# Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene*

Peter D. Keightley,*[,1] Ana Pinharanda,[2] Rob W. Ness,[1] Fraser Simpson,[3] Kanchon K. Dasmahapatra,[3,4] James Mallet,[3,5] John W. Davey,[2] and Chris D. Jiggins[2]

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

[2]Department of Zoology, University of Cambridge, Cambridge, United Kingdom

[3]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[4]Department of Biology, University of York, York, United Kingdom

[5]Department of Organismic and Evolutionary Biology, Harvard University

***Corresponding author:** E-mail: peter.keightley@ed.ac.uk.

**Associate editor**: John Parsch

## Abstract

**We estimated the spontaneous mutation rate in *Heliconius melpomene* by genome sequencing of a pair of parents and 30 of their offspring, based on the ratio of number of de novo heterozygotes to the number of callable site-individuals. We detected nine new mutations, each one affecting a single site in a single offspring. This yields an estimated mutation rate of $2.9 \times 10^{-9}$ (95% confidence interval, $1.3 \times 10^{-9}$–$5.5 \times 10^{-9}$), which is similar to recent estimates in *Drosophila melanogaster*, the only other insect species in which the mutation rate has been directly estimated. We infer that recent effective population size of *H. melpomene* is about 2 million, a substantially lower value than its census size, suggesting a role for natural selection reducing diversity. We estimate that *H. melpomene* diverged from its Müllerian comimic *H. erato* about 6 Ma, a somewhat later date than estimates based on a local molecular clock.**

***Key words:*** **mutation, *Heliconius*, genome sequencing.**

Understanding the process of spontaneous mutation is central for many of the most important questions in evolutionary genetics. The neutral nucleotide diversity expected within a species ($\theta$) is proportional to the product of the spontaneous mutation rate per nucleotide site ($\mu$) and the effective population size ($N_e$). Variation in the mutation rate is therefore expected to contribute to the large range of variation in neutral nucleotide diversity that has been observed in natural populations (Leffler et al. 2012). Conversely, if nucleotide diversity for putatively neutral sites of a population has been estimated, and the mutation rate is known, it is possible to estimate $N_e$. Effective population size is an important factor determining the effectiveness of natural selection, and selection effects on diversity at linked sites could limit diversity in the genome (Charlesworth B and Charlesworth D 2010). Species split times can be estimated using between-species neutral nucleotide divergence, which is also expected to be proportional to the mutation rate. This can be useful if fossil evidence-based dates of species divergence are not available. Estimates of the mutation rate for a range of species across the tree of life are therefore needed in order to better understand patterns of diversity in relation to $N_e$ and the influence of natural selection on variation. However, at present, only a handful of direct mutation rate estimates are available, for a small number of model species.

Mutation rate estimation has until recently depended on assaying rates of mutation at specific loci or on the between-species nucleotide divergence at putatively neutral sites, such as synonymous sites (Drake et al. 1998). There are, however, drawbacks to these approaches, including uncertainty about species divergence dates and nonneutral synonymous site evolution (Chamary et al. 2006). This has led to efforts to directly estimate the mutation rate by sequencing mutation accumulation (MA) lines or outbred parents and their offspring. The MA line approach suffers from potential difficulties, however. For example, recessive mutator alleles might become fixed by inbreeding in the MA line progenitor. Furthermore, its practical applicability is limited, because inbred lines cannot be produced for most species. Sequencing parents and their offspring and searching for de novo mutations in the offspring are more generally applicable, but to date experiments have only been carried out in humans (Roach et al. 2010; Conrad et al. 2011; Kong et al. 2012; Michaelson et al. 2012) and *Drosophila melanogaster* (Keightley et al. 2014).

Both humans and *D. melanogaster* have "finished" genome sequences, but the genomes of most sequenced species are incomplete or "draft" and it is unknown whether parent–offspring sequencing can be applied in such cases. Widely used sequencing technologies produce short reads and spurious base calls arise due to the mismapping of paralogs. Here, we apply parent–offspring genome sequencing to a tropical butterfly species, *Heliconius melpomene*, whose genome sequence is currently draft (Heliconius Genome Consortium 2012). *Heliconius melpomene* has become a focal organism for genome-based studies of speciation and hybridization







Letter

**Table 1.** Mutations Called and Depth of Sequencing Coverage Statistics for the Wild Type (WT) and Mutant (Mut) Bases in the Mutant Focal Offspring Along with Average Read Depth in the Parents and Focal Offspring.

| Contig | Position | Individual | Base Call | | Depth | | Mean Depth | |
|---|---|---|---|---|---|---|---|---|
| | | | WT | Mut | WT | Mut | Parents | Offspring |
| HE671270 | 80778 | 118 | A | T | 5 | 15 | 33.5 | 26.8 |
| HE670334 | 71590 | 33 | T | C | 13 | 11 | 33 | 28.7 |
| HE670118 | 16036 | 103 | G | A | 11 | 14 | 17 | 25.0 |
| HE670855 | 10858 | 37 | A | G | 25 | 11 | 30.5 | 32.7 |
| HE668834 | 189330 | 1 | G | A | 26 | 23 | 58 | 35.6 |
| HE671384 | 187868 | 4 | T | A | 29 | 17 | 40.5 | 30.8 |
| HE672075 | 836004 | 118 | G | A | 4 | 12 | 17 | 15.6 |
| HE679870 | 4463 | 33 | G | C | 24 | 13 | 53 | 34.2 |
| HE671010 | 9591 | 4 | T | A | 21 | 21 | 24.5 | 22.0 |

(Martin et al. 2013), and an accurate estimate of the mutation rate will have several immediate applications. We use deep sequencing of parents and 30 offspring to produce an estimate of the mutation rate that is close to one recently obtained by similar means in *D. melanogaster* (Keightley et al. 2014).

## Results

Among the 30 offspring, we sequenced 13 focal offspring at a high depth (mean = 26.3, SD = 7.1; supplementary table S1, Supplementary Material online) and 17 "bait offspring" at a lower depth (mean = 12.7, SD = 3.9). Bait offspring were used to remove regions prone to alignment errors that generate false positives by excluding sites at which any of these individuals had an alternate base call. Mutations were not called in the bait offspring nor did they contribute to the number of site-individuals. In 4,309 scaffolds of the draft genome assembly, there are $2.70 \times 10^8$ sites, of which $1.23 \times 10^8$ (46%) were estimated to be callable, yielding $1.60 \times 10^9$ site-individuals.

We used the Genome Analysis Toolkit (GATK; Depristo et al. 2011) to call mutations. We assume that reads having the alternate allele at a site present in multiple individuals are due to mismapping. This arises when a paralogous locus present in the sample, but not in the reference genome, contributes reads that map to the wrong place (Li 2011). We assume that such mismapping is equally likely to occur at mutated and unmutated sites.

We applied the mutation calling rules described in Materials and Methods to the GATK genotype calls, yielding 15 candidate mutations appearing as de novo heterozygotes in up to two focal offspring (supplementary table S2, Supplementary Material online). We first examined each candidate using the Integrative Genomics Viewer (Thorvaldsdóttir et al. 2012) to determine whether there are single nucleotide polymorphisms (SNPs) in complete association with the alternate base calls at the candidate sites, a characteristic of mismapped paralogs (Li 2011; Keightley et al. 2014). Two closely linked candidates 17 bp apart on contig HE668478 (individual 110; supplementary fig. S1, Supplementary Material online) and candidates HE671028 and HE669561 (individual 110; supplementary figs. S2 and S3, Supplementary Material online) met this criterion. Furthermore, in each case reads containing the alternate allele are truncated and have unmapped mate pairs. We judged these four candidates as likely false positives caused by mismapping. We then attempted to verify the 11 remaining candidates by Sanger sequencing. Ten gave clearly interpretable chromatograms, confirming that eight are genuine mutations, and that candidates HE671439 and HE672001 are false positives (supplementary table S1, Supplementary Material online). Further attempts at sequencing the remaining candidate (HE671010) were unsuccessful, but mutant-bearing reads are well aligned and have aligned mate pairs (supplementary fig. S4, Supplementary Material online), suggesting that it is genuine. Thus, there are nine apparently genuine de novo mutations, eight confirmed by Sanger sequencing (table 1), each one affecting a single site and present in a single focal individual.

Among the nine de novo mutations, the number of transitions exceeds the number of transversions, as is usually observed in eukaryotes. The number of mutations divided by twice the number of callable site-individuals yields an estimated mutation rate (uncorrected for false negatives) of $2.8 \times 10^{-9}$ (95% confidence interval = $1.3 \times 10^{-9}$–$5.3 \times 10^{-9}$, assuming that the number of mutations is Poisson distributed).

To estimate the frequency of false negatives, we simulated synthetic mutations by modifying sequencing reads for randomly selected sites in the focal offspring. We realigned and analyzed the modified data using the same procedures as for the real data. Among 1,000 synthetic mutations, 456 occurred at callable sites where all other focal offspring, both parents and all bait offspring were pure. Of the callable mutations, 436 (96%) were called. The small proportion of uncalled mutations presumably reflects mutant-bearing reads mapping less frequently than reference reads (fig. 1). A corrected estimate for the mutation rate is therefore $2.9 \times 10^{-9}$ (95% confidence interval = $1.3 \times 10^{-9}$–$5.5 \times 10^{-9}$).

## Discussion

We estimated the mutation rate per base pair by genome sequencing of parents and offspring in *H. melpomene*. The incomplete state of the genome causes difficulties in identification of de novo mutations, because paralogous reads map

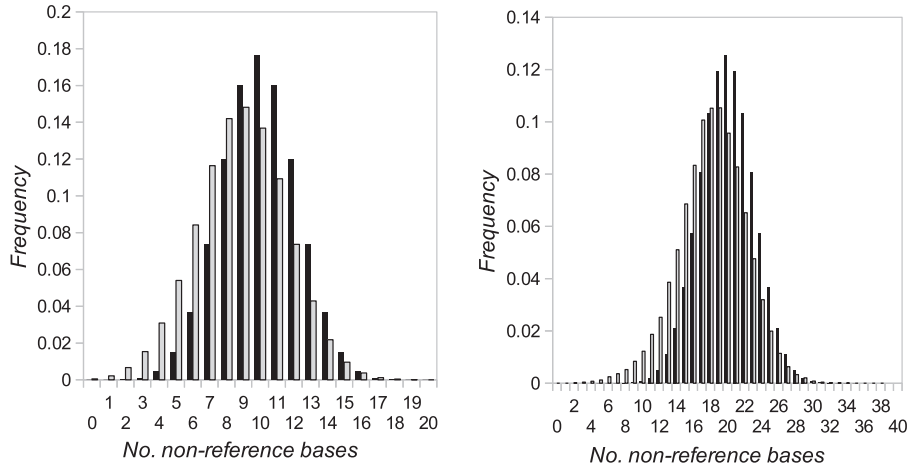


**Fig. 1.** Examples of observed frequency distributions (red or gray) and binomial distributions with parameter 0.5 (blue or black) of alternate (i.e., nonreference) base number at heterozygous sites in the focal offspring having (*A*) 20 reads and (*B*) 40 reads.

more frequently to the wrong location, often yielding false heterozygote calls. Disregarding impure sites affecting any bait offspring and more than two focal offspring effectively addressed this problem. It has been estimated that approximately 20% of spontaneous recessive sex-linked lethal mutations in *D. melanogaster* males occur as premeiotic clusters (Woodruff and Thompson 1992). In the present experiment, we detected no mutation clusters affecting two focal offspring, but in view of the small number of mutation events detected, the sequences of many more individuals will be needed to accurately estimate the rate of premeiotic cluster mutations in *H. melpomene*. The draft state of the genome precluded the detection of large-scale de novo variants, such as rearrangements and duplications, which are particularly sensitive to mismapping.

Autosomal nucleotide diversity ($\pi$) at 4-fold degenerate sites in *H. melpomene* is approximately 2.4% (Martin SH, unpublished data). Assuming neutral synonymous sites evolution, and equating $\pi$ to $4N_e\mu$, $N_e$ for the species is therefore approximately 2 million. This will be an underestimate if selection reduces diversity at 4-fold sites. However, estimates of $N_e$ for *D. melanogaster* based on this approach are of similar magnitude (Keightley et al. 2014), but they are orders of magnitude smaller than both species' census population sizes. Similar diversities and effective population sizes are consistent with the small range of genetic diversity across eukaryotes (Leffler et al. 2012), suggesting a role for processes such as genetic draft limiting diversity (Maynard-Smith and Haigh 1974; Gillespie 2001; Leffler et al. 2012).

Estimates of $\mu$ can also be used to date species divergences, assuming that neutral nucleotide divergence $d = 2\mu t$, where $t$ is the divergence time in generations. For example, synonymous divergence between *H. melpomene* and its Müllerian comimic *H. erato* corrected for diversity within *H. melpomene* is 14% (Martin SH, unpublished data), yielding $t = 23$ million generations, which will be an underestimate if selection reduces synonymous site divergence. Assuming four generations per year, the divergence date is approximately 6 Ma, which is somewhat more recent than that estimated from a fossil-calibrated phylogeny of 10–13 Ma (Kozak et al. 2014). Although our data suggest that current estimates of the age of the *Heliconius* radiation are approximately correct, further work will be required to reconcile these estimates. It remains to be seen whether the hypothesis that the early radiation of *Heliconius* coincided with a time of rapid uplift in the Andes about 10 Ma is supported.

This is only the second direct estimate of the mutation rate per base pair in insects, and the first in Lepidoptera. There have been several direct estimates in *D. melanogaster*, by Denaturing High Performance Liquid Chromatography (Haag-Liautard et al. 2007), by whole-genome sequencing of MA lines (Keightley et al. 2009; Schrider et al. 2013), and most recently by parent–offspring sequencing (Keightley et al. 2014). There is significant variation among these estimates, but most are close to $3 \times 10^{-9}$, which is remarkably close to our estimate of $2.9 \times 10^{-9}$ for *Heliconius*. We have demonstrated that it is possible to estimate the mutation rate by offspring–parent genome sequencing for the case of a draft genome sequence. It should soon be possible to address the question of whether this lack of variation in the mutation rate extends to other arthropod groups whose draft genome sequences are now becoming available.

## Materials and Methods

### Cross Sequencing

The cross was previously used to produce chromosomal scaffolds of the *H. melpomene* genome (Heliconius Genome Consortium 2012, supplementary material section S4). After four generations of inbreeding, a male *H. melpomene melpomene* from the same lineage as for the *H. melpomene* reference genome was crossed with a female *H. melpomene rosina* from a laboratory strain established from Gamboa, Panama. DNA from two $F_1$ parents and 30 of their $F_2$ offspring was extracted using the DNeasy Blood and Tissue Kit (Qiagen). Illumina TruSeq libraries (300 bp insert size) were sequenced using 100-bp paired-end reads on an Illumina HiSeq2500.

### Alignment to Reference Genome

Reads for parents and offspring were aligned to version 1.1 of the *H. melpomene* genome (available on Ensembl and from http://butterflygenome.org, last accessed November 5, 2014) using SMALT version 0.7.0.1 with default options. Output sequence alignment/map (SAM) files were converted to binary format (BAM) files, sorted and annotated with read groups using Picard version 1.84 (http://picard.sourceforge.net/, last accessed November 5, 2014).

### Genotype Assignment

Each individual's BAM file was processed to remove duplicates using Picard tools, then to realign indels using GATK. SNPs were called using the GATK UnifiedGenotyper across all individuals simultaneously to produce a variant call format (VCF) file, assuming a heterozygosity parameter of 0.01. For high read depth, genotype calls are insensitive to this parameter (Depristo et al. 2011; Ness et al. 2012).

### Mutation Calling

We processed the VCF by a similar algorithm as described previously (Keightley et al. 2014) filtering sites as follows:

1) Not marked as low quality (GATK LowQual).
2) Read depth of both parents $\geq$10, both homozygous references, containing no alternate allele reads.
3) None of the 17 bait offspring contains alternate allele reads.
4) The genotypes of all 13 focal offspring are defined (i.e., are called by GATK).
5) At most two focal offspring are called as heterozygous by GATK.
6) No other focal offspring contains alternate allele reads.

Our method excludes sites containing alternate alleles in either parent, which precludes the identification of mutations at polymorphic sites. Assuming that mutations are not more frequent at polymorphic sites, this should reduce the number of mutations and callable sites proportionally. There was no filtering carried out on read depth of the focal or bait offspring. Heterozygotes called among the focal offspring were marked as candidate mutations.

### Synthetic Mutations

We estimated the proportion of false negatives (genuine mutations we failed to call) by simulating mutations in the *Heliconius* data, running our pipeline, and calculating the fraction of simulated mutations called. Synthetic mutations were simulated by modifying the reads overlapping a random site in a focal offspring. We sampled the number of reads to be altered from empirical distributions of numbers of nonreference base calls at heterozygous sites (see below). For each synthetic mutation, we randomly sampled a genomic position *b* and a focal offspring. We sampled a random integer *y* from the frequency distribution of nonreference base number for the individual's read depth (e.g., see fig. 1). We then changed *y* reference bases to a different randomly selected base by modifying reads overlapping position *b* in the individual's BAM file.

We generated 1,000 synthetic mutations in the BAM files of focal individuals, extracted all reads from the modified BAM files, and aligned the modified reads to the reference genome by the procedure used for the original data. We then applied the mutation-calling algorithm, exactly as described above, with the exception that filters were not applied to the focal offspring carrying the synthetic mutation. The fraction of callable simulated mutated sites estimates the fraction of callable sites in the genome. Uncallable sites will include, for example, sites of low mapping quality and sites where genotypes are undefined in one or more focal offspring.

### Frequency Distributions of Nonreference Read Number in Heterozygotes

To produce the distributions used to generate synthetic mutations, we identified a set of sites heterozygous for natural polymorphisms, regardless of the genotypes called from the sequencing data, taking advantage of the lack of recombination in *Heliconius* females. $F_2$ offspring receive whole chromosomes from the $F_1$ mother, so SNPs from the same chromosome have identical segregation patterns in the offspring, and segregation patterns for each of the 21 *H. melpomene* chromosomes for this cross are known (Heliconius Genome Consortium 2012). We identified SNPs from each chromosome by compiling sites called as heterozygous in the F1 mother, homozygous in the F1 father, and matching one of the chromosome segregation patterns for the bait offspring. Heterozygous focal offspring could then be identified based on segregation pattern, without reference to their sequenced genotype. We used these heterozygous focal offspring to tabulate frequency distributions of numbers of nonreference base calls for read depths 1, . . . 100 (see fig. 1).

### Sanger Sequencing

With the exception of four candidates ruled out by inspection (see Results), we checked all candidates by Sanger sequencing. We sequenced the focal individual and one control individual on both strands. If the initial sequencing failed, an alternative primer pair was tried.

## Data Accessibility

Whole-genome sequence data for the parents and the 30 offspring from the mapping cross used for this study are available from the European Nucleotide Archive, study accession PRJEB7581.

## Supplementary Material

Supplementary figures S1–S4 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The authors thank Krzysztof Kozak, Simon Martin, and two anonymous referees for helpful comments. This work was funded by Biotechnology and Biological Sciences Research

Council (BBSRC) (H01439X/1) to C.J., the Herchel Smith Fund to J.W.D., and the BBSRC to P.D.K. and R.W.N.

## References

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7: 98–108.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Greenwood Village (CO): Roberts & Co.

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet.* 43:712–714.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491–498.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148:1667–1686.

Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55:2161–2169.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445:82–85.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487:94–98.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196:313–320.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 19: 1195–1201.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471–475.

Kozak KM, Wahlberg N, Neild A, Dasmahapatra KK, Mallet J, Jiggins CD. Forthcoming 2014. Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* 10: e1001388.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 11:1817–1828.

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

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151: 1431–1442.

Ness RW, Morgan AD, Colegrave N, Keightley PD. 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics* 192:1447–1454.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* 194:937–954.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14:178–192.

Woodruff RC, Thompson JN Jr. 1992. Have premeiotic clusters of mutation been overlooked in evolutionary theory? *J Evol Biol.* 5: 457–464.