Supporting information

Hellsten et al. 10.1073/pnas.1319032110

Sequencing and Assembly of the Reference Sequence *Mimulus guttatus* V1

A high-quality draft reference genome for *Mimulus guttatus* was produced using a conventional Sanger-based whole-genome shotgun approach, using DNA from the *M. guttatus* IM62 inbred line. Multiple libraries with a range of insert sizes were generated and sequenced at both ends using Sanger dideoxy sequencing by the Department of Energy Joint Genome Institute, Walnut Creek, CA, using standard protocols (www.jgi.doe.gov/sequencing/protocols/prots_production.html). Three 3.3-kb insert-size plasmid libraries (3,558,614 reads, 2.4588 Gb total sequence), two 6.6-kb insert-size plasmid libraries (1,805,879 reads, 1.2301 Gb total), and one 7.9-kb insert-size plasmid library (499,072, 331.6 Mb total) were sequenced at both ends. Six fosmid libraries (1,198,751 reads, 758.6 Mb), 33.7–36.1-kb insert size, and two BAC libraries (147,456 reads, 101.7 Mb), 64.9- and 105.2-kb average insert size, were sequenced on both ends with Sanger sequencing for a total of 7,209,772 Sanger reads of 4.8808 Gb of high-quality bases.

Genome Assembly and Construction of Pseudomolecule Chromosomes. The sequence reads were assembled using our modified version of Arachne v.20071016 (1) with parameters maxcliq1 = 100, maxcliq2 = 50, remove_duplicate_reads = True, and BINGE_AND_PURGE = True. This produced 5,588 scaffold sequences, with a raw scaffold N50 of 1.0 Mb and a raw contig N50 of 41.9 kb, 418 scaffolds larger than 100 kb, and total genome size of 348.7 Mb. Scaffolds were screened against bacterial proteins, organelle sequences, and the National Center for Biotechnology Information nonredundant nucleotide sequence database and removed if found to be a contaminant. Additional scaffolds were removed if they (*i*) consisted of >95% 24mers that occurred four other times in scaffolds larger than 50 kb, (*ii*) contained only unanchored RNA sequences, (*iii*) contained only unanchored alternative haplotypes consisting of >95% shared content with chromosomes, or (*iv*) were less than 1 kb in length. The final assembly contains 2,216 scaffolds (17,831 contigs) that cover 321.7 Mb of the genome with a final contig N50 of 45.5 kb and a final scaffold N50 of 1.1 Mb.

Completeness of the euchromatic portion of the genome assembly was assessed using 140,292 Sanger-sequenced *M. guttatus* ESTs. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The *M. guttatus* ESTs were aligned to the assembly using BLAT (2) (Parameters: −*t* = dna −*q* = rna −extendThroughN), and alignments ≥90% bp identity and ≥85% coverage were retained. The screened alignments indicate that 131,495 (94.9%) of the ESTs aligned to the assembly.

Annotation of *Mimulus* V1. Three complementary approaches were used to comprehensively identify and characterize transposable elements: (*i*) structure-based methods for full-length LTR retrotransposons (3), Helitrons (4), and nonautonomous cut-and-paste transposable elements (TEs) including MITEs (5); (*ii*) homology-based methods for autonomous cut-and-paste TEs and LINEs [the DDE domain of the five superfamilies and the LINE RT domain were retrieved from previous studies (6, 7) and used as query in tblastn (8) search]; and (*iii*) de novo repeat detection (RepeatModeler; www.repeatmasker.org/RepeatModeler.html) for SINEs and other TEs missed from the previous two approaches. All full-length elements were then classified into families, and exemplar sequences from each family were generated to mask the entire genome by RepeatMasker (www.repeatmasker.org).

TEs make up 49.8% of the *M. guttatus* genome. They represent all known plant TE types, including LTR retrotransposons (superfamily *Copia* and *Gypsy*), non-LTR retrotransposons (LINEs and SINEs), five superfamilies of cut-and-paste DNA TEs (*Tc1-mariner*, *hAT*, *CMC* [*CACTA*], *PIF-Harbinger*, and *MULE*), and *Helitrons*. The most abundant TEs in the *M. guttatus* genome include *Copia* and *Gypsy* LTR elements (each group accounts for ~10% of the genome content), *MULE* cut-and-paste TEs (11.56% of the genome), and *Helitrons* (6.29%). The high abundance of *Helitrons* is exceptional among plant genomes, where *Helitrons* usually occupy less than 3% of the genome space. Also noteworthy is the dominance of nonautonomous *MULE*s (including miniature inverted-repeat transposable elements [MITEs]) over elements in the *Tc1-mariner* and *PIF-Harbinger* superfamilies ("stowaway" and "tourist," respectively), which are numerically the most abundant TEs in well-studied grass genomes.

Protein-Coding Genes. Transcript assemblies were made by PASA (9) from *Mimulus guttatus* ESTs (~555K, IM62 and DUN, including some 454 EST reads) and ESTs of related species (~256K, other asterids). Loci were determined by BLAT alignments of the above transcript assemblies and/or BLASTX alignments of peptides from *Arabidopsis thaliana*, rice, and grape genomes or a few tomato, potato, and tobacco peptides to repeat-soft-masked *M. guttatus* genome. Gene models were predicted by homology-based predictors, mainly by FGENESH+ (10), with an additional one by GenomeScan (11) if FGENESH+ misses the locus. Predicted genes were UTR-extended and/or improved by PASA. Filtered gene set was made from PASA-improved gene models based on ESTs support or peptide homology support subjected to filtering of repeats/transposable elements. Protein domains (PFAM and Panther) were assigned to filtered gene models, and gene models whose protein contains ≥30% TE domains were further filtered out to yield 26,718 loci of protein-coding genes. Additional statistics on this annotation (v1.1) are included in Tables S1 and S2.

Population Sample Collection. We collected seeds from a total of 98 individual plants in May 2008. The plants were collected from four locations within a 15-km radius near Copperopolis, CA, in the Sierra Nevada foothills (latitude 37.9° N, longitude 120.6° W) from two populations growing in soils contaminated by copper mining from 80 to 150 y ago and from two populations growing in ephemeral stream environment uncontaminated by copper. Seeds were sown into 2.5′′ pots with Fafard 4P potting soil, and after germination, we maintained only a single seedling from each maternal family. The plants were grown in the Duke University greenhouse with 18-h day lengths at a constant temperature of 20 °C. Thirty days after germination, we collected tissue from individual plants. We pooled equal amounts of bud tissue from 20 to 31 individual (diploid) plants for each population. We extracted genomic DNA from the pooled samples using a CTAB/chloroform protocol (12).

Data Analysis. Pooled DNA was sequenced using Illumina paired-end sequencing with insert size of ~200 bases and read length of 75 bases. The reads were aligned to the IM62 reference genome assembly using bwa (13), and only bases of quality (Q) 30 or better within reads aligning as proper pairs with mapping quality $q \geq 29$ were used in the analysis. Ninety four and a half percent of all reads could be mapped, with 79.6% mapping as proper pairs. The mean coverage of the genome by such reads is about 255×. The coverage varies substantially with position on the genome as exemplified by a representative 50-kb region in Fig. S1 and, with respect to annotated genes, in Fig. S2. Regions with uncalled bases ("Ns," e.g., the large gap around position 750 kb) are not represented; neither are extremely repetitive regions with low mapping score or regions where large-indel polymorphisms with respect to the reference sequence prevent alignment. Positions in the vicinity of such regions also have depressed depths due to our strict requirement that the reads align as proper pairs, as some of the paired mates fail to align. Finally, regions with higher than average GC content tend to be underrepresented by Illumina sequencing (data not shown), which also leads to variation in depth. This is expected behavior when dealing with short read alignment of a diverse population. The main point of including this figure is to illustrate that useable bases for SNP calling and recombination testing are ubiquitously available in the genome, within at most a few kilobases from any position in the assembled part of the genome.

We detected SNPs by parsing a file in the samtools mpileup format (14) of these alignments. SNPs were defined as sites with Q30 depths ranging from 58 to 450 (determined by looking at the depth histogram), exhibiting exactly two variants, with no reads suggesting any insertions or deletions at the site, and a minor allele frequency (MAF) of at least 5%. More than 9.43 million sites with MAF ≥ 5% were found (14.9 million if we allow MAF ≥ 2%). The size of the assembled genome that could be tested for SNPs under these criteria is 111.3 Mb or 37% of all bases in the assembled genome (excluding Ns). The full folded allele frequency spectrum of this population is shown in Fig. S5. SNP calling in this manner, in particular for low MAF, is vulnerable to incorrect alignments around indels. In particular, bwa will by default not open a gap while aligning bases within 5 bp from the start or end of a read but would either soft clip or align with mismatches, the latter which could lead to a false-positive variant call. Such potentially problematic SNPs are detected and eliminated at the next step in the analysis described below.

Sequencing of pooled data can be viewed as sampling the underlying 2N = 196 haploid genomes with replacement, causing the actual number of independent genomes to be smaller than the read depth: $2N_{\text{sampled}} = 2N\,[1 - \exp(-d/2N)]$, where 2N = 196 and d is the read depth. To sample the equivalent of 50 independent genomes, the read depth must be about 58. (We note that although it is standard practice in handling diploid genotypes to test SNPs for Hardy–Weinberg equilibrium, this is not possible with our pooled samples.)

Parsing the aligned reads in the SAM format, we next identified pairs of SNPs that were covered by individual reads to allow pairwise haplotypes to be determined. We required that both variants of each SNP be confirmed by reads with interior aligned bases, i.e., bases not within the first or last five bases in the read. This eliminated any potential false-positive SNP calls due to misalignments of read ends near indels. A subset of these SNP pairs were used for four-gamete testing if they met the following additional requirements:

i) None of the SNPs must be amino acid changing (to eliminate obvious candidates for positive selection).
ii) Pairwise haplotype information exists for at least 50 independent chromosomes (haploid genomic regions), corresponding to a read depth of at least 58. For uniformity in interpretation and comparison with population genetic models, SNP pairs with haplotype information for more than 50 chromosomes were down-sampled to 50.
iii) $50 \times \text{MAF}_1 \times \text{MAF}_2 \geq 1$. That is, at least one copy of the rarest haplotype would be expected in the sample, if it exists. This last requirement eliminates many combinations of rare SNPs with little or no power to detect recombination.

With these criteria, 11.5 million SNP pairs can be defined across the genome, 3.5 million of which span nonoverlapping regions on the genome. Three important parameters that summarize a SNP pair are

i) the distance between the SNPs (d_H)
ii) the frequency of the least common allele $f_{\min}$ of the four alleles present, and
iii) whether the SNP pair passes the four-gamete test.

We expect the probability (*F4*) of passing the four-gamete test to increase with d_H and $f_{\min}$. The larger the distance is between the SNPs, the larger the chance of a gene conversion tract boundary falling between them. Also, lower values of $f_{\min}$ imply that the SNP pair samples more recent history, because rarer alleles, on average, result from more recent mutations, and a SNP pair cannot start registering recombination events until the most recent of the two mutations that created the SNPs has happened.

Testing for Method Robustness. Several conditions in the data could, in principle, conspire to produce the observed profiles in Figs. 2 and 4 in the main article, even if there were no variation in recombination rates:

i) Systematic variation in coverage depth at positions relative to genes. For example, if the coverage is significantly lower in introns than exons, this might be due to certain haplotypes being unaligned in introns, which could depress the number of observed recombinations.
ii) Locally increased mutation rates at CpG dinucleotides could cause inflated estimates of recombination in regions such as exons which have higher than average CpG content.
iii) "Nested SNP pairs," i.e., two or more SNP pairs crossing the same gene conversion boundary, could cause multiple counting of a single event.
iv) For SNP pairs with $50^*\text{MAF}_1{}^*\text{MAF}_2 \sim 1$, i.e., close to the predicted detection limit (condition 3 above), the fourth haplotype could be absent by chance, reducing the power of the four-gamete test.
v) Systematic variation of the minimum allele frequency $f_{\min}$ of a SNP pair could cause *F4* to exhibit variation, even at a constant underlying recombination rate.

To investigate whether *i*) is an issue, we calculated average coverages upstream of coding DNA sequence (CDS) starts, introns and exons relative to the same annotated gene set used for Fig. 4 in the main paper. The results are shown in Fig. S2. Introns have on average 20–30% less coverage than coding exons. We argue that this difference is not the cause of the systematically lower recombination in introns shown in Figs. 4 and S3, for the following reasons: first, if omission of haplotypes by alignment were common, we would expect a large number of SNP pairs to exhibit fewer than three 2-SNP haplotypes. In the actual data, more than 93% of the 11.5 million SNP pairs feature at least three haplotypes. Second, recombination rates show a strong gradient with respect to gene position, which is not observed in the coverage variation. Finally, a recombination event resulting in a new haplotype (and passing of the four-gamete test) will not in itself reduce the probability of corresponding reads being able to align to the reference sequence, unless accompanied by additional deletions or mutations.

To address *iii*) through *iv*) we selected a very strict subset of 699,685 SNP pairs, or 6% of the 11.5 million pairs, that only features SNPs with "transversions" (which are immune to CpG-enhanced mutation rates), were nonoverlapping, and satisfied a stricter criterion of $50 \times \text{MAF}_1 \times \text{MAF}_2 \geq 2$. Fig. S4 below shows the equivalent of Fig. 2 in the main article, but using only the stricter *F4* as a function of position from nearest CDS start.

For increased detail, we display only the region of ±2 kb around CDS starts. Clearly, the observed peak around the CDS start is still present and not due to any of the potential artifacts. In even more detail, Fig. S3 compares the results in Fig. 4 of the main paper to those based on a subset of these SNP pairs featuring only transversions. The characteristic pattern of polarity and elevated recombination in exons relative to introns remains.

To address *v*) we also show in Fig. S4 the average minimum allele frequency $f_{\min}$ as a function of distance to CDS start. Here we have used the entire set of 11.5 million SNP pairs. As can be seen, this frequency has a very small, but visible increase as we move into the first coding exon. This is probably related to nonneutral evolution of some synonymous mutations due to codon bias-related selection. However, as can be seen, $f_{\min}$ remains largely constant at positions above +300 bp from the CDS start, whereas *F4* continues to drop. Hence, *v*) cannot be the main cause of the observed *F4* profile.

Coalescent Simulations. To place observations into the context of a population genetic evolutionary model, we developed a PERL script to perform coalescent simulations represented as samples of 50 individual pairs of nucleotides separated by a distance of d_H bases and with a Kimura two-parameter mutation model with transition/transversion rate $\kappa = 4$. The effective population size history is chosen so as to be consistent with observed values of the population diversity θ and the allele frequency spectrum (AFS) (Fig. S5), the latter which is consistent with an effective population size growing exponentially from an origin in the distant past: $\mathrm{N} = \mathrm{N}_0\ \mathrm{e}^{-\beta t}$, $\beta = 4$. The mutation rate μ is not entered as separate input to the simulations, as only the product $\mu\mathrm{N}_0$ is fixed by the observed θ, and the recombination rate is specified in dimensionless units as ρ/μ, i.e., in units of the mutation rate. Should the need arise for absolute values of the recombination rates, effective population size, or coalescence times in years, which is not essential for the present work, a reasonable value of the mutation rate would be $\mu \sim 10^{-8}$ base^{-1} gen^{-1}, with a generation time of 1 y.

The exponentially growing effective population size simply means that the gene tree shows a more star-like genealogy than that suggested by models with constant N (Fig. S6). The underlying cause for the steeper AFS is likely population substructure caused by geographic constraints. Our choice of model parameters enables us to reproduce genealogies consistent with the observed AFS, regardless of cause.

The one remaining free parameter is then the recombination rate, defined as the probability per generation that the two sites in the probing SNP pair will become separated due to the cut and paste of a gene conversion tract. This can be written as $(\rho/\mu)d_H$, where ρ is the recombination rate per base per generation. We performed simulations for a large grid of values of $(\rho/\mu)d_H$ which we transformed into a lookup table of *F4*, the probability of observing four gametes in the SNP pair, as a function of $(\rho/\mu)d_H$ and $f_{\min}$, where $f_{\min}$ is the allele frequency of the least frequent allele for the pair of SNPs under consideration. This table was subsequently used in assigning average recombination rates for subsets of SNP pairs, as the recombination rate that, given the actual values of d_H and $f_{\min}$ for the probes ,would predict an expected number of SNP pairs with four haplotypes that equals the number observed.

Hotspot and Cold Spot Detection. Local recombination rates were inferred using sliding windows of nonoverlapping SNP pairs, and for each such set the recombination rate that best matches the observed F4 was inferred from the coalescent lookup table. As a compromise between achieving good spatial resolution and limiting the sampling noise, we chose window sizes of 15 SNP pairs, which typically span over 200–300 bases. The average recombination rate, defined as the rate that would be consistent with the total F4 for all SNP pairs in the genome, is $\rho/\mu = 0.8$. As a null hypothesis of constant recombination rate we assigned pass/fail values to all probes based on Monte Carlo simulations using probabilities from the coalescent lookup table and $\rho/\mu = 0.8$. The subsequent sampling in 15-base windows of the simulated data demonstrates a relatively high noise level for such small window sizes, but the real data show many more high-recombination areas than the simulated, inconsistent with constant recombination rates. If we then define hotspots as local peaks with values $\rho/\mu \geq 5$ and including areas of both side of the peak where $\rho/\mu \geq 2$, an automated scan of the data results in 21,501 hotspots, whereas the simulated data yield 8,455, suggesting a false positive rate of 39% and about 13,000 bona fide detectable hotspots. Higher specificity comes at a cost of lower sensitivity. For example, using $\rho/\mu \geq 15$ as the cutoff, the false positive rate is 3.5%, with a total of 3,235 hotspots detected. In a similar manner, we cataloged cold spots, defined as regions with $\rho/\mu \geq 0$, including flanking regions with $\rho/\mu \leq 0.4$, and requiring lengths of at least 200 bases.

1. Jaffe DB, et al. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13(1):91–96.
2. Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12(4):656–664.
3. McCarthy EM, McDonald JF (2003) LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* 19(3):362–367.
4. Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci USA* 106(31):12832–12837.
5. Han Y, Wessler SR (2010) MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* 38(22):e199.
6. Yuan YW, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci USA* 108(19):7884–7889.
7. Kapitonov VV, Tempel S, Jurka J (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* 448(2):207–213.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
9. Haas BJ, et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654–5666.
10. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 10(4):516–522.
11. Yeh R-F, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Res* 11(5):803–816.
12. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19:11–15.
13. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
14. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

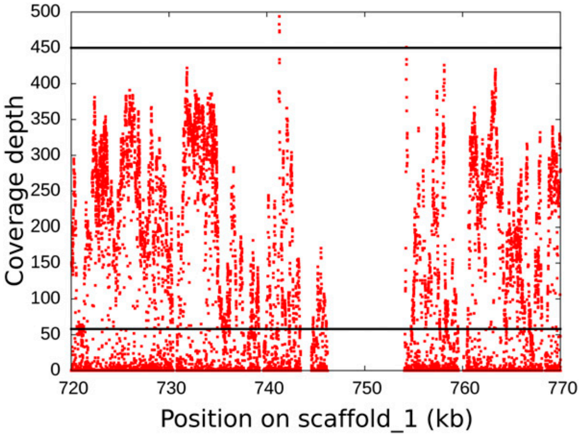

Fig. S1. An example of the variation of population sequence aligment coverage of perfectly aligned read-pairs with mapping quality $q \geq 29$ and base quality $Q \geq 30$ across a 50 kb segment of the reference genome. This is intended to exemplify the typical behavior across the entire genome. Solid lines indicate cutoffs for sites used for SNP calling.

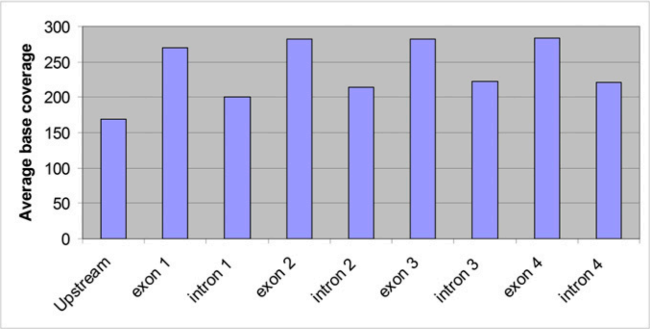

Fig. S2. Average coverage at positions relative to genes. Coverage is somewhat lower outside coding exons due to the greater variation and possibilities of indels here.

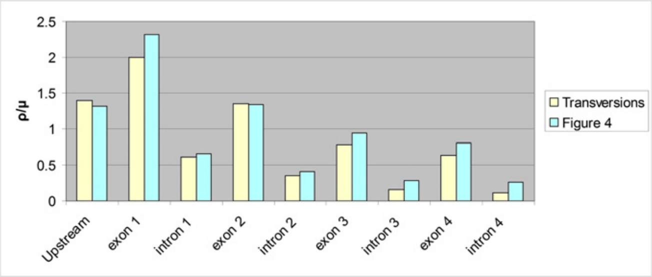

Fig. S3. A comparison of the data in Fig. 4 in the main article to a similar calculation restricted to the much smaller dataset of transversions only (non-overlapping SNP pairs).

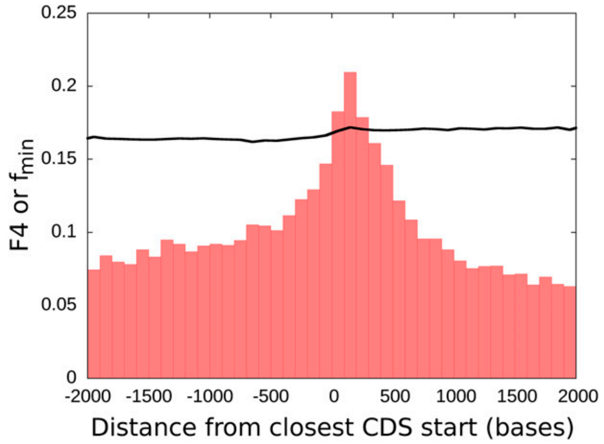

Fig. S4. A remake of Fig. 2 in the main article using only nonoverlapping SNP pairs featuring transversions only and with higher predicted power. The profile is suggested by the larger data set. Also shown in this figure (black line) is the mean frequency of the least common allele (using all SNP pairs; see discussion in *SI Text*).

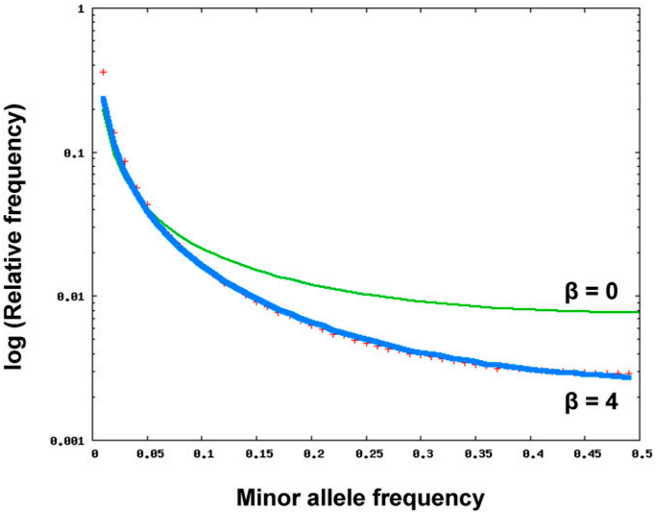

Fig. S5. Allele frequency (red +), with best fit to constant (green) and exponentially growing (blue) population size. Red crosses show observed folded allele frequency spectrum in the *Mimulus* population. Green curve shows predicted spectrum for a coalescent model with constant effective population size N_{eff} ($\beta = 0$), whereas the much better fitting blue curve assumes exponentially increasing N_{eff} ($\beta = 4$).

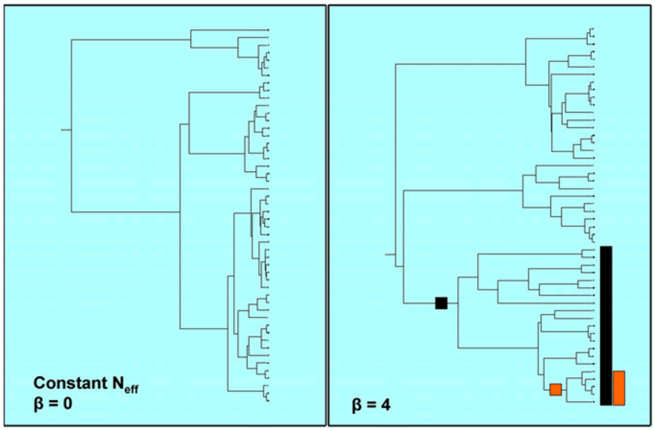

Fig. S6. Examples of coalescent trees with constant and growing population size. With growing population size, trees for the Northern California *Mimulus* population have more star-like form, with fewer short terminal branches. It is an inherent property of the bifurcating tree structure that mutations occurring at early epochs tend to be represented by alleles of higher frequency in a population sample than more recent mutations (black and orange bars).

Table S1. Summary statistics for *Mimulus guttatus* annotation v1.1

Feature	Count
Primary transcripts (loci)	26,718
Alternative transcripts	1,564
Total transcripts	28,282
Average number of exons	5.2
Median exon length (bases)	148
Median intron length (bases)	129

The intron and exon numbers are for primary transcripts only.

Table S2. Gene model support for annotation v1.1

Type of evidence or support	Number of gene models
Any EST support	16,329
EST support over 100% of their lengths	8,114
EST support over 95% of their lengths	9,095
EST support over 90% of their lengths	9,655
EST support over 75% of their lengths	10,713
EST support over 50% of their lengths	12,503
Peptide similarity coverage of 100%	1,149
Peptide similarity coverage of over 95%	9,809
Peptide similarity coverage of over 90%	11,853
Peptide similarity coverage of over 75%	13,983
Peptide similarity coverage of over 50%	15,084
Pfam annotation	18,710
Panther annotation	14,733
KOG annotation	11,069
KEGG Orthology annotation	4,181
E.C. number annotation	2,228