Supplementary Methods

Genome sequencing and alignment

Prior to alignment, the FASTQ quality scores of the reads from all four runs of the Illumina GA II were converted from Illumina 1.3+ to Sanger format. Using the BWA software, suffix array coordinates for the sequence reads were constructed by allowing a maximum edit distance equal to 5% of the total seed length of 30 bp. The resulting reference-based assemblies for all four lanes were then ordered by reference genome coordinates and merged into a single compressed sequence alignment/map (BAM) file, with labeled individual read groups and samples using the SAMTOOLS software (Li *et al.* 2009). Lastly, PCR duplicates were removed from the alignment using the PICARD software package (http://picard.sourceforge.net). A BAM file of the entire alignment can be downloaded from http://kimura.biology.rochester.edu/seqdata/ngs/simclade/simclade.bam.

For a site to be considered a true variant in the alignment, the Phred-scaled variant score must be greater than or equal to 25 in one or more of the three species. The Phred-scaled variant score is calculated by considering the minimum of either the base quality or mapping quality for read *i*. This minimum quality is the probability of error associated with each base, ε_i (Li *et al.* 2008). Given that the reference base occurs *k* times in a sample of *n* reads and the most frequent alternative base occurs *j* times, the probability that the consensus base at that site is identical to the reference is binomially distributed as

$$P_{ref} = \binom{n}{j} \prod_{i=1}^{j} \varepsilon_i \times \prod_{i=1}^{k} 1 - \varepsilon_i \ . \qquad [1]$$

The Phred-scaled variant score is then equal to $-10 \times \log_{10}(P_{ref})$ (Ewing and Green 1998). Therefore, a variant score of 25 corresponds to a type I error rate of approximately 3.16×10^{-3}. Lastly, individual sites with a derived mutation called as heterozygous in a single species are excluded from the analysis.

Phylogenetic analysis

Phylogeny was inferred using the RAxML software using a general time reversible model of nucleotide substitution. To measure node support, we performed 1000 nonparametric bootstrap replicates. We did not include a parameter estimating the proportion of invariant sites because the gamma distribution already allows for sites with low substitution rates, previous work has shown the interaction between these two parameters can lead to inaccurate estimates due to non-independent optimization (Sullivan *et al.* 1999). For each data set that was analyzed, we performed a likelihood ratio test to compare the maximum likelihood inferred topology against the null model of a polytomy (*i.e., D. simulans, D. sechellia* and *D. mauritiana* with zero-length internodes). Because RAxML cannot calculate the likelihood of multifurcating trees, we calculated the likelihood of the maximum likelihood and polytomy trees using the R package PHANGORN (Schliep 2011), using the same parameters as RAxML. The probability that the polytomy better fit the data than the maximum likelihood tree was determined by calculating the ratio of the polytomy and ML tree likelihoods which was then evaluated as χ^2 distributed with 2 degrees of freedom.

Global allopatric model of species divergence

The exact likelihood of the global allopatric model from the site type frequencies can be calculated using coalescent theory. The expected proportions of each of the six different site types are derived under the null allopatric model. For example, the probability of a site being exclusive to the outgroup (P_{EO}) can be written as,

$$P_{EO} = \frac{xT_2 + 1 - \frac{e^{T_1 - T_2}}{3}}{T_1(\alpha + \beta + 1) + [(T_2 - T_1)(x+1)] + 3}, \qquad [2]$$

where

$$x = \begin{cases} 1 & \text{if outgroup is } D.\,simulans \\ \alpha & \text{if outgroup is } D.\,mauritiana \\ \beta & \text{if outgroup is } D.\,sechellia \end{cases}$$

Similarly, the probability of a site type that is exclusive to either of the ingroup species is

$$P_{EI} = \frac{yT_1 + 1 - \frac{e^{T_1 - T_2}}{3}}{T_1(\alpha + \beta + 1) + [(T_2 - T_1)(x+1)] + 3}, \quad [3]$$

in which y =1 when the ingroup species probability being considered is *D. simulans*, and $y = \alpha$ when the ingroup species is *D. mauritiana*, and $y = \beta$ when it is *D. sechellia*. The probability of the two different types of site that are shared between the outgroup and either of the ingroup species is

$$P_{SO} = \frac{\frac{e^{T_1 - T_2}}{3}}{T_1(\alpha + \beta + 1) + [(T_2 - T_1)(x+1)] + 3}. \quad [4]$$

While the probability of a shared derived site between the two ingroup species is

$$P_{SI} = \frac{(T_2 - T_1) + \frac{e^{T_1 - T_2}}{3}}{T_1(\alpha + \beta + 1) + [(T_2 - T_1)(x+1)] + 3}. \quad [5]$$

By considering the probability of a site type as the expected fraction of the length of the total coalescent genealogy and assuming an infinite sites model of mutation, it is no longer necessary to account for the mutation rate at a particular site (Nielsen 2000). It should also be noted that when $T_1 = T_2$, it is expected that $P_{SO} = P_{SI}$ and that when $T_1 = T_2$ and $\alpha = \beta = 1$, then $P_{EO} = P_{EI}$. The likelihood of the allopatric model parameters can then be calculated as a multinomial probability using the observed site type counts and the probabilities from equations 2-5. The analytical approach presented above assumes all sites in the genome are independent and, therefore, in linkage equilibrium.

For the simulation based approach to estimating the likelihood of the divergence model in windows across the genome, prior probability distributions were assigned to two nuisance variables, θ (population mutation rate) and ρ (population recombination rate) per window (site type frequencies in different windows are assumed to be independent). A gamma prior distribution was used for both variables, $\theta \sim$ Gamma(20, 2) and $\rho \sim$ Gamma(5, 0.25). Likewise, the species trees were also sampled with equal probability. The parameter space was covered by

10^8 coalescent replicates, each consisting of 2×10^4 windows. Each replicate represents a random draw from the bounded parameter space. Furthermore, each window needed a minimum of 20 variable sites to be considered for inclusion in the analysis. The coalescent simulations were generated using a modified version of the MS program (Hudson 2002), and it should be noted that the α and β parameters in this case are no longer the ratios of effective population size, but are ratios of lineage-specific mutation rates. Lastly, both of the above approaches were implemented separately for the autosomes and the *X* chromosome, due to the presumed difference in effective population size between these two compartments of the genome.

Test of complex speciation

Likelihood ratio tests of the global and local parameters of the divergence model were carried in both 1-kb and 5-kb windows across the genome. Likelihoods were estimated using the exact probabilities given in equations 2-5. This analytical approach to estimating the local likelihood was employed because the simulation-based heuristic approach to estimating likelihood was not feasible due to the required computational time. The likelihood ratio statistic is assumed to be chi-square distributed with five degrees of freedom because there are six parameters that are free to vary in the locally fit model: each of the three possible species tree topologies, the time to exchangeability (τ), α, and β. Source code for the C++ program to perform the likelihood based analysis is available from the website http://kimura.biology.rochester.edu/software/4sp/4sp.tar.gz. Lastly, to account for variability among regressors in the relative importance analysis of site type frequency and *p*-values, the bootstrapping approach of Davison and Hinkley (1997) was used.

False discovery rate

Rather than adopting the approach of Williamson *et al.* (2007), which is a method for selecting a tuning parameter in the model of Storey and Tibshirani (2003), we estimate the overall proportion of null *p*-values (π_0) using both an lower (λ_1) and an upper bound (λ_2),

$$\pi_0 = \frac{\#\{p_i > \lambda_1; i = 1, \ldots w\}}{w\left(1 - \frac{\lambda_1}{\lambda_2}\right)}, \qquad [6]$$

in which *w* is the total number of windows that occur below the upper bound, λ_2, of *p*-values.

For most distributions of *p*-values we used, the region between $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$ appeared uniformly distributed, and thus most representative of null *p*-values. After implementing equation 6, the FDR with threshold *t* can be estimated as

$$\mathrm{FDR}(t) = \frac{\pi_0 \, wt}{\#\{p_i \le t\}}. \qquad [7]$$

Finally, the expected proportion of false positives among all windows that are more extreme than a window with a given *p*-value is

$$q(p_i) = \min_{t \ge p_i} \mathrm{FDR}(t). \qquad [8]$$

To estimate the above *q*-value, the FDR at 1000 equally spaced points for *t* were evaluated for each observed *p*-value. The above procedure allows us to express the number of significant windows as a function of a given critical *p*-value.

Coding sequence analysis

The 8,563 single-copy ortholog sequences were downloaded from FlyBase (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/melanogaster_group.guide_tree.longest.cds.masked.tar.gz). After alignment of these CDS to the genome alignment of the three *D. simulans* clade species, orthologous coding sequences for each species were generated by concatenating the high-scoring segment pairs (HSPs) within each blast hit and inserting gaps as appropriate. Following alignment of HSPs from the three species, insertions relative to the *D. melanogaster* sequence were deleted to preserve reading frame and sequences with alignment errors were identified by translating the reading frame and filtering out those with premature stop codons. Sites that differ within the *D. simulans* clade were checked against the *D. melanogaster* outgroup to determine the ancestral state. If a variable site had no match to *D. melanogaster*, that site was also excluded from the analysis. Finally, any codons with missing/ambiguous data or alignment gaps were excluded from the analysis. We excluded 17 genes due to lack of a significant BLAST hit in at least one of the *D. simulans* clade species, 217 due to alignment errors, and 87 due to updates in gene prediction models, such as collapsing two CDS into a single transcript or withdrawals from FlyBase.

For analysis of d_N/d_S ratios, pseudocounts were used for any d_N or d_S equaling zero, such that all genes could be included in the analysis. Specifically, a single synonymous or nonsynonymous substitution was assumed and a Jukes-Cantor distance was calculated.

Codon usage bias was assessed by identifying synonymous substitutions between the reconstructed ancestral sequence and the extant sequence for each of the three *D. simulans* clade species. Synonymous substitutions were categorized as either "preferred", "unpreferred", or "equivalent". However, for codons with more than one substitution between the ancestor and the extant species sequences, the averages for each type of change were computed across all accessible mutational pathways (*i.e.*, excluding pathways that include stop codons). To quantify the relative rate of change in each lineage for nonsynonymous and each of the three synonymous mutation types, we calculated the relative excess of changes (λ) between species *i* and species *j* at each of the four coding mutation types:

$$\lambda = \frac{s_i n_j}{s_j n_i} - 1, \qquad [9]$$

in which *s* is the number of changes along each of the lineages, and *n* is the number of coding sites in each lineage.

References

Davison A, Hinkley DV. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8**: 186-194.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337-338.

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

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.

Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**: 931-942.

Schliep KP. 2011. PHANGORN: phylogenetic analysis in R. *Bioinformatics* **27**: 592-593.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**: 9440-9445.

Sullivan J, Swofford DL, Naylor GJP. 1999. The effect of taxon samplling on estimating rate-heterogeneity parameters of maximum-likelihood models. *Mol Biol Evol* **16**: 1347-1356.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *Plos Genet* **3**: e90.