

## RESEARCH ARTICLE

# Population Genomic Inferences from Sparse High-Throughput Sequencing of Two Populations of *Drosophila melanogaster*

Timothy B. Sackton,<sup>\*1</sup> Rob J. Kulathinal,<sup>\*1</sup> Casey M. Bergman,<sup>†</sup> Aaron R. Quinlan,<sup>‡§</sup>  
Erik B. Dopman,<sup>\*</sup> Mauricio Carneiro,<sup>\*</sup> Gabor T. Marth,<sup>‡</sup> Daniel L. Hartl,<sup>\*</sup> and Andrew G. Clark<sup>||</sup>

<sup>\*</sup>Department of Organismic and Evolutionary Biology, Harvard University; <sup>†</sup>Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom; <sup>‡</sup>Department of Biology, Boston College; <sup>§</sup>Department of Biochemistry & Molecular Genetics, University of Virginia School of Medicine; and <sup>||</sup>Department of Molecular Biology and Genetics, Cornell University

Short-read sequencing techniques provide the opportunity to capture genome-wide sequence data in a single experiment. A current challenge is to identify questions that shallow-depth genomic data can address successfully and to develop corresponding analytical methods that are statistically sound. Here, we apply the Roche/454 platform to survey natural variation in strains of *Drosophila melanogaster* from an African ( $n = 3$ ) and a North American ( $n = 6$ ) population. Reads were aligned to the reference *D. melanogaster* genomic assembly, single nucleotide polymorphisms were identified, and nucleotide variation was quantified genome wide. Simulations and empirical results suggest that nucleotide diversity can be accurately estimated from sparse data with as little as  $0.2\times$  coverage per line. The unbiased genomic sampling provided by random short-read sequencing also allows insight into distributions of transposable elements and copy number polymorphisms found within populations and demonstrates that short-read sequencing methods provide an efficient means to quantify variation in genome organization and content. Continued development of methods for statistical inference of shallow-depth genome-wide sequencing data will allow such sparse, partial data sets to become the norm in the emerging field of population genomics.

## Introduction

With the recent emergence of new sequencing approaches that enable biologists to sample genomes at an unprecedented scale (Mardis 2008), a new challenge arises to develop research programs that best leverage these technologies for the next generation of evolutionary questions. For population geneticists studying multiple samples from a single species, such rapid and reliable high-throughput sequencing has the potential to provide unprecedented levels of genome-wide polymorphism data at relatively low cost. In parallel, the advancement of computational resources, including increased memory capacity, storage accessibility, multinode processing, and advanced bioinformatics workflows, has enabled biologists to manage massive genomic data sets. The transition from either full sequencing of selected loci or genotyping many previously ascertained single nucleotide polymorphisms (SNPs), to sequencing entire genomes across many individuals, is an important step in population genetic inference. One advantage is that the potential for ascertainment biases both from surveying previously ascertained SNPs (Clark et al. 2005) and from sampling limited genomic regions (Mousset and Derome 2004) is greatly reduced. Early resequencing efforts (e.g., Andolfatto 2001) were often biased toward regions that are conserved, found in single copy, or thought to show nonneutral patterns of variation when studied as electrophoretic variants. Genome-wide surveys of variation also provide internal controls for the consequences of demography because changes in population size or mating patterns should affect the entire genome,

unlike selective forces that are locus specific (Wall et al. 2002; Ometto et al. 2005). Ultimately, a population genomics approach will provide information about the relative effects of genetic drift versus natural selection, and the inference of these evolutionary forces can be globally normalized against the effects of bottlenecks, subdivision, and demography.

Although next-generation sequencing presents a formidable advance in population genomics, a significant limitation in sampling depth remains for organisms with large genomes. Because sequence reads are typically short and genomic regions are not targeted, random sequencing results in a loose patchwork of sparsely aligned regions. Further confounding the problem is that the error rate per nucleotide for a given single read can be considerably higher than Sanger sequence traces (Mardis 2008; Quinlan et al. 2008). This paper determines how well one can make population genetic inferences with shallow read depth using a modest number of individual genomes from the same species. Using a Roche/454 GS-20 platform, we sequenced nine inbred genomes of *Drosophila melanogaster*, representing two populations: an African population from Malawi ( $n = 3$ ) and a North American population from North Carolina ( $n = 6$ ). These data can be readily placed into a rich context, as much is known about the population genetics and biology of *D. melanogaster* (Powell 1997), particularly these two divergent populations (Andolfatto 2001). In addition, its genome ( $\sim 180$  MB in total, of which 120 MB is euchromatin) is well assembled and expertly annotated (Celniker et al. 2002; Misra et al. 2002; Wilson et al. 2008).

*Drosophila melanogaster* has long been a model system for studying how patterns of population genetic variation are shaped by demography and selection (Begun and Aquadro 1994; Andolfatto 2001; Glinka et al. 2003; Orenge and Aguade 2004; Haddrill et al. 2005; Hutter et al. 2007; Singh et al. 2007). The sparse short-read data

<sup>1</sup> These authors contributed equally to this work.

Key words: *Drosophila*, population genomics, next-gen sequencing, transposable elements, copy number polymorphism, nucleotide diversity.

E-mail: tsackton@oeb.harvard.edu.

*Genome. Biol. Evol.* 1(1):439–455. 2009

doi:10.1093/gbe/evp048

Advance Access publication November 18, 2009

set presented here provides the first opportunity to examine patterns of natural variation on a genome-wide scale in *D. melanogaster* and complements recent population genomic studies in *Drosophila simulans* based on low-coverage Sanger sequencing (Begun et al. 2007). In particular, we focus on population genomic variation across inferred functional classes of nucleotides, chromosomes, and geographic populations. We also discuss correlates of variation across the genome, including recombination rate, and take advantage of a large body of previous work that allows for a robust validation of sparse data inference. Additionally, we investigate the utility of sparse short-read data for studying structural variants such as transposable element (TE) sequences and copy number polymorphisms (CNPs). This work clearly shows the high value of sparse short-read data for population genomic inference (Branscomb and Predki 2002) and raises many important considerations for the use of next-generation sequencing technologies in population genetics, particularly in contexts (such as for nonmodel organisms and organisms with large genomes) where sequencing many genomes to high coverage is not yet feasible.

## Materials and Methods

### *Drosophila* Lines and Libraries

A set of six highly inbred (20 generations) isofemale lines collected in Raleigh, North Carolina (RAL-301, RAL-303, RAL-306, RAL-358, RAL-375, RAL-732) and three extracted chromosome lines from Malawi (western Africa; MW28-5, MW56-4, MW63-5) were used in this study. The Malawi lines have wild origin chromosomes 2 and 3, whereas the X and fourth chromosomes may include balancers. Adult genomic DNAs were extracted as follows (Bingham et al. 1981): each line was expanded, and nuclei were isolated from adult males and females (in roughly equal proportions). The nuclei were resuspended in CsCl and ultracentrifuged to buoyant density equilibrium. The viscous fractions were dialyzed against TE. CsCl purified and resulting samples were sent to the Washington University Genome Sequencing center by Dr Charles Langley (University of California, Davis).

### Roche/454 GS-20 Sequencing

Genomic DNA was fragmented by nebulization according to standard Roche/454 protocols. Nebulized DNA was analyzed by agarose gel electrophoresis, and fragments within a size range of 500 bp were collected. The collected fragments were linker ligated with a mixture of the two 454-specific linkers, one species of which is biotinylated. An enrichment step was performed to remove fragments with the same species of unbiotinylated linker at both ends, by capturing those with biotinylated linkers on streptavidin magnetic beads. Next, the fragments on the beads were denatured, and the nonbiotinylated strand was reclaimed from the supernatant. First, the released single-stranded DNA fragments were run on an Agilent Bioanalyzer to calculate yield, then coupled to sepharose beads that carry covalently linked oligonucleotides comple-

mentary to the linkers ligated onto the nebulized DNA fragments. The input concentration of DNA fragments was adjusted to give, on average, a 1:1 association between beads and DNA fragments. The mixture was then emulsified in an oil suspension containing aqueous polymerase chain reaction (PCR) reactants, and emulsion PCR (emPCR) enabled the amplification of millions of unique fragment-bead combinations in a large batch PCR format. After combining the emPCR reactions for the library, sepharose beads that contained amplified DNA were isolated via streptavidin magnetic beads in order to capture the biotinylated ends of amplified fragments. Following enrichment, the biotinylated strand was melted away by the addition of NaOH, and sequencing primers were annealed to the bead-bound amplicons.

Primer- and polymerase-bound sepharose beads were loaded into a PicoTiterPlate (PTP) device, composed of hundreds of thousands of fused fiber optic strands, the ends of which are hollowed out to a diameter sufficient to contain a single sepharose bead. Smaller magnetic beads, to which pyrosequencing (sulfurylase and luciferase) enzymes are covalently attached, were pipetted into the PTP subsequently, and a centrifugation step packed them around each sepharose bead. The PTP fits into a flow-cell device that positions it against a high-sensitivity CCD camera in the 454 GS-20 sequencing instrument. Pyrosequencing follows, whereby sequential flows of each deoxyribonucleotide triphosphate, separated by an imaging step and a wash step take place. At each well address in the PTP, the incorporation of one or more nucleotides into the synthesized strand on each bead was captured by the CCD camera, which records positional information about each well address that reports a signal during the initial flow cycles and then monitors all addresses throughout the sequencing process. Separate runs were performed for each of the nine lines. All sequences are available from the National Center for Biotechnology Information's Short Read Archive under the submission accession, SRA009784 (study accession, SRP001156); basic statistics for each strain are presented as supplementary table 1 (Supplementary Material online).

### Base-Calling and Synthetic Assembly to *D. melanogaster* Reference

Because the native quality scores produced by the Roche/454 GS-20 platform have been shown to underestimate the accuracy of each sequenced nucleotide, we recalled all sequencing reads with the Pyrobayes base-calling algorithm prior to alignment, which empirically estimates error rates from 454 reads of the sequenced strain (iso-1), as previously described (Quinlan et al. 2008). Empirical error rates are observed to be 0.29% for insertions, 0.09% for deletions, and 0.017% for substitutions (which is more than one order of magnitude smaller than our estimates of  $\theta$ ) and are assumed to be homogenous across runs. We subsequently aligned each 454 read to the *D. melanogaster* Release 5 euchromatic genome (flybase.org) using the Mosaik alignment algorithm (<http://bioinformatics.bc.edu/marthlab/Mosaik>). Mosaik uses a hash-based approach for a fast initial read placement, followed by

an exhaustive local Smith–Waterman–Gotoh pairwise alignment (Smith and Waterman 1981; Gotoh 1982). We employed a relaxed gap opening penalty when aligning portions of 454 reads containing homopolymers. This minimizes spurious SNP calls owing to misalignment in homopolymer regions where the 454 technology is most prone to nucleotide over- or undercalls. We allowed each aligned read to differ from the reference sequence by up to 5% of the read length (e.g., up to five mismatches, insertions, and/or deletions for a 100 bp read). Mosaik examines all possible mapping locations for each read. In an effort to reduce false positive SNP calls that might arise from paralogous alignments, we only retained reads that aligned to a single locus within our 5% divergence threshold. In other words, we excluded reads where two alignments existed with less than 5% divergence relative to the reference genome. Predictably, increasing the tolerance for mismatches increases the number of SNPs we observe (supplementary table 2, Supplementary Material online); All aligned reads from all nine inbred lines were then multiply aligned and converted into ACE format so that polymorphic loci could be identified.

#### Identifying SNPs—Probabilistic PyroBayes

SNPs were called among the aligned sequences from all nine inbred lines using the GigaBayes polymorphism discovery algorithm (Quinlan A, Marth G, in preparation). Although the Bayesian SNP calling framework in GigaBayes is fundamentally similar to the PolyBayes algorithm (Marth et al. 1999), GigaBayes has been rewritten in C++ and is designed to efficiently process large data sets produced by next-generation sequencing platforms. GigaBayes assigned a posterior SNP likelihood to all reference genome positions where mismatches in the aligned reads were observed. The posterior SNP likelihood is derived from the PyroBayes base quality estimates assigned to the aligned bases at each putatively variant site. Owing to fourth and X chromosomes from balancer stocks harbored in maintained African lines, two independent data sets were generated: 1) all strains from both populations but without fourth and X chromosomes and 2) all chromosomes from the North Carolina population. GigaBayes makes a preliminary screen for alignment positions where there is at least one alternate allele aligned with a PyroBayes quality score greater than 5. For all putative polymorphic loci, GigaBayes first computes the posterior likelihood of the three most likely diploid genotypes for each aligned strain based on the observed alleles and quality scores for that strain. GigaBayes then computes the posterior SNP likelihood from the computed genotype likelihoods for each aligned strain. The derived posterior likelihood is based on an initial estimate of  $\theta$ , which for this study, was estimated at 1/200.

For estimates of  $\theta$ , we include only sites with reads from at least two different strains; this filtering is necessary to ensure that all polymorphisms we include are polymorphic within our sample and do not represent differences between our sample and the reference genome. Furthermore, we exclude any site that is inferred to be heterozygous in a single strain.

#### Estimating $\theta$ Using Partial Data

To estimate nucleotide diversity (measured as Watterson's  $\theta$ ) from the aligned short-read data, we used a modification of the approach described in Hellmann et al. (2008). We need to correct both for variation in coverage across the alignment and for sequencing errors. For each 50-kb window, we calculate  $\theta$  per site as follows: for each alignment segment of length  $L$  with a constant coverage, the number of segregating sites were estimated as the sum of the posterior probability of all putative SNPs, effectively weighting each observed SNP by the probability it is a true positive. We then use the standard Watterson's estimator to calculate  $\theta$  for each segment. To calculate  $\theta$  for a 50-kb window, then, we sum across the segments of length  $L$  with constant depth, each weighted by length.

To test the efficacy of this method for correcting for coverage variation, we simulated data using ms under the standard neutral coalescent and then generated simulated short-read sequencing data sets using custom Perl scripts. For each simulated data set, subsampled sites were generated to an expected coverage of 0.1x, 0.25x, 0.5x, 1x, 2x, or 4x per line, using the observed mean and variance of read lengths in our data set. We then calculated  $\theta$  for each simulated data set both before and after the sparse sampling.

#### Identification of TE Sequences

To estimate the overall TE content by class and family in 454 sequencing reads from nine strains of *D. melanogaster*, we concatenated files ending in \*TCA.454R-reads.fna into a single fasta file for each strain. A custom RepeatMasker library was constructed by modifying the Berkeley Drosophila Genome Project's TE data set (v9.4.1) to 1) include the class/subclass of each family and 2) remove any TE sequences not from *D. melanogaster*. TE sequences were detected in 454 reads using RepeatMasker open-3.1.6 (parameters: -s -xsmall -gff -no\_is) with WU-BlastN (v 2.0) as the search engine. Summary statistics on overall TE abundance per strain were obtained from RepeatMasker .out files. We note that overall estimates of TE content by class and family are based on all chromosomes, and thus African strains were excluded from this analysis because of the nonisogenicity of the fourth and X chromosome in these strains.

To identify individual TE insertions in the nine strains of *D. melanogaster*, we filtered for individual 454 reads that span both TE sequences and non-TE sequences in the genome. Reads had nonzero length after removing all TE nucleotides were used in a BLAT v32x1 search against the *D. melanogaster* Release 5 genome sequence. We refer to these sequences as TE "flank tags." Initial results demonstrated that many flank tags overlapped annotated TEs, which represent unmasked TE sequence in the 454 reads because of sequencing error and/or divergence from the TE consensus sequence. We therefore removed all flank tags that overlapped annotated TEs on genomic coordinates. We also observed that many flank tags hit multiple locations in the genome, which may result from segmental duplications, reference genome misassemblies, or incomplete definition of the consensus TE query sequence.

Because we cannot unambiguously determine the location of these flank tags, we removed all flank tags that hit more than one genomic location.

The final set of “unique flank tags” (UFTs) was used to assess the presence or absence of known and novel TE insertions in all nine strains. This analysis only included regions of the Release 5 genome sequence with TEs annotated and curated in (Quesneville et al. 2005) and omits uncurated repeat regions in the heterochromatic extensions to Release 5 not present in Release 4. To verify the presence of known TE insertions, UFTs were overlapped against the 100 bp upstream and downstream of each TE in *D. melanogaster* annotated and curated in (Quesneville et al. 2005). Reads that had a UFT mapping within 100 bp of an annotated TE for which RepeatMasker annotated the same TE family in the TE portion of the read were identified as supporting the presence of the annotated TE in that strain.

### Identifying CNPs

Filtering the reads for structural mutation analysis comprised several steps. First, we performed individual BLAT searches of 30 bp flanks from 5' and 3' ends of each 454 read against the *D. melanogaster* Release 5 genome. We then identified reads containing flanks that each possess a single unique hit in the genome and where there exists a difference in length between their position on the actual read versus their alignment on the genome. Second, we applied a homopolymer filter in which reads were removed if the total homopolymer length was greater than twice the difference of read and mapped lengths. Third, duplicated reads were eliminated, and only those reads whose BLAT products consisted of two highly significant hits ( $E$  value  $< 1 \times 10^{-6}$ ) and one or no hits covering 75% of the read were retained. This provided us with a conservative list of reads that potentially map to structural polymorphism. Some real events were likely excluded by our filter, including mutations in repetitive regions. Reads were subsequently divided into different classes based on mapping position and orientation of read ends. Classes include those in which read ends: 1) map to different chromosome arms, 2) are reverse complement (i.e., + and - strand) but map to the same arm, 3) are in wrong order (e.g., the 5' end of the read maps 3' to the 3' end of the read) but are on the same strand and chromosome arm, and 4) map to the same chromosome arm, strand, and are in the correct order. We confirmed putative deletions and duplications detected by Emerson et al. (2008) by filtering reads to those with ends that map to CNP ends (within 500 bp).

### Genomic Annotations and Statistical Analysis

Gene annotations (e.g., coding sequence [CDS], introns, intergenic regions, etc.) were extracted from FlyBase Release 5.4 (flybase.org). Divergence between *D. melanogaster* and *D. simulans* were parsed from chain files generated by the University of California Santa Cruz Genome Browser (genome.ucsc.edu). Tables are available upon request from R.J.K. Statistical analyses were carried out in R version 2.8. R scripts are available upon request from T.B.S.

## Results

### Developing Methods to Infer Patterns of SNP Variation with Short-Read Data

The initial challenges to a population genomic analysis of short-read data are read mapping, local assembly, and SNP identification. For the initial mapping, we used the software package Mosaik, which uses a BLAT-like approach to align short sequencing reads to a reference genome (Quinlan A, Marth G, in preparation for further details, see Materials and Methods). Reads from all runs (in both populations) were aligned together. Overall, 66.7% of the *D. melanogaster* assembled reference genome is covered by at least one sequencing read, and 33.4% of the reference is covered by at least two aligned sequencing reads. Alignment coverage varies considerably across the genome (fig. 1A) and is highly correlated across the two populations (fig. 1B). Regions with very low alignment coverage in both populations likely represent either repetitive regions with little sequence that can be uniquely mapped or regions that are difficult to sequence with 454 technology. In each population, the fraction of bases that are not uniquely aligned fit very closely to expectations from the Lander–Waterman model (Lander and Waterman 1988): for the North Carolina population (mean alignment coverage = 0.893), 40.3% of bases in the assembled reference genome are not uniquely mapped, compared with an expectation of 40.9%, and for the Malawi population (mean alignment coverage = 0.274), 76.2% of bases in the assembled reference genome are not uniquely mapped, compared with the expectation of 76.0% (fig. 1C).

We used the software package GigaBayes (Quinlan A and Marth G, in preparation) to call SNPs. Briefly, GigaBayes uses a Bayesian approach to estimate a posterior probability that an observed segregating site represents a true SNP, based on the prior probability of observing a SNP and on the 454 sequencing error model (for more details, see Materials and Methods). This method then allows us to propagate the uncertainty in SNP calls through to subsequent population genetic parameter estimates and hypothesis tests. In general, most detected SNPs are called with high confidence (fig. 2).

Traditional tools of population genetic inference are typically not robust to substantial missing data, such as arises from sparse alignments with low-coverage next-generation sequencing projects. Recent work has begun to develop statistical frameworks for sparse coverage population genomics, focusing in particular on estimating nucleotide diversity ( $\theta$ ) when coverage is low and variable (Hellmann et al. 2008; Jiang et al. 2009; Lynch 2009). In sequence data from heterozygous organisms, the challenge is not only to correct for variation in coverage in the sample but also to estimate accurately the number of alleles sampled at each position given that multiple reads from a single individual can represent one or two alleles. The *D. melanogaster* data presented here, sequenced from highly inbred strains, are slightly simpler, as we can assume that each strain carries only one allele at each position in the genome (an assumption violated to the extent that residual heterozygosity may exist in our population, but generally considered reasonable for highly inbred *Drosophila*

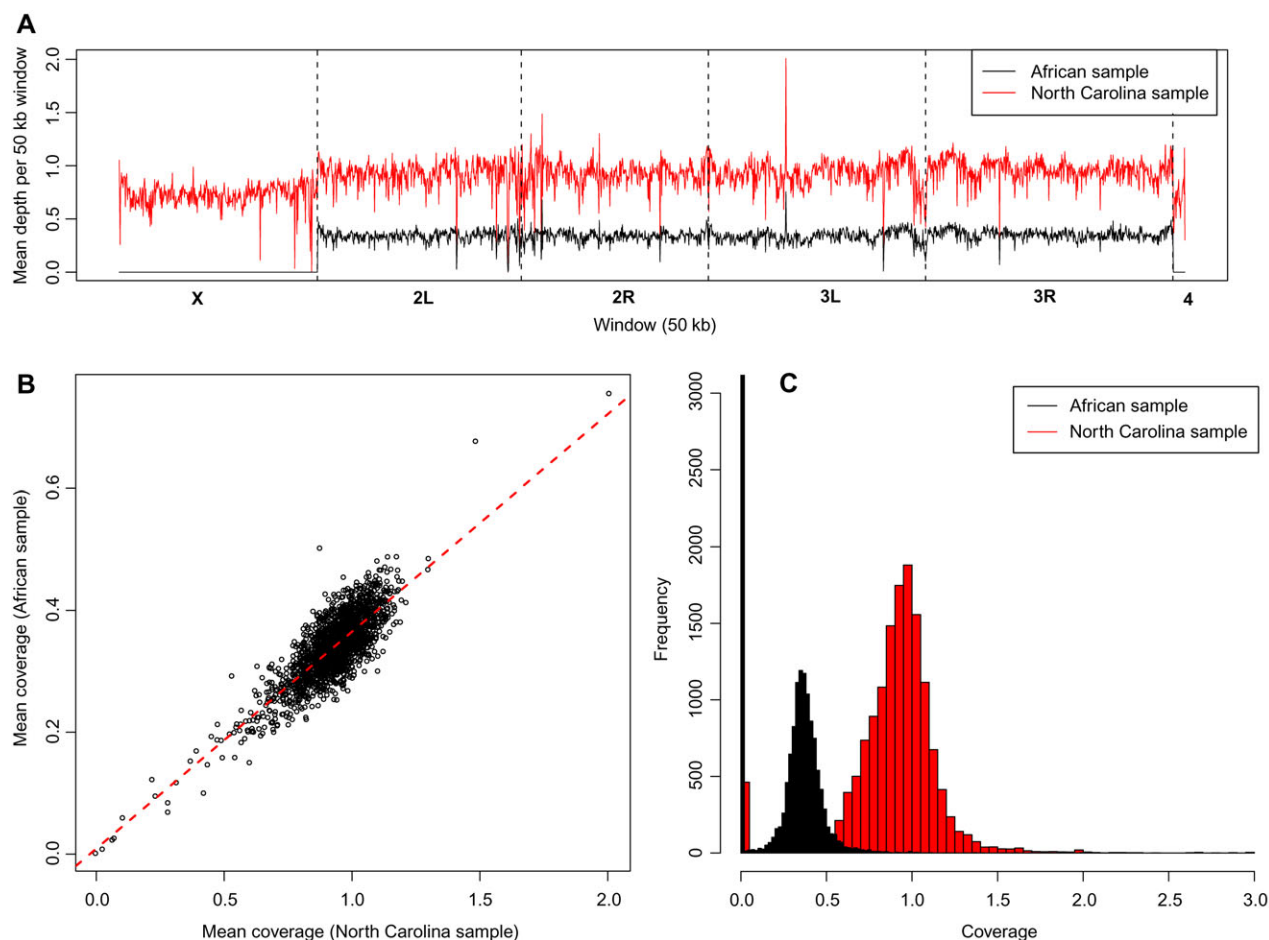


FIG. 1.—Genomic patterns of coverage. (A) Mean coverage in 50-kb windows across the genome. Chromosome arms are separated by dashed lines and labeled. X and fourth chromosomes are only shown for North Carolina populations. (B) Correlation in mapped coverage between African (Malawi) and North American (North Carolina) samples, measured by mean coverage in orthologous 50-kb windows. (C) Frequency distribution of mean coverage in African and North American samples across 50-kb windows.

strains). In practice for the application of GigaBayes, we begin with a strong prior that each site within each line of flies is homozygous.

In order to estimate  $\theta$ , we modified the approach initially described by Hellmann et al. (2008). Specifically, we consider a given alignment to consist of discrete segments of constant sample size, with observed segregating sites estimated based on the Bayesian posterior probability of a SNP at each position calculated by GigaBayes as described above. For a given segment with  $i$  segregating sites and with a constant alignment depth, we calculate  $\theta$  using the standard Watterson equation (Watterson 1975). The number of segregating sites in a segment of  $n$  sites is estimated as  $\sum_{i=1}^n \Pr(S_i)$ , where  $\Pr(S_i)$  represents the posterior probability that the  $i$ th site is segregating. We then sum over segments weighted by the length of each segment.

In order to verify the behavior of our estimator under a range of coverage conditions, we simulated data under a variety of different  $\theta$  values, coverage depths, and recombination rates based on the empirical properties of our 454 reads (mean and variance of length) and assuming that reads are randomly distributed. Across a range of simulated coverage depths, our estimator is unbiased, although the variance of the estimator increases dramatically at lower

coverage (fig. 3). Furthermore, coverage is uncorrelated with  $\theta$  in our simulations, suggesting that our method adequately corrects for coverage variation across the genome.

We also wanted to test the possible influence of misspecification of the posterior probabilities of segregating sites on our results. Although previous work has suggested that the GigaBayes error model is accurate based on calibration to known sequence data (Quinlan et al. 2008), we considered the effect of two “worse-case” extremes on the magnitude of our  $\theta$  estimates: a conservative worse-case scenario and a permissive worse-case scenario. In the conservative scenario, we consider the effect of assuming that all SNPs with a posterior probability of less than 99% are false positives. On average, this results in a  $\theta$  estimate reduced by 51.9% relative to the standard case: median overall  $\theta$  for the combined autosomal sample is reduced from 0.0049 to 0.0024. Restricting the sample to synonymous sites or noncoding sites produces a very similar pattern (data not shown). Restricting the sample to nonsynonymous sites results in a slightly larger reduction in median  $\theta$  (by 62.1% relative to the standard case), apparently due to a slightly lower percentage of observed segregating sites with high posterior probability in the nonsynonymous sample. For the permissive case, we assume that all

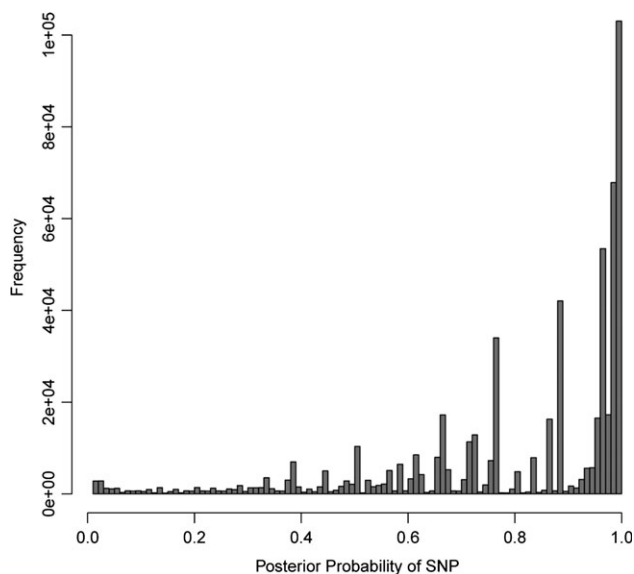


FIG. 2.—Distribution of SNP probabilities. Histogram of posterior probabilities for SNPs, based on estimates from the combined (North Carolina + Malawi) alignments. Results are similar for North Carolina or Malawi alone.

observed segregating sites, regardless of their posterior probability, represent true SNPs. On average, this results in a combined overall  $\theta$  estimate increased by 32.7% relative to the standard case (median overall  $\theta$  increases to 0.0065 from 0.0049). Restricting the sample to nonsynonymous sites or to synonymous sites results in roughly similar increases in  $\theta$  (by 40.4% and 29.8%, respectively). It is likely that assuming a uniform prior across site classes, as GigaBayes does, is at least partly responsible for the variation in the degree to which different assumptions about sequencing error appear to affect different site classes.

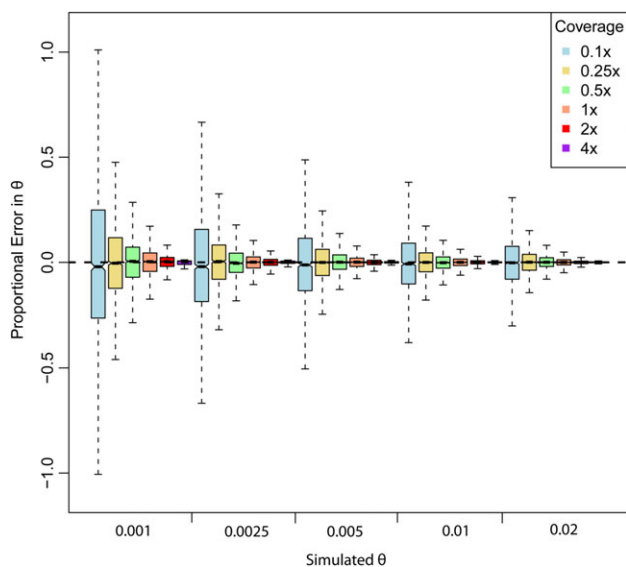


FIG. 3.—Deviation from true  $\theta$  in simulations. Data were simulated using *ms* and custom scripts under several different true  $\theta$ s and with varying degrees of coverage. Boxplot shows the proportional error in  $\theta$  (difference between  $\theta$  assuming perfect coverage and  $\theta$  after sampling, divided by real  $\theta$ ) under several coverages for each simulated  $\theta$ .

**Table 1**  
Estimates of  $\theta$  for Each Site Class (All Sites, 4-Fold Degenerate Sites, Noncoding Sites, Nonsynonymous Sites, and Synonymous Sites) from Each Population of *Drosophila melanogaster* and Chromosome Combination

Site class	North Carolina			Malawi
	Autosomes	X chromosome	Fourth chromosome	
All	0.004121	0.002910	0.002342	0.004773
4-fold degenerate	0.008717	0.005717	0.001366	0.011151
Noncoding	0.004393	0.003064	0.002640	0.005037
Nonsynonymous	0.001247	0.000787	0.000391	0.001609
Synonymous	0.010328	0.006853	0.001723	0.013331

NOTE.—Major autosomes are pooled. Due to the presence of segregating balancer chromosomes, only autosomes were sampled from the Malawi population.

However, overall, it appears that assuming badly misspecified sequencing error models can result in  $\theta$  estimates at most 50% higher or lower than what we observe.

These results suggest that future work should focus on the development of models that allow for different priors for different site classes and that allow computation of full data likelihoods instead of just presence/absence of segregating sites in order to continue to improve inference from low-coverage data. Nonetheless, provided the error model is reasonably good, our simulations show that reliable estimates of  $\theta$  are possible from even extremely sparse data.

#### Genomic Landscape of Population Genetic Variation

We calculated the average nucleotide diversity ( $\theta$ ) for autosomes in both populations, North Carolina X chromosomes and North Carolina fourth chromosomes in 50-kb windows for all sites as well as four subsets of site classes: noncoding (intergenic and intronic), synonymous, 4-fold degenerate, and nonsynonymous (table 1). Our estimates of  $\theta$  range from a low of 0.000391 at nonsynonymous sites on the North Carolina fourth chromosome to a high of 0.013331 at synonymous sites on Malawi autosomes. Average nucleotide diversity is lowest at nonsynonymous sites across all populations and chromosomes, as expected. With the exception of the North Carolina fourth chromosome, average nucleotide diversity is substantially higher at synonymous and 4-fold degenerate sites than noncoding sites. Although the data we present here are unweighted means of 50-kb windows, diversity estimates from 20 kb and 100 kb windows are highly consistent, as are results when we weight diversity by the fraction of bases in each window sequenced to a depth of at least two (data not shown). For consistency with previous studies, and in order to focus primarily on the classes of sites where variation is least likely to be affected by selection, we primarily focus on synonymous diversity (and intergenic diversity, although that measure is less likely to represent neutral variation) in the following sections.

#### African Versus North American Diversity

Previous studies of DNA sequence variation in *D. melanogaster* have shown reduced variation in derived

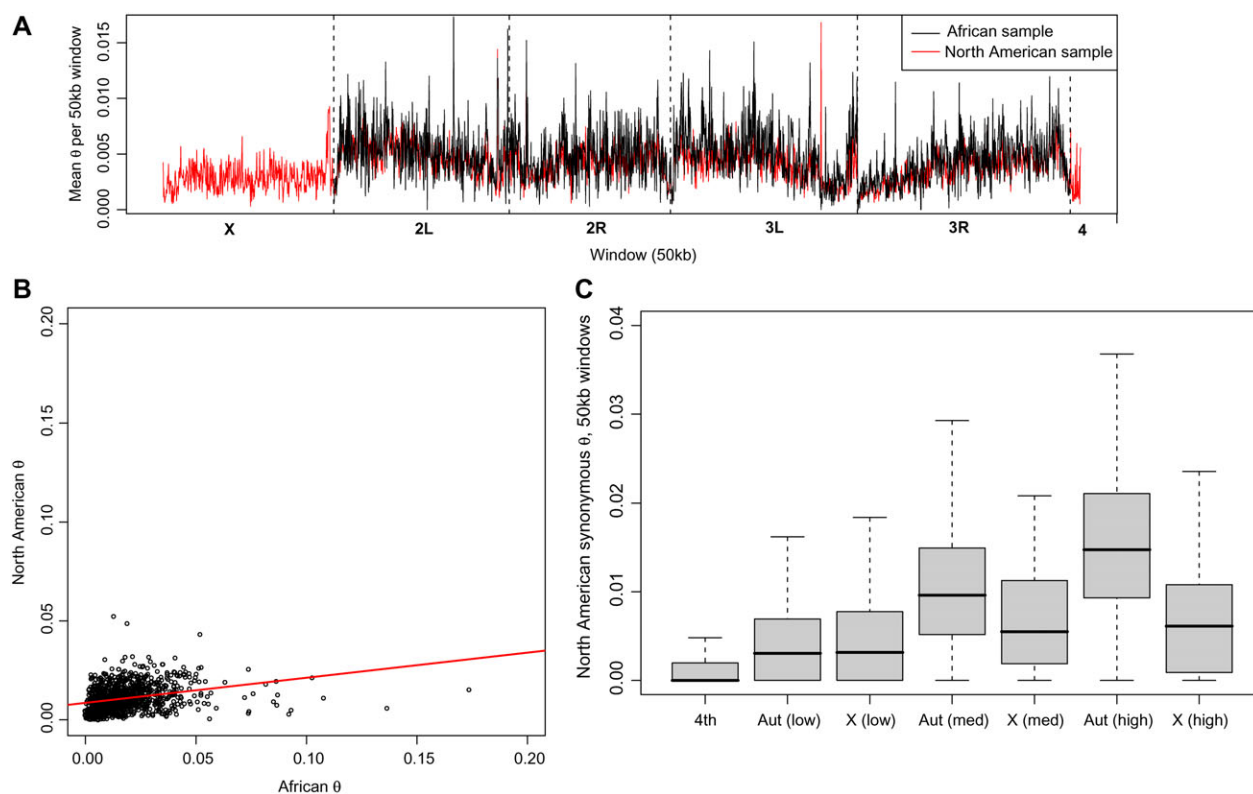


FIG. 4.—Genomic patterns of diversity. (A) Estimates of  $\theta$  (for all sites in 50-kb windows) across chromosome arms. Chromosome arms are separated by dashed lines and labeled. (B) Correlation between African (Malawi) and North American (North Carolina)  $\theta$ , measured for all autosomal sites in orthologous 50-kb windows. (C) Boxplot of North Carolina synonymous  $\theta$  for different chromosomes and recombination rates. Low recombination, bottom quartile of recombination rate estimates; medium recombination, middle two quartiles; high recombination, top quartile.

non-African populations relative to ancestral populations in Africa, although the magnitude of the reduction in variation varies considerably among studies (Begun and Aquadro 1993; Glinka et al. 2003; Haddrill et al. 2005; Hutter et al. 2007; Singh et al. 2007). Genome-wide nucleotide variation (among 50-kb windows with at least 200 bp of useable sequence from each population; fig. 4A) is significantly reduced in the North Carolina population sampled here relative to the Malawi population at both synonymous sites (median  $\theta_{\text{NC}}/\theta_{\text{AF}} = 0.793$ , Wilcoxon signed-rank test,  $P = 9.832 \times 10^{-8}$ ) and noncoding sites (median  $\theta_{\text{NC}}/\theta_{\text{AF}} = 0.892$ , Wilcoxon signed-rank test,  $P < 2.2 \times 10^{-16}$ ).

Across orthologous 50-kb windows, both synonymous  $\theta$  and noncoding  $\theta$  are highly positively correlated between Malawi and North Carolina populations (synonymous:  $\rho = 0.617$ ,  $P < 2.2 \times 10^{-16}$ ; noncoding:  $\rho = 0.384$ ,  $P < 2.2 \times 10^{-16}$ ; fig. 4B). Numerous variables which may influence patterns of nucleotide diversity across the genome are correlated between the two populations, including gene density and divergence (which are essentially identical), as well as coverage which is highly correlated ( $\rho = 0.740$ ,  $P < 2.2 \times 10^{-16}$ ; fig. 1B) and recombination rate, which is expected to be very highly correlated except to the extent that polymorphic inversions affect recombination rates. The correlation between Malawi and North Carolina in regional levels of diversity across the genome could be explained by an underlying correlation with any of these factors that are correlated across populations.

#### X Chromosome Versus Autosome Diversity

Using the North American lines, we can compare diversity on the autosomes with the X chromosome. Synonymous X chromosome diversity is significantly reduced compared with autosomal diversity (X:A ratio = 0.663; Mann–Whitney  $U$ ,  $P < 1.78 \times 10^{-15}$ ; the effect is much stronger in regions of high recombination than low recombination, fig. 4C). If sex ratios are equal, X chromosome diversity is expected to be three-fourths that of autosomal diversity due to differences in effective population sizes. To test if the reduction we observe in X chromosome diversity deviates from this expectation, we normalized X chromosome diversity by multiplying by 4/3; this normalized X chromosome diversity is still significantly lower than autosomal diversity (normalized X:A ratio = 0.884, Mann–Whitney  $U$ ,  $P = 6.17 \times 10^{-4}$ ). We find very similar patterns when comparing noncoding diversity between X chromosomes and autosomes (data not shown).

Both selective and demographic effects have been proposed to explain reduced diversity on the X chromosome (Begun and Whitley 2000; Wall et al. 2002; Singh et al. 2007; Pool and Nielsen 2008). If the average new positively selected mutation is at least partially recessive, hitchhiking is more efficient on the X chromosome, which can lead to reduced X-linked diversity at linked neutral sites under many conditions (Charlesworth et al. 1987) but not all (Orr and Betancourt 2001; Betancourt et al. 2004). In principle, male-skewed sex ratios could explain this effect, as

**Table 2**  
**Previous Estimates of  $\theta$  for African and Non-African Populations of *Drosophila melanogaster***

Study	Population	Chromosome	$\theta$
Singh et al. (2007)	Af	Auto	0.01392
Hutter et al. (2007)	Af	Auto	0.01140
Andolfatto (2001)	Af	Auto	0.00599
Singh et al. (2007)	non-Af	X	0.00473
Hutter et al. (2007)	non-Af	X	0.00470
Andolfatto (2001)	non-Af	X	0.00333
Singh et al. (2007)	non-Af	Auto	0.01325
Hutter et al. (2007)	non-Af	Auto	0.00686
Andolfatto (2001)	non-Af	Auto	0.00541

NOTE.—Data sets are Hutter et al. (2007), Singh et al. (2007), and Andolfatto (2001). Singh et al. (2007) and Hutter et al. (2007) report on noncoding sites primarily; Andolfatto (2001) reports on synonymous sites primarily. Abbreviations: Af, African; non-Af, non-African; Auto, autosomes; X, X chromosomes.

they will reduce the effective population size of the X chromosome relative to autosomes, and thus reduce the expected level of segregating neutral polymorphism (Caballero 1995; Charlesworth 2001). However, it is unlikely that sex-ratio skew can completely account for the reduction in X-linked diversity we observe: to produce the X:A diversity ratio we observe, the population would need twice as many males as females, which is unlikely in the absence of a sex-ratio distorter. Alternatively, Pool and Nielsen (2007, 2008) have shown that population size changes can produce X:A diversity ratios consistent with what is observed in *Drosophila* populations. Finally, sex-biased mutation rates could easily produce the X:A diversity ratios we observe, although no good evidence currently exists for sex-biased mutation rates in *Drosophila* (Bauer and Aquadro 1997).

In our data, the sequencing coverage on the X chromosome would be expected to be only be three-fourths of autosomal coverage because the whole-genome shotgun sampling was from a mixed sex (50:50) genomic extraction. Empirically, our data fit this expectation closely (actual X:A coverage in the North America lines 0.778; fig. 1A). It is unlikely however that the reduced polymorphism on the X chromosome is due to lower sampling coverage because, as discussed in the previous section, our simulation results suggest that our  $\theta$  estimator is unbiased with respect to coverage. Nevertheless, it may be advisable to only extract DNA from adult females in future studies to circumvent this issue. In short, our sparse-data results appear consistent with previous studies with respect to X:autosome polymorphism levels.

#### *Diversity on the Fourth Chromosome*

The fourth chromosome in *D. melanogaster* recombines at very low levels and has been previously observed to be lacking in nucleotide polymorphism (Berry et al. 1991; Wang et al. 2002, 2004; Sheldahl et al. 2003). We find that the overall level of nucleotide polymorphism on the fourth chromosome is only 56.8% of that on the major autosomes (Mann–Whitney  $U$ ,  $P = 1.135 \times 10^{-7}$ ) and 80.5% of that on the X chromosome (Mann–Whitney  $U$ ,  $P = 0.001366$ ). Levels of synonymous diversity are reduced even more severely to 16.7% and 24.4% of auto-

somal or X chromosome levels, respectively (autosomes: Mann–Whitney  $U$ ,  $P = 1.773 \times 10^{-9}$ ; X: Mann–Whitney  $U$ ,  $P = 1.141 \times 10^{-5}$ ). As discussed below, it is likely that the reduction of diversity in regions of low recombination explains some substantial portion of the reduced variability on the fourth chromosome. We compared levels of diversity in very low recombination regions of autosomes and the X chromosome to diversity on the fourth chromosome. In both cases, synonymous diversity on the fourth chromosome is reduced to about 33% of that observed in low recombination regions of the autosomes or X chromosome; Although still a significant reduction (Mann–Whitney  $U$ ,  $P = 0.0003$  for autosomes and 0.002 for X chromosomes), it is less reduced than when compared with high recombination regions of the autosomes. These results support previous studies that suggest the presence of recurrent selective sweeps (Maynard-Smith and Haigh 1974; Kaplan et al. 1989) and/or background selection (Charlesworth et al. 1993, 1995; Charlesworth 1996) on this small, essentially nonrecombining chromosome.

#### *Comparisons with Previous Studies*

Numerous studies in the past decade have investigated patterns of population genetic variation in *D. melanogaster*, providing an ideal opportunity to investigate the correspondence between genome-wide sparse short-read analysis and traditional sampling and Sanger sequencing approaches (Andolfatto 2001; Haddrill et al. 2005; Hutter et al. 2007; Singh et al. 2007). We summarize estimates of  $\theta$  from previous comparable studies in table 2. In general, we observe lower estimates of nucleotide diversity in our sample than in previous studies, which we explore in more detail below.

Many early studies of genetic variation in *D. melanogaster* were limited to a small number of loci by the cost of generating extensive sequence data sets; however, a large recent multilocus survey of genetic variation in one African and one European population provides an ideal data set for comparisons to our genome-wide data (Hutter et al. 2007). In order to compare our estimates of  $\theta$  to those from a traditional sequencing approach, we matched loci from Hutter et al. (2007) to windows in our study. For each locus sequenced in Hutter et al. (2007), we identify the window that contains the studied locus and compare our estimates of  $\theta$  from African and North American populations to Hutter et al. (2007) estimates of  $\theta$  from African and European populations. Although the populations sampled are not the same (Malawi vs. Zimbabwe for African populations; North Carolina vs. The Netherlands for non-African populations), we would still expect broadly similar patterns of nucleotide diversity across the two studies.

We do find that our estimates of  $\theta$  for North Carolina and African populations are significantly correlated with estimates from Hutter et al. (2007) (North Carolina vs. Netherlands:  $\rho = 0.133$ ,  $P = 6.089 \times 10^{-16}$ ; Malawi vs. Zimbabwe:  $\rho = 0.154$ ,  $P = 0.0086$ ). However, the median difference within a window between our estimate and the previous estimate (Hutter et al. 2007) is significantly greater than zero for both the non-African and African samples, with higher estimates of  $\theta$  in Hutter et al.

**Table 3**  
**Type III Sums of Squares and *P* Values from Analysis of Variance for Three Different Models**

	Sum of squares	df	<i>F</i> value	Pr (> <i>F</i> )
Autosomes, North Carolina				
Intercept	0.000595	1	7.83	0.005
Gene density	0.000006	1	0.08	0.77
Recombination rate	0.026796	1	352.59	$<2.2 \times 10^{-16}$
Divergence	0.002352	1	30.95	$3.03 \times 10^{-8}$
Residuals	0.140747	1852		
Autosomes, Malawi				
Intercept	0.00236	1	8.39	0.0038
Gene density	0.00020	1	0.73	0.39
Recombination rate	0.02629	1	93.6	$<2.2 \times 10^{-16}$
Divergence	0.00218	1	7.74	0.0054
Residuals	0.48623	1730		
X chromosome, North Carolina				
Intercept	0.0013436	1	28.42	$1.606 \times 10^{-7}$
Gene density	0.0000168	1	0.35	0.55
Recombination rate	0.0002458	1	5.20	0.0231
Divergence	0.0000368	1	0.778	0.38
Residuals	0.0196677	416		

NOTE.—In each case, the predictor variables are recombination rate (based on [http://petrov.stanford.edu/cgi-bin/recombination-rates\\_updateR5.pl](http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl)), divergence to *D. simulans* (based on UCSC genome-to-genome alignments), and gene density. The response variable is synonymous  $\theta$  for North Carolina autosomes, Malawi autosomes, and North Carolina X chromosomes, respectively. In all cases, similar results are obtained when using alternative measures of recombination rate (data not shown). df, degree of freedom.

(2007) (median non-African difference: 0.0015, Wilcoxon signed-rank test,  $P < 2.2 \times 10^{-16}$ ; median African difference: 0.0058, Wilcoxon signed-rank test,  $P < 2.2 \times 10^{-16}$ ). These data are consistent with several possibilities, including unaccounted for sequencing errors in the (Hutter et al. 2007) data set inflating estimates of  $\theta$ , a too conservative correction for sequencing errors in our data set reducing estimates of  $\theta$ , or differences between the populations sampled in our study and by Hutter et al. (2007).

### Genomic Correlates of Diversity across Windows

With genome-wide data, we have a unique opportunity to examine genomic correlates of variation in diversity within populations (fig. 4A). We focus on three particular parameters: divergence to *D. simulans*, recombination rate, and gene density. Polymorphism and divergence are predicted to be correlated to the extent that mutation rates vary across the genome; it is important to include divergence in our model in order to correct for variation in mutation rate. Under models of hitchhiking (Maynard-Smith and Haigh 1974; Kaplan et al. 1989) or background selection (Charlesworth et al. 1993, 1995; Charlesworth 1996), recombination rate and polymorphism are expected to be negatively correlated, with reduced polymorphism in regions of low recombination. To the extent that higher gene density correlates with more targets for selection, we also expect windows with more coding sequence to have lower levels of linked neutral polymorphism.

To test these hypotheses, we fit a linear model with either autosomal North Carolina  $\theta$ , X chromosome North Carolina  $\theta$  or autosomal African  $\theta$  calculated based on synonymous sites as the independent variable and recombination rate (measured using one of four different approaches),

divergence to *D. simulans*, and gene density (as the fraction of CDS in a window) as the dependent variables. In all cases, we find that recombination rate is a significantly positively correlated with  $\theta$ , divergence is significantly positively correlated with autosomal  $\theta$  but not X chromosomal  $\theta$ , and gene density is not significantly correlated with  $\theta$  (table 3). This is consistent with previous results in *D. melanogaster* (Begun and Aquadro 1992; Andolfatto 2007; Haddrill et al. 2007) and other species (Kulathinal et al. 2008), although recent genomic studies in *D. simulans* suggest a negative correlation between coding density and polymorphism (Begun et al. 2007) which we do not find here. The fact that we fail to observe a significant correlation with gene density is consistent with recent observations that suggest a substantial fraction of noncoding sequence in *Drosophila* may be under selection (Bergman and Kreitman 2001; Andolfatto 2005; Halligan and Keightley 2006), which would imply that coding density is not a reliable proxy for density of selective targets in a region.

Interestingly, the extent to which recombination rate is a good predictor of nucleotide diversity varies considerably among North Carolina autosomes, Malawi autosomes, and North Carolina X chromosomes. On the North Carolina autosomes, recombination rate explains between 5.5% and 15.7% of variation in synonymous polymorphism, depending on the measure of recombination rate used. For African autosomes, recombination rate only explains between 1.9% and 5.0% of variation in synonymous polymorphism, and for North American X chromosomes, recombination rate only explains between 1.0% and 1.7% of the variation in synonymous polymorphism.

One possible explanation for this pattern is that recombination rates are different in North American and African populations due to inversions segregating in African populations. Although it is difficult to directly test whether recombination rates are consistent between populations, we can test whether or not regions of high inconsistency in diversity between African and North American populations are correlated with recombination rate: A negative correlation between  $\theta_{AF} - \theta_{NC}$  and recombination rate would suggest that regions of high recombination in North American populations may have low recombination rates in Africa and vice versa. However, we do not observe a significant correlation between  $\theta_{AF} - \theta_{NC}$  and recombination rate ( $\rho = 0.0211$ ,  $P = 0.3797$ ). We believe a more likely explanation for this pattern is that recombination rates on the *D. melanogaster* X chromosome have recently evolved (Takano-Shimizu 1999). Previous work has shown that correlations between codon bias and recombination rate have shifted on the *D. melanogaster* X chromosome compared with the autosomes (Singh, Davis, and Petrov 2005), which is also consistent with a shift in X chromosome recombination patterns in recent evolutionary time.

### TE Distributions

#### Unbiased Genome-Wide Estimates of TE Abundance

Estimates of genome-wide TE content in *D. melanogaster* vary widely among different reports. Manning et al. (1975) estimated from reassociation kinetics that 12% of the *D. melanogaster* genome is middle repetitive

**Table 4**  
**Abundance of Natural TE Sequences in North Carolina Lines of *Drosophila melanogaster* Estimated from Roche/454 Light-Shotgun Data**

Strain	Total read length (bp)	Total TE (bp)	Total TE (%)	LTR TE (bp)	LINE TE (bp)	DNA TE (bp)	LTR TE (%)	LINE TE (%)	DNA TE (%)
RAL-301	46018941	5978658	12.99	3285515	2037690	655453	7.14	4.43	1.42
RAL-303	35035499	4566780	13.03	2396756	1678637	491387	6.84	4.79	1.40
RAL-306	32822554	3799168	11.57	2055061	1330296	413811	6.26	4.05	1.26
RAL-358	37903202	4507858	11.89	2366965	1626406	514487	6.24	4.29	1.36
RAL-375	34779820	3850776	11.07	1998471	1449718	402587	5.75	4.17	1.16
RAL-732	29001706	3734127	12.88	1906083	1416674	411370	6.57	4.88	1.42

NOTE.—The total amount and percent of TE sequences, as well as subtotals for each of the major subclasses of TE (LTR, LINE-like, and DNA elements) are shown relative to the total amount of 454 sequences processed.

DNA, whereas Young (1979) calculated that 16–17% of the genome is middle repetitive in nature. Spradling and Rubin (1981) estimated that three-fourths of middle repetitive DNA, or a total of 9–12% of the genome, would be comprised of families of dispersed repeats like TEs. Estimates of genome-wide TE content based on analysis of sequenced euchromatic and heterochromatic sections of the *D. melanogaster* genome have ranged from 7.5% of the genome (Bartolome et al. 2002) to 20–22% (Kapitonov and Jurka 2003; Quesneville et al. 2005) to 28% (Biemont and Vieira 2005; Smith, Edger, et al. 2007). The potentially unbiased nature of sampling regions from across the genome using 454 sequencing offers a new source of data to address the question of genome-wide TE content in *D. melanogaster*.

Genome-wide TE content for the six North Carolina strains of *D. melanogaster* estimated from 454 light-shotgun reads is shown in table 4. Because this analysis does not place reads to genomic locations, we excluded African strains because of the nonisogenicity of their X and fourth chromosomes. The proportion of sequences matching known TE families ranges from 11.1% in RAL-375 to 13.0% in RAL-303, with a median value of 12.4% across all six strains. In general, these results indicate that genome-wide TE content based on 454 sequencing is more consistent with the results of reassociation kinetic studies than previous analyses based on traditional clone-based Sanger-sequence genome assemblies. Moreover, the overall TE content from 454 sequencing (~12%) for all strains is clearly higher than that estimated for the euchromatic component of the reference genome sequence (~3.3%; Bergman et al. 2006), indicating that a substantial proportion of 454 reads are from TE-rich heterochromatic regions. Given that 1/3 of the *D. melanogaster* genome is thought to be heterochromatic (Adams et al. 2000), we estimate that ~30% of all heterochromatic regions are TE sequence. This figure is lower than estimated TE content in heterochromatin based on direct analysis of currently cloned and sequenced regions (50–70%; Hoskins et al. 2002; Smith, Shu, et al. 2007). This discrepancy in results based on 454 and Sanger sequencing may imply that currently uncloned, unsequenced portions of the heterochromatin may have substantially lower TE content than cloned and sequenced heterochromatic regions. This inference is consistent with observations of rare “islands” of complex TE sequences in “seas” of simple repeats in centromeric regions of the X chromosome (Sun et al. 2003).

The rank order of abundance for the major classes/subclasses of TEs is the same for all six North American strains (table 4). Long terminal repeat (LTR) retrotransposons (~6.4%) are the most abundant, followed by LINE-like retrotransposons (~4.4%) and DNA transposons (~1.4%), as has been observed in the *D. melanogaster* reference genome sequence and other genome sequences in the genus *Drosophila* (Kaminker et al. 2002; Bergman et al. 2006; Clark et al. 2007). Additionally, sequences for all but 4 of the 125 known *D. melanogaster* TE families are found in each of the North American strains, including families that are absent from the reference genome such as the *P element*. Only the *BS3*, *Helena*, *Helitron*, and *Stalker3* families are not observed in all North American strains, which may simply reflect an effect of sparse shotgun coverage because these families are detected in low abundance in at least two North American strains. At the individual family level, the LINE-like retrotransposon *R1 element* occupies the largest fraction of the genome followed by the LTR retrotransposon *roo*, which is the most abundant family in the euchromatic reference genome sequence (Kaminker et al. 2002).

#### *Presence in Natural Strains of TE Insertions Annotated in Reference Genome*

We also investigated which of the individual known TE insertions annotated in the *D. melanogaster* genome sequence were present in any of the nine wild-type strains from both populations, excluding TEs on the X and fourth chromosome. To do this, we identified 454 reads containing both TE and unique sequence that mapped to the flanking regions of the 3,961 TEs on chromosomes 2 and 3 in regions of the Release 5 reference assembly that were annotated by Quesneville et al. (2005). We required these UFTs to map to only one position in the genome. A summary of TEs on chromosomes 2 and 3 in the reference genome with supporting UFTs in these nine wild-type strains can be found in supplementary file 1 (Supplementary Material online), coordinates of UFTs overlapping these TE flanking regions can be found as supplementary file 2 (Supplementary Material online), and an illustration of UFTs that overlap known TEs on chromosome arm 3L is shown in figure 5A.

In total, each light-shotgun sequence produced UFTs that overlapped 90–203 TEs (minimum in MW28-5/MW56-4 and a maximum of in RAL-301) on the major autosomes, corresponding to 2.3–5.1% of known TEs on

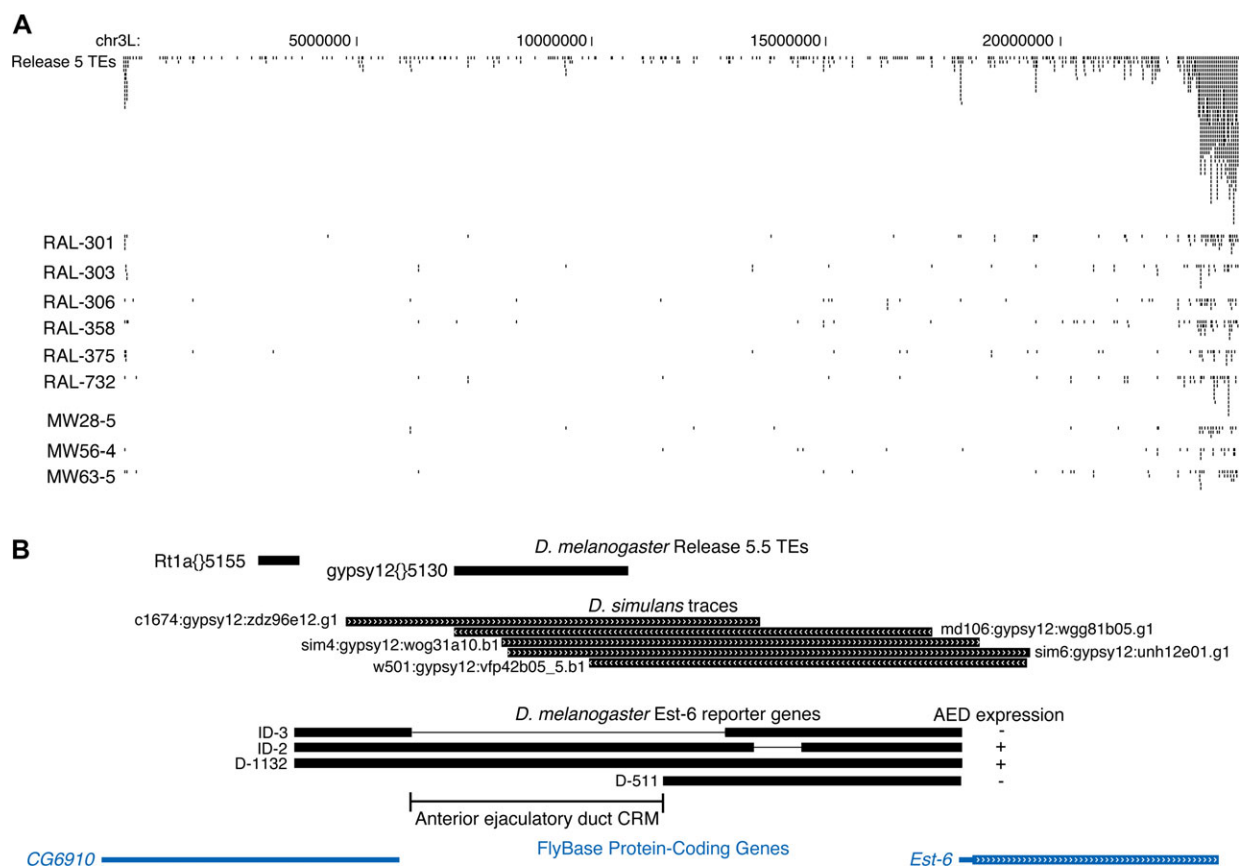


FIG. 5.—TE distribution across lines. (A) Overview of unique flank tags (UFTs) from *Drosophila melanogaster* Roche/454 sequences that overlap known TEs on chromosome 3L. (B) An example of a fixed ancestral TE, gypsy12{5130 (FBti0063191) that co-localizes with a *cis*-regulatory module (CRM) from the *Esterase-6* (*Est-6*) gene that drives reporter gene expression in the male anterior ejaculatory duct (AED). Reporter gene constructs are from Ludvig et al. (1993); Tamarina et al. (1997)

chromosomes 2 and 3 in the reference genome. We find that the presence of 22.3% (885/3961) of TEs on chromosomes 2 and 3 annotated in the euchromatic genome sequence is supported by at least one UFT in one or more wild-type strain. The majority of TEs are supported by UFTs in only one strain (71.6%, 634/885). Previous estimates based on population sampling of a limited number of loci suggest that >40% of annotated TEs are present in natural strains (Lipatov et al. 2005), and thus this sample of unique 454 TE flank tags likely underestimates the proportion of annotated TE insertions segregating in nature. This underestimate likely arises from the incomplete coverage of these light-shotgun data, omission of the TE-rich fourth chromosome, and the requirement for UFTs to map uniquely in the genome, which prevent UFTs from mapping to TE dense regions that are enriched in segmental duplications (Bergman et al. 2006; Fiston-Lavier et al. 2007).

The majority of known TE insertions on chromosomes 2 and 3 found in natural strains are from the *INE-1* family of elements (55.6%, 492/885), for which 32.2% (492/1529) of *INE-1* elements in the reference genome on chromosomes 2 and 3 are found in one of the nine wild-type strains. This result is consistent with the facts that *INE-1* is the most numerous TE annotated in *D. melanogaster* genome sequence (Quesneville et al. 2005) and that insertions for this family are thought to have fixed in the genome prior

to speciation from the common ancestor with *D. simulans* (Singh and Petrov 2004; Singh, Arndt, and Petrov 2005; Wang et al. 2007; see below). If we assume that all *INE-1* elements in *D. melanogaster* are fixed, then the proportion of *INE-1* insertions discovered in natural strains should represent the proportion of true positive TEs detected using the UFT method in this sample of 454 reads. Using this correction factor applied to the number of TEs observed in natural strains yields 2748 (885/0.322) TEs on chromosomes 2 and 3 that would be discovered in nature given full genome coverage of these strains, which would convert to 69.4% (2748/3961) of all TEs annotated on chromosomes 2 and 3 in the reference genome sequence. This number slightly exceeds the total number of TEs that are annotated in low recombination regions on chromosomes 2 and 3 ( $n = 2673$ ). Given the fact that the majority of annotated TEs found in these strains are in low recombination regions (73.8%, 653/885), the data are compatible with a scenario in which essentially all TEs in *D. melanogaster* low recombination regions are fixed or segregating at high frequencies in natural populations (Bartolome and Maside 2004), with only  $\sim 6\%$  [(2748 - 2673)/(3961 - 2673)] of insertions in the reference genome at appreciable frequency in high recombination regions.

Our data also indicate that the likelihood of a TE insertion in the reference genome to be found segregating

at appreciable frequency in nature depends on its transposition mechanism, with DNA-based transposons found with higher probability than RNA-based retrotransposons. Approximately 31.3% of DNA-based elements on chromosomes 2 and 3 are found in natural strains (659/2106), whereas only 15.7% of LINE (126/801) or 9.5% of LTR (100/1054) elements are present on chromosomes 2 and 3 of natural strains. This effect is not due solely to the highly abundant *INE-1* family, as 28.9% (167/577) of non-*INE-1* DNA elements on chromosomes 2 and 3 are also present in natural strains. Thus, the abundance of a TE class across the *D. melanogaster* reference genome is inversely related to the likelihood of its presence in natural strains. With complete genome coverage, we estimate ~90% of DNA elements would be shown to be segregating at appreciable frequencies in nature, in contrast to ~65% of annotated LINE or ~30% of LTR elements. The higher population frequency of DNA and non-LTR elements relative to LTR elements may be related to the fact that they are in general shorter (Kaminker et al. 2002) and that shorter TE insertions are predicted to be less deleterious under models that posit TE evolution is controlled by genome compaction (Fontanillas et al. 2007) or ectopic exchange (Petrov et al. 2003). Alternatively, this pattern may reflect the differences in the historical activity of TE classes associated with worldwide colonization (Bergman and Bensasson 2007) perhaps due to recent horizontal transfer (Bartolome et al. 2009).

#### Identification of Putatively Ancestral TE Insertions

Finally, we applied the same UFT strategy to the Sanger shotgun traces from seven strains of *D. simulans* generated by the Drosophila Population Genomics Project (Begun et al. 2007), which revealed a set of TE insertions in the *D. melanogaster* genome that are likely to have inserted in the common ancestor of these species. As a positive control that cross-species UFT mapping is effective, we can recover three TEs that have been reported previously in the literature as ancestral (FBti0064176 [Bartolome and Maside 2004]; FBti0019203, FBti0020079 [Bergman and Bensasson 2007]). In total, 19.3% (1039/5385) of all Release 5 TEs overlapped with UFTs from *D. simulans*. The overwhelming majority of ancestral insertions are from the *INE-1* family (86.4%, 898/1039), which is thought to have undergone transposition prior to the divergence of these two species (Singh and Petrov 2004; Singh, Arndt, and Petrov 2005; Wang et al. 2007). Here, we provide direct evidence of ancestral insertion for 40.2% (898/2235) of all annotated *INE-1* elements. The remainder of the putatively ancestral non-*INE-1* TE insertions falls roughly evenly among the three major TE classes (58 DNA elements, 49 LINE elements, and 34 LTR elements) with nearly equal numbers in high (61) and low (80) recombination regions.

We investigated in detail the set of four putatively ancestral TE insertions that are present in the maximal number of strains observed (5 of 7) in high recombination regions and identified 2 potential cases of molecular domestication: 1) a *TART-A* element (FBti0061962) that spans the terminal two exons of the nuclear RNA export factor 2 (*nx2*) gene, and 2) a *gypsy12* element (FBti0063191) that co-localizes

precisely with the *Est-6* anterior ejaculatory duct *cis*-regulatory module (Ludwig et al. 1993; Tamarina et al. 1997; fig. 5B). Intriguingly, although present in *D. melanogaster* and *D. simulans* genomes, this *gypsy12* element appears to be absent from the *Drosophila yakuba* genome (data not shown), which correlates perfectly with the presence or absence of *Est-6* expression in the male ejaculatory duct in these species (Richmond et al. 1990). Because *Drosophila pseudoobscura*, an outgroup to these three species, also lacks the ejaculatory duct regulatory element and expression pattern (Tamarina et al. 1997), we propose that the insertion of *gypsy12* into the *Est-6* promoter region (either directly or indirectly) led to the gain of *Est-6* ejaculatory duct expression in the ancestor leading to the *melanogaster-simulans* lineage after divergence from the rest of the *melanogaster* species subgroup.

#### CNPs in the *D. melanogaster* Genome

It is becoming clear that differences in the number of copies of large-scale DNA segments represent a substantial source of genetic variation across diverse species. Indeed, in humans and in model genetic organisms, including *D. melanogaster*, CNP can account for variation in ~10% of the genome and ~8% of genes (Redon et al. 2006; Dopman and Hartl 2007; Stranger et al. 2007; Emerson et al. 2008; Henrichsen et al. 2009). Genome-wide experimental surveys of CNPs have been possible by using array comparative genomic hybridization (aCGH) (Pollack et al. 1999), with probes of various size (from 25-mer oligos to bacterial artificial chromosomes) and genome density (in CDS or tiling array). Although ideal for initial characterization, at least two limitations of aCGH exist. First, array platform differences can produce biased, nonoverlapping data sets (Redon et al. 2006; Henrichsen et al. 2009). Second, resolving each of the thousands of CNP breakpoints that are discovered by aCGH requires extensive experimental follow-up work (e.g., through genome walking and sequencing of CNP ends). As an unbiased alternative to aCGH, several recent studies have demonstrated the power of high-throughput sequencing to resolve known CNPs at the level of individual base pairs and to discover novel CNPs and other types of structural variation (Korbel et al. 2007; Campbell et al. 2008; Daines et al. 2009).

To detect CNPs with our 454 sequencing data, we developed a pipeline to identify reads that span breakpoints of CNP events, depicted in figure 6A. We initially performed a BLAT search of the 5' and 3' ends (30 bp) of each read against the *D. melanogaster* genome to identify reads where each end of the read has a single unique hit in the genome and where the genomic distance between hits on the reference assembly differs from the length of the read. We then screened these initial candidates with two additional filters: First, we removed any read where the total length of homopolymer sequence was greater than twice the inferred insertion or deletion size. This filter conservatively removes size differences that could be explained by errors in determining the length of homopolymer runs, which can be a substantial problem in 454 sequencing data. Second, we BLASTed the full sequence of the remaining candidates against the

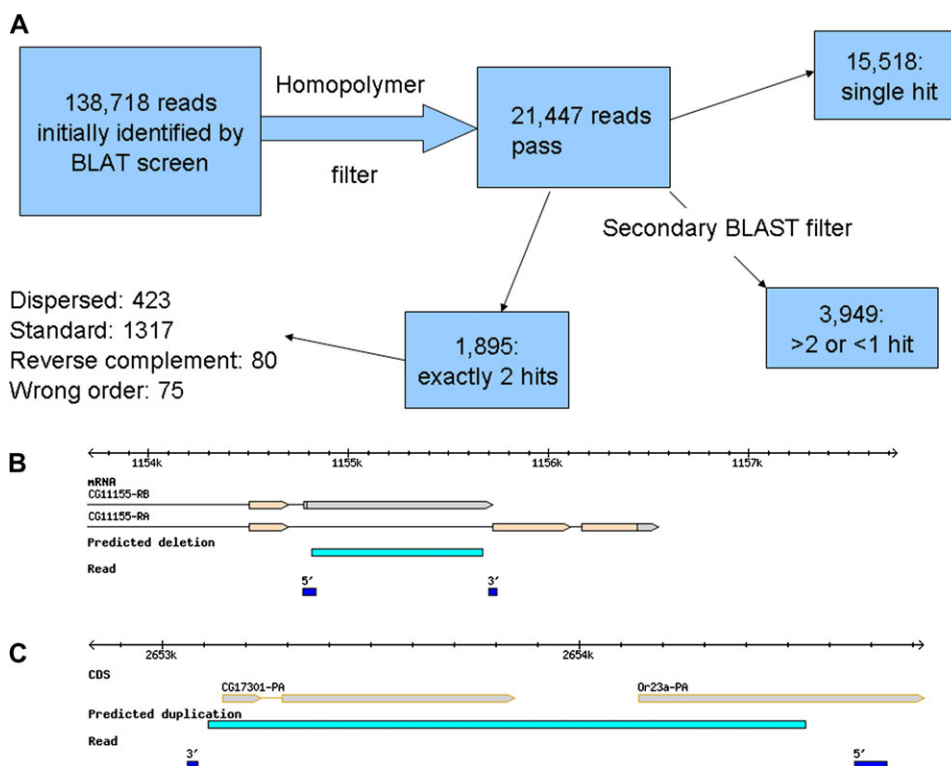


FIG. 6.—Diversity in CNP. (A) Flowchart of CNP detection pipeline. (B) An example of an observed deletion in the sampled line. The sequenced read is shown in dark blue, mapped to the genome; the predicted deletion is shown in light blue. (C) An example of an observed duplication. The sequenced read is shown in dark blue (note that the 3' end of the read maps 5' of the 5' end, as would be predicted for a tandem duplication), with the predicted region that is duplicated shown in light blue.

*D. melanogaster* genome and removed reads where the Blast result and the initial BLAT results of the ends are inconsistent. Finally, we then screened the final set to remove duplicate reads. This workflow produced a conservative estimate of the extent of CNP diversity among 454 sequences. Some events cannot be detected by this pipeline, such as insertions of repetitive sequence in the sequenced read relative to the reference, which would lead to one end of the read having nonunique Blast hits. Nonetheless, this approach can detect a number of structural mutations, including deletions of any size, small insertions, translocations, inversions, and tandem duplications (supplementary fig. 1, Supplementary Material online).

To facilitate discrimination among different structural mutation types, the 1,898 reads that passed our filtering criteria were divided into four classes based on mapping position and read-end orientation (fig. 6A). Classes include those in which read ends: 1) map to different chromosome arms (“dispersed”), 2) are oriented on plus and minus strands but map to the same arm (“reverse complement”), 3) are in the incorrect order (e.g., the 5' end of the read maps 3' to the 3' end of the read) but are on the same strand and chromosome arm (“wrong order”), and 4) map to the same chromosome arm, strand, and are in the correct order (“standard”). These different configurations are likely to represent different mutational events: translocation events will produce signatures consistent with the dispersed pattern, duplications and deletions with the standard pattern, tandem duplications with the wrong order pattern,

and inversions with the reverse complement pattern (supplementary fig. 1, Supplementary Material online).

Of reads that fit into a standard conformation, half (947) are consistent with a deletion (relative to the reference genomic sequence). The median size of these deletion events is 32 bp; most deletion events we detect are small, with only 15.2% of all deletions larger than 1 kb and only 4.1% larger than 10 kb. However, we do observe a few very large events (deletions in excess of 1 Mb), which are likely explained by the presence of dispersed duplicates on the same chromosome arm, chimeric sequences produced during adapter ligation (a problem for all singleton reads), or an inadequacy in our filtering or pipeline. A smaller number of small-sized duplications/insertions (370) occur in the data set and are much smaller in length (mean = 20.05 bp), suggesting that they represent a more homogeneous group of polymorphisms. A total of 423 reads have dispersed ends that map to different chromosome arms: these reads could represent translocations or chimeric sequences produced during adapter ligation. We tested for a skew in the distribution of these putative translocations across the assembled *D. melanogaster* genome. If translocations are randomly dispersed, the probability of movement between chromosome arms will depend on their relative sizes. Of 361 dispersed reads located in assembled regions, 359 have at least one end in a euchromatic region and 37 have at least one end in a heterochromatic region. This is consistent with random expectations, considering that ~93% of the *D. melanogaster* genome sequence is euchromatin

(goodness-of-fit  $G$ -test;  $P > 0.2$ ). We also find that the total number of translocations associated with each chromosome arm is consistent with their relative size (goodness-of-fit  $G$ -test;  $P > 0.3$ ). Furthermore, for a given arm, translocations are dispersed randomly among the other arms (goodness-of-fit  $G$ -test;  $P > 0.1$ ). Overall, these observations are consistent with an absence of bias in gene segment “traffic,” in contrast to the pattern previously reported for newly retroposed genes (Betran et al. 2002); this suggests the possibility that chimeric sequences produced during adapter ligation may explain most or all of these putative translocation events.

CNPs previously characterized by aCGH in *D. melanogaster* (Emerson et al. 2008) were also uncovered by 454 sequencing, in spite of low sequence coverage ( $\sim 0.2\times$  per line) and the use of different fly strains. Although few of the 1,000's of CNPs uncovered by aCGH could be verified in our data set, shared polymorphisms suggest that these mutations are either weakly deleterious (or neutral) or are sites of recurrent mutation. Tandem duplications and deletions are among the easiest structural mutation class to recognize with short-read sequence data because of their distinctive sequence characteristics (supplementary fig. 1, Supplementary Material online). Five reads show patterns consistent with the presence of a tandem duplication break point and are near four aCGH characterized duplications (supplementary table 3, Supplementary Material online). Two tandem duplications occur in coding regions. One 1,677 bp duplication contains the entire coding sequence for *CG17301* and part of a neighboring gene, *Odorant Receptor 23a (Or23a)* (fig. 5B). The second coding region duplication is 4,580 bp long and comprises the 3' end of the alternative transcript, *Sap-r PA*. A total of 12 deletions in reads correspond to presumptive aCGH deletions. Of these, three are  $>50$  bp in length and include an 858 bp deletion in the 3' UTR of an alternative transcript for *CG11155 (CG11155-RB)*; fig. 6C), a 480 bp deletion in an intron for *Nipped-A*, and a 146 bp deletion in an intron for *dnc*. We found no evidence for inverted tandem duplicates that corresponded to aCGH duplications.

## Discussion

To sample genotypic space within species, empirical population genetics has closely followed the current state of the art molecular techniques for surveying genetic variation (Avice 1994). From the early days of protein electrophoresis (Lewontin and Hubby 1966), to DNA sequencing (Kreitman 1983), to surveys of microsatellite variation (Schlotterer et al. 1997; Irvin et al. 1998), to large-scale resequencing screens (Hutter et al. 2007), two major goals in population genetics have been to characterize patterns of genetic variation in natural populations and, subsequently, to infer processes of evolutionary change. Efforts to unravel evolutionary processes at a finer scale have motivated the development of tools to increase both the number of sampled individuals and the fraction of the genome covered. New sequencing technologies (Mardis 2008) are now instigating a step change in the scope of population genetics by generating sample coverage and depth at a much higher

scale than ever before. As the quantity of data available for population genetic analysis grows to multiple full genome proportions (Liti et al. 2009), new techniques and analytical approaches will be required. Several studies have already begun to develop methods for estimating nucleotide diversity from sparse data (Hellmann et al. 2008; Jiang et al. 2009; Lynch 2009), but none have yet applied these approaches to real short-read data sets. As data begin to appear from large-scale resequencing projects such as the 1000 Genomes Project in humans, the 1001 Genome Project in Arabidopsis, and the Drosophila Genetic Reference Panel Project in flies, understanding the practical application and limitations of these short-read methods will become increasingly important.

Here, we present the first attempt to make population genetic inference on a genomic scale from low-coverage alignments. Using two populations of *D. melanogaster*, sampled at different coverage levels, we employed stringent approaches and criteria, including conservative alignments, probabilistic SNP models, and a correction to estimate nucleotide diversity. In many cases, we recapitulate patterns of SNP variation previously observed in *Drosophila*: reduced diversity on the X chromosome relative to autosomes, reduced diversity in non-African populations relative to ancestral African populations, and positive correlations between recombination rate and diversity. We also report novel results that depend on broad-scale sampling, in particular our observation that correlations between recombination rate (based on the standard genetic map of *D. melanogaster*) and diversity appear to be stronger for non-African autosomes than other populations and chromosomes.

However, our approach also suffers from important limitations. Our estimates of  $\theta$  appear to be influenced by the conservative choices made during alignment and SNP calling: we tend to observe lower estimates of  $\theta$  than previously reported. In future studies, it will be important to recognize that alignment and SNP calling methods can have significant impacts on downstream estimates of diversity. Additionally, given current methods and the sparse nature of our data set, we cannot make inferences that depend on frequency-based statistics. Deeper coverage and methods that allow for the calculation of full data likelihoods (as opposed to just the probability of a site being a SNP or not, relative to the reference) will be necessary to fully capture allele frequency information in sparse data sets.

Sampling entire genomes from natural populations via an increasing number of new sequencing platforms is likely to become the norm in population genomics. As sequencing significantly decreases in cost, it may soon be feasible to generate full, high-coverage resequencing data for model organisms with relatively small genomes. However, sparse data sets such as the one we describe here will undoubtedly become the norm in nonmodel organisms and in organisms with large genomes. It is therefore imperative that we continue to develop rigorous statistical methods that deal with this onslaught of random genomic sequences. In this paper, we highlight the potential problems of sparse coverage population genomics, which include alignment issues, sequencing quality, variable depth of coverage, and missing sites. We show that solutions to these problems—a conservative

Mosaik assembly incorporating sequencing errors, Bayesian model for SNP identification, and unbiased estimators of nucleotide diversity (i.e.,  $\theta$ )—allow us to infer the expected patterns of variation from even very sparse coverage across two populations of *D. melanogaster*, although further work will be required to develop methods to allow inference based on allele frequencies and to address the challenges inherent in a probabilistic approach to alignment and data quality.

Even current methods demonstrate the ample promise of short-read population genomics, especially for organisms where resources for high-quality and deep-coverage resequencing projects are not available. Sparse-coverage population genomic projects will always face some limitations: de novo assembly of low-coverage data is not feasible, and thus any population genomic study of this sort will require a reference genome for mapping purposes. Although the reference genome need not necessarily be the same species as the surveyed populations, more distant reference genomes will reduce mapping efficiency. Mapping efficiency is also likely to be reduced in organisms with very large genomes and especially those with high repetitive DNA content, as repetitive sequences generally cannot be uniquely mapped to the reference. However, beyond the availability of a suitable reference, we believe that sparse-coverage short-read approaches provide a cost-effective and accessibility way to survey genome-wide variation in a wide range of organisms. The method for inferring  $\theta$  described here is easily applicable to heterozygous organisms (Hellmann et al. 2008), obviating the need for inbreeding prior to sequencing. Furthermore, genome-wide sampling has important advantages over alternative approaches, such as sequencing targeted genomic regions: as we demonstrate, a single experiment can provide information about SNPs, CNPs, and variation in TE content.

Population genetics has historically focused on mutational variants comprised of single nucleotide change. By utilizing random sequences aligned to a reference assembly, new genomic data hold the promise to provide a richer snapshot of extant genetic variation beyond single nucleotide variants. With genome-wide data amassed on a population scale, we can also characterize such patterns of genomic variation as TE diversity and CNP. By sampling structural and sequence variation in an integrated manner and by providing cost-effective ways for population-genomic inference in nonmodel organisms, next-generation sequencing is ushering us into a new era in population genomics that will allow comprehensive insights into the molecular variation underlying all genome and organismal evolution.

### Supplementary Material

Supplementary files 1 and 2, figure 1, and tables 1–3 are available at *Genome Biology and Evolution* online ([http://www.oxfordjournals.org/our\\_journals/gbe/](http://www.oxfordjournals.org/our_journals/gbe/)).

### Acknowledgments

We would like to thank the National Human Genome Research Institute for covering the costs of the initial se-

quencing as a pilot for assessing the utility of shallow short-read sequencing for genome-wide SNP discovery. Drs Charles F. Langley (University of California, Davis) and Trudy F.C. Mackay (North Carolina State University) suggested the initial set of lines for the project and Charles F. Langley provided the DNA samples for the study. Elaine Mardis and the Washington University Genome Sequencing Center performed the 454 sequencing. We also thank Chip Aquadro, John Wakeley, Dan Garrigan, and Sarah Kingan for insightful discussion. This work was supported by the National Institutes of Health (NRSA 1F32GM086950-01 to T.B.S., NRSA 1F32GM080090-01 to E.B.D., R01 AI064950 to A.G.C.] and the Natural Environment Research Council (NERC NE/G000158/1 to C.M.B.)

### Literature Cited

- Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science*. 287:2185–2195.
- Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol*. 18:279–290.
- Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*. 437:1149–1152.
- Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res*. 17:1755–1762.
- Avise J. 1994. Molecular markers, natural history, and evolution. New York: Chapman & Hall.
- Bartolome C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol*. 10:R22.
- Bartolome C, Maside X. 2004. The lack of recombination drives the fixation of transposable elements on the fourth chromosome of *Drosophila melanogaster*. *Genet Res*. 83:91–100.
- Bartolome C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol*. 19:926–937.
- Bauer VL, Aquadro CF. 1997. Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. *Mol Biol Evol*. 14:1252–1257.
- Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*. 356:519–520.
- Begun DJ, Aquadro CF. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature*. 365:548–550.
- Begun DJ, Aquadro CF. 1994. Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. *Genetics*. 136:155–171.
- Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.
- Begun DJ, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc Natl Acad Sci USA*. 97:5960–5965.
- Bergman CM, Bensasson D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 104:11340–11345.
- Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in

- intergenic and intronic sequences. *Genome Res.* 11:1335–1345.
- Bergman CM, Quesneville H, Anxolabehere D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* 7:R112.
- Berry AJ, Ajioka JW, Kreitman M. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* 129:1111–1117.
- Betancourt AJ, Kim Y, Orr HA. 2004. A pseudohitchhiking model of X vs. autosomal diversity. *Genetics* 168:2261–2269.
- Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12:1854–1859.
- Biemont C, Vieira C. 2005. What transposable elements tell us about genome organization and evolution: the case of *Drosophila*. *Cytogenet Genome Res.* 110:25–34.
- Bingham PM, Levis R, Rubin GM. 1981. Cloning of DNA sequences from the white locus of *D. melanogaster* by a novel and general method. *Cell* 25:693–704.
- Branscomb E, Predki P. 2002. On the high value of low standards. *J Bacteriol.* 184:6406–6409. discussion 6409.
- Caballero A. 1995. On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics.* 139:1007–1011.
- Campbell PJ, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet.* 40:722–729.
- Celniker SE, et al. 2002. Finishing a whole genome shotgun sequence assembly: release 3 of the *Drosophila* euchromatic genome sequence. *Genome Biol.* 3:RESEARCH0079.
- Charlesworth B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet Res.* 68:131–149.
- Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet Res.* 77:153–166.
- Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat.* 130:113–146.
- Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.
- Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141:1619–1632.
- Clark AG, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature.* 450:203–218.
- Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496–1502.
- Daines B, et al. 2009. High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. *Genetics.* 182:935–941.
- Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci USA.* 104:19920–19925.
- Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science.* 320:1629–1631.
- Fiston-Lavier AS, Anxolabehere D, Quesneville H. 2007. A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res.* 17:1458–1470.
- Fontanillas P, Hartl DL, Reuter M. 2007. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *PLoS Genet.* 3:e210.
- Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics.* 165:1269–1278.
- Gotoh O. 1982. An improved algorithm for matching biological sequences. *J Mol Biol.* 162:705–708.
- Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* 8:R18.
- Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15:790–799.
- Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16:875–884.
- Hellmann I, et al. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* 18:1020–1029.
- Henrichsen CN, et al. 2009. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet.* 41:424–429.
- Hoskins RA, et al. 2002. Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol.* 3:RESEARCH0085.
- Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics.* 177:469–480.
- Irvin SD, Wetterstrand KA, Hutter CM, Aquadro CF. 1998. Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*. Evidence for founder effects in new world populations. *Genetics.* 150:777–790.
- Jiang R, Tavare S, Marjoram P. 2009. Population genetic inference from resequencing data. *Genetics.* 181:187–197.
- Kaminker JS, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* 3:RESEARCH0084.
- Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci USA.* 100:6569–6574.
- Kaplan NL, Hudson RR, Langley CH. 1989. The “hitchhiking effect” revisited. *Genetics.* 123:887–899.
- Korbel JO, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science.* 318:420–426.
- Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature.* 304:412–417.
- Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MA. 2008. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc Natl Acad Sci USA.* 105:10051–10056.
- Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics.* 2:231–239.
- Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural population of *D. pseudoobscura*. *Genetics.* 54:595–609.
- Lipatov M, Lenkov K, Petrov DA, Bergman CM. 2005. Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. *BMC Biol.* 3:24.
- Liti G, et al. 2009. Population genomics of domestic and wild yeasts. *Nature.* 458:337–341.
- Ludwig MZ, Tamarina NA, Richmond RC. 1993. Localization of sequences controlling the spatial, temporal, and sex-specific

- expression of the esterase 6 locus in *Drosophila melanogaster* adults. *Proc Natl Acad Sci USA*. 90:6233–6237.
- Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*. 182:295–301.
- Manning JE, Schmid CW, Davidson N. 1975. Interspersion of repetitive and nonrepetitive DNA sequences in the *Drosophila melanogaster* genome. *Cell*. 4:141–155.
- Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 9:387–402.
- Marth GT, et al. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat Genet*. 23:452–456.
- Maynard-Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. *Genet Res*. 23:23–35.
- Misra S, et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol*. 3:RESEARCH0083.
- Mousset S, Derome N. 2004. Molecular polymorphism in *Drosophila melanogaster* and *D. simulans*: what have we learned from recent studies? *Genetica*. 120:79–86.
- Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol*. 22:2119–2130.
- Orengo DJ, Aguade M. 2004. Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. *Genetics*. 167:1759–1766.
- Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics*. 157:875–884.
- Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol*. 20:880–892.
- Pollack JR, et al. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*. 23:41–46.
- Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution*. 61:3001–3006.
- Pool JE, Nielsen R. 2008. The impact of founder events on chromosomal variability in multiply mating species. *Mol Biol Evol*. 25:1728–1736.
- Powell JR. 1997. *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford: Oxford University Press.
- Quesneville H, et al. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol*. 1:e22.
- Quinlan AR, Stewart DA, Stromberg MP, Marth GT. 2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods*. 5:179–181.
- Redon R, et al. 2006. Global variation in copy number in the human genome. *Nature*. 444:444–454.
- Richmond RC, Nielsen KM, Brady JP, Snella EM. 1990. Physiology, biochemistry and molecular biology of the *Est-6* locus in *Drosophila melanogaster*. In: Barker JSF, Starmer WT, McIntyre RJ, editors. *Ecological and evolutionary genetics of Drosophila*. New York: Plenum Press. p. 273–292.
- Schlotterer C, Vogl C, Tautz D. 1997. Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. *Genetics*. 146:309–320.
- Sheldahl LA, Weinreich DM, Rand DM. 2003. Recombination, dominance and selection on amino acid polymorphism in the *Drosophila* genome: contrasting patterns on the X and fourth chromosomes. *Genetics*. 165:1195–1208.
- Singh ND, Arndt PF, Petrov DA. 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics*. 169:709–722.
- Singh ND, Davis JC, Petrov DA. 2005. Codon bias and noncoding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. *J Mol Evol*. 61:315–324.
- Singh ND, Macpherson JM, Jensen JD, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. *BMC Evol Biol*. 7:202.
- Singh ND, Petrov DA. 2004. Rapid sequence turnover at an intergenic locus in *Drosophila*. *Mol Biol Evol*. 21:670–680.
- Smith CD, et al. 2007. Improved repeat identification and masking in Dipterans. *Gene*. 389:1–9.
- Smith CD, Shu S, Mungall CJ, Karpen GH. 2007. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science*. 316:1586–1591.
- Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol*. 147:195–197.
- Spradling AC, Rubin GM. 1981. *Drosophila* genome organization: conserved and dynamic aspects. *Annu Rev Genet*. 15:219–264.
- Stranger BE, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 315:848–853.
- Sun X, Le HD, Wahlstrom JM, Karpen GH. 2003. Sequence analysis of a functional *Drosophila* centromere. *Genome Res*. 13:182–194.
- Takano-Shimizu T. 1999. Local recombination and mutation effects on molecular evolution in *Drosophila*. *Genetics*. 153:1285–1296.
- Tamarina NA, Ludwig MZ, Richmond RC. 1997. Divergent and conserved features in the spatial expression of the *Drosophila pseudoobscura* esterase-5B gene and the esterase-6 gene of *Drosophila melanogaster*. *Proc Natl Acad Sci USA*. 94:7735–7741.
- Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics*. 162:203–216.
- Wang J, Keightley PD, Halligan DL. 2007. Effect of divergence time and recombination rate on molecular evolution of *Drosophila* INE-1 transposable elements and other candidates for neutrally evolving sites. *J Mol Evol*. 65:627–639.
- Wang W, Thornton K, Berry A, Long M. 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science*. 295:134–137.
- Wang W, Thornton K, Emerson JJ, Long M. 2004. Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. *Genetics*. 166:1783–1794.
- Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256–276.
- Wilson RJ, Goodman JL, Strelets VB. 2008. FlyBase: integration and improvements to query tools. *Nucleic Acids Res*. 36:D588–D593.
- Young MW. 1979. Middle repetitive DNA: a fluid component of the *Drosophila* genome. *Proc Natl Acad Sci USA*. 76:6274–6278.

Yoshihito Niimura, Associate Editor

Accepted November 14, 2009