together with their offspring are full genome sequenced[17,18]. This has the advantage that the observed mutational spectrum includes also recessively deleterious mutations as they appear heterozygously in the offspring. Limitations arise from the large number of offspring needed to be screened for an appreciable number of mutations and the requirement to know the parents[6], which is difficult in some species.

In our estimation of μ in the non-biting midge *Chironomus riparius*, we drew on the advantages of both approaches. We established ten MA-lines over five generations and deep sequenced the genomes of a single individual per MA-line. Because individual parenthood is difficult to determine in the swarm breeding *C. riparius,* we compared these individuals with the pooled full-sibling offspring from the single egg-clutch the MA-lines were established from. This yielded an appreciable number of mutations, allowing an accurate estimation of μ.

Results

The mean sequence coverage ranged from 23.8x to 30.6x for the ten MA-line individuals and was 69.9x for the reference pool, representing the parents (Table S1). The number of callable sites for the MA line individuals ranged between 113 and 130Mb, covering up to 85% of the high complex regions of the genome. Overall, we identified 51 mutations (range 4-11 per individual) of which 26 were SPMs (range 2-6, Table 1). Seven mutations were found to be homozygous at least in some of the individuals that were Sanger-sequenced for confirmation. This is not significantly different from the expected mean as inferred from a simulation ($\chi^2 = 0.343$, $p = 0.56$, Table S2). One mutation was in an intron of an annotated gene. This is less than expected from the extent of the gene space (15%) in the draft genome.

Eighteen SPMs were transitions (Ts) and twelve transversions (Tv), resulting in a Ts/Tv ratio of 1.50. There were twice as many G/C to A/T mutations than vice versa (14 to 7), but this difference was not significant ($\chi^2 = 2.333$, $p = 0.13$). We observed eight insertions and 17 deletions, a difference that only marginally deviated from random expectations ($\chi^2 = 3.240$, $p = 0.07$). With few exceptions, all indel mutations were associated with an A or T monomer stretch of at least four positions length (Table S1). With Sanger sequencing, we could confirm the presence of 23 mutations (17 SPM, 2 insertions, 4 deletions) out of 27 selected candidates. By experimental design (Figure S1), we could not confirm mutations arisen in the last generation (Table S2), the proportion of unconfirmed mutations was therefore within expectations. Interestingly, two positions in the Sanger sequences showed mutations to two different bases in different individuals (A2 scaffold 31: 1191434, G>C and G>A and A7 185:145141, G>T and G>A).

Our estimate of the SPM rate was $\mu = 4.23 \times 10^{-9}$ (95% CI: $2.78 \times 10^{-9} – 6.05 \times 10^{-9}$, Figure 1A). Using coalescence estimates of theta from 70 unlinked non-genic regions in European wild populations between 0.023 and 0.034 (unpublished data), we estimate N_e for these populations to range between 1.36×10^6 and 2.01×10^6. The rate for single base deletions was $\mu_{del} = 2.77 \times 10^{-9}$ (95% CI = $1.63 \times 10^{-9} – 4.31 \times 10^{-9}$) and for insertions $\mu_{ins} = 1.30 \times 10^{-9}$ (95% CI = $5.69 \times 10^{-10} – 2.44 \times 10^{-9}$). The total mutation rate for all single base mutations was thus 8.30×10^{-9} (95% CI = $6.26 \times 10^{-9} – 1.08 \times 10^{-8}$, Figure 1A).

Discussion

We here present a direct estimate of the spontaneous mutation rate in *C. riparius*, a valuable resource from a non-model dipteran as additional representative for the vast biodiversity of insects. We were able to confirm the expected proportion of mutation candidates by Sanger sequencing, suggesting a very low false positive rate. Together with the known low false negative rate of the applied bioinformatics pipeline[19], the presented values likely present accurate estimates. The estimated SPM rate reported here is in the range, although at the upper margin, of estimates from both MA-lines and single generation parent-offspring approaches in *D. melanogaster* (2.8×10^{-9} to 5.49×10^{-9})[19 20 21] or *H. melpomene* (2.9×10^{-9})[22], with broadly overlapping confidence intervals. It is however, lower than the recently reported rate for *A. mellifera* [11](, Figure 1B). The estimate of the *C. riparius* effective population size is comparable to both, *D. melanogaster* and *H. melpomene* (~1.4×10^6 or ~2×10^6 for *D. melanogaster* and ~2×10^6 for *H. melpomene*[19,22]). This and the similarity of μ may be taken as support for the drift-barrier hypothesis [23], stating that the realised μ of a species is determined by the balance between selection and drift. The slight increase of the *C. riparius* μ might result from presence of mutator alleles in the more or less inbred laboratory population used, as discussed for one *D. melanogaster* Florida line[21]. However, as the occurrence of mutator and antimutator alleles is stochastic and N_e is subject to changes through time[23], our estimate could as well present normal interspecific variability.

Even though not significantly different, the observed bias towards G/C > A/T mutations is in line with the high A/T content of the *C. riparius* genome (GC content 31%[24]). The phenomenon of two different mutations at the same site in different individuals was recently reported also in *D. melanogaster*[22] and plausibly explained by a mutation cluster early in the germline with a subsequent error-prone repair. Our results indicate that this may be a relatively frequent process, meriting future attention.

The here presented experimental set-up combines short term mutation accumulation lines with the information of the parental genotypes and is thus an efficient approach to estimate μ in many organisms even without high quality reference genomes. While identifying a substantial number of mutations, the effort in terms of time (5 generations of about 28 days each) and deep sequencing of ten individuals and the reference pool appeared reasonable. Yet, over such a short period, all mutations can still occur in heterozygous fashion, thus revealing the full mutational spectrum (Table S2). The applied experimental design is furthermore promising to determine the influence of demographic, and environmental factors and/or anthropogenic substances on the evolutionary relevant germline mutation rate.

Material and Methods

We used a strain of *C. riparius* that was established from a field population in Western Germany and kept since several decades in various laboratories for genetic and ecotoxicological research ("Laufer population"), from which also the *C. riparius* reference draft genome was sequenced[24]. Larvae from a single egg-clutch were raised at 20°C under most permissive conditions concerning space and food[25] to avoid selection. The offspring (F_1) was allowed to reproduce. After successful reproduction, the adults of this first generation were collected to produce the reference pool (see below). Twenty of the clutches were used to establish as many mutation accumulation lines. These lines were reared as described above for additional four generations, always bringing only a single egg-clutch into the next generation. In the fifth generation, a single individual from each of ten randomly chosen MA lines was retained for genome sequencing (Figure S1). Siblings of each sequenced individual were kept for experimental mutation confirmation.

Due to swarm fertilisation of females, it is not possible to unequivocally determine the parents of a particular egg-clutch. To obtain a baseline against which to identify DNMs, we pooled 190 individual head capsules of their F_1 offspring and sequenced their pooled DNA to an expected mean coverage of 60X as 150bp paired-end library on a Illumina HiSeq2500 platform, because the allelic composition of the F_0 parents should be mirrored in the allele frequency of their offspring.

One female individual of each of the ten MA-line was whole-genome sequenced to an expected mean coverage of 25X on an Illumina HiSeq4000 platform. Library preparation of single individuals was performed with the KAPA HyperPrep Kit (KR0961, KAPA Biosystems) to yield enough DNA. The 150bp paired-end reads were individually adapter clipped and quality trimmed, using Trimmomatic[26]. The cleaned reads of MA line individuals and the reference pool were then processed with the GATK-pipline[27], i.e. mapped with bwa *mem* (v0.7.10-r789, [28]) against the reference genome draft (NCBI accession number to be provided), duplicates marked with Picard (v1.119 available at http://picard.sourceforge.net), realignment around indels and recalibration of bases with GATK. The bam-files for all individuals and the F1 pool can be found at (ENA-accession number to be provided).

We established a multistep pipeline in line with, [6] to identify potential DNMs and minimise false positives. Initially, we applied variant calling with the GATK UnifiedGenotyper[29] individually for each MA-line. The reference pool, reflecting the joint genotype of two diploid parent individuals, was treated alike with the exception that variant calling parameters were set as for a tetraploid individual (see above). We then intersected all ten resulting vcf-files amongst themselves and with the reference pool simultaneously, retaining the unique variants for each MA-line, using bcftool *isec* (v1.3, htslib 1.3, available at https://github.com/samtools/BCFtools).

Raw mutation candidates were then quality filtered following the GATK best practices [27]. Only candidate positions with an overall quality score (GQ) concerning base calling quality and position in

read above 90 were considered. Variants with indication for substantial strand bias (SOR > 4) were removed. Since we wanted to concentrate on point mutations and single base indels, indel length was restricted to two. To assure sufficient coverage on the one hand and minimise the effects of mismapped duplicated, paralogous regions on the other, coverage depth was restricted to a range between 15 and 44 reads. A minimum allele count of 5 was required for the non-reference allele to be retained.

The resulting list of candidate positions was then used to create pileups between the respective individual and the reference pool using samtools *mpileup* (SAMtools utilities version 1.1[30]). A custom Python-script screened the reads in the reference pool mapped to this position for presence of the alternative allele in the MA-line individual and, if successful, removed it from the candidate list. All surviving candidate positions were manually curated by visualising each candidate position along with the reference pool in IGV (v2.3.68[31,32]). This was necessary, because candidate positions contained several false positives due to paralog mismapping[6], PCR artefacts escaping duplicate removal and wrongly emitted variant calls.

The described approach has been shown to yield negligible false negatives[6]. Twenty-seven candidate mutations (53%) were checked for presence in the respective MA-line by designing primers for the mutated region and Sanger-sequencing the resulting PCR products for ten full-siblings of the deep sequenced individual. Mutations occurred up to F_3 should show in this sample at least once with a probability of 0.9989. Assuming that probability of occurrence is equal in all generations, we expected to confirm only about 80% of all mutations because those occurring in the germline of the F_4 generation will occur only in a single individual in F_5 (Table S2).

Sites where a mutation could in principal be called according to our criteria were calculated separately for each individual. Assuming a Poisson distribution, we used a Maximum Likelihood method as described in[16] to estimate μ and associated confidence intervals.

References

1 Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**, 195-205 (2009).

2 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

3 Hiroshi Akashi, R. M. K. & Eyre-Walker, A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Mutation and Evolution* **7**, 49-60 (2012).

4 Ho, S. Y. The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.* **29**, 496-503 (2014).

5 Denver, D. R. *et al.* A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *P Natl Acad Sci USA* **106**, 16310-16314, doi:10.1073/pnas.0904895106 (2009).

6 Keightley, P. D., Ness, R. W., Halligan, D. L. & Haddrill, P. R. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a *Drosophila melanogaster* Full-Sib Family. *Genetics* **196**, 313-+, doi:10.1534/genetics.113.158758 (2014).

7 Keightley, P. D. *et al.* Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene*. *Molecular biology and evolution* **32**, 239-243, doi:10.1093/molbev/msu302 (2015).

8 Keightley, P. D. *et al.* Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* **19**, 1195-1201, doi:10.1101/gr.091231.109 (2009).

9 Keith, N. *et al.* High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome Res* **26**, 60-69, doi:10.1101/gr.191338.115 (2016).

10 Ossowski, S. *et al.* The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* **327**, 92-94, doi:10.1126/science.1180677 (2010).

11 Yang, S. H. *et al.* Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463-U187, doi:10.1038/nature14649 (2015).

12 Lynch, M. The Lower Bound to the Evolution of Mutation Rates. *Genome Biol Evol* **3**, 1107-1118, doi:10.1093/gbe/evr066 (2011).

13 Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**, 704-714, doi:10.1038/nrg.2016.104 (2016).

14 Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat Rev Genet* **17**, 422-433, doi:10.1038/nrg.2016.58 (2016).

15 Mukai, T. & Cockerham, C. C. Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **74**, 2514-2517 (1977).

16 Haag-Liautard, C. *et al.* Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**, 82-85, doi:10.1038/nature05388 (2007).

17 Roach, J. C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**, 636-639, doi:10.1126/science.1186802 (2010).

18 Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-U137, doi:10.1038/ng.862 (2011).

19 Keightley, P. D., Ness, R. W., Halligan, D. L. & Haddrill, P. R. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a *Drosophila melanogaster*Full-Sib Family. *Genetics* **196**, 313-320 (2014).

20 Keightley, P. D. *et al.* Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research*, gr. 091231.091109 (2009).

21 Schrider, D. R., Houle, D., Lynch, M. & Hahn, M. W. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* **194**, 937-954 (2013).

22 Keightley, P. D. *et al.* Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol. Biol. Evol.*, msu302 (2014).

23 Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**, 704-714 (2016).

24 Oppold, A.-M. *et al. Chironomus riparius* (Diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence. *bioRxiv*, 080721 (2016).

25 OECD. *Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water*. (OECD Publishing, 2004).

26 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

27 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

28 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

29 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-+, doi:10.1038/ng.806 (2011).

30 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

31 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

32 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (grant number PF390/8-1). We are grateful to Dennis Lüders for Sanger sequencing and Andreas Wieser for statistical support.

Author contributions

MP and AMO designed the study, AMO carried out the experiments, MP performed the bioinformatics analyses, MP and AMO drafted the manuscript.

Materials and Correspondence

Markus Pfenninger (Pfenninger@bio.uni-frankfurt.de)

Table 1. Summary information on the number of callable sites, the total number of single base mutations, resulting mutation rate (μ) per generation and site, the number of single point mutations (SPM) and the associated rate (μ SPM), number of insertions, deletions, transitions (Ts) and transversions (Tv) per mutation accumulation (MA) line.

MA line	number of callable sites	number of mutations	μ	SPM	μ SPM	insertions	deletions	TS	TV
A1	113428649	7	1.23E-08	2	3.53E-09	3	2	3	1
A2	126279134	4	6.34E-09	2	3.17E-09	1	1	2	0
A3	125198906	3	4.79E-09	1	1.60E-09	0	2	1	0
A4	119583116	7	1.17E-08	4	6.69E-09	2	1	2	2
A5	100483488	9	1.79E-08	4	7.96E-09	2	3	0	5
A6	129246425	4	6.19E-09	2	3.09E-09	0	2	0	2

A7	127455137	3	4.71E-09	1	1.57E-09	0	2	2	0
A8	128191107	7	1.09E-08	4	6.24E-09	0	3	2	2
A9	130004076	3	4.62E-09	2	3.08E-09	0	1	2	0
A10	129317412	4	6.19E-09	4	6.19E-09	0	0	4	0
Sum	1229187450	51		26		8	17	18	12

Figure 1. A) Mean and confidence interval estimates for the total single base mutation rate, the single point mutation (SPM) rate, the deletion and insertion rate. B) Comparison of the known SPM rates for insects with confidence intervals where available