The Drosophila Genome Nexus: a population genomic resource of 605 *Drosophila melanogaster* genomes, including 197 genomes from a single ancestral range population

Justin B. Lack[*,1], Charis M. Cardeno[†], Marc W. Crepeau[†], William Taylor[‡], Russell B. Corbett-Detig[§,**], Kristian A. Stevens[†], Charles H. Langley[†,1], John E. Pool[*,1]*

*Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, [†]Department of Evolution and Ecology, University of California, Davis, CA 95616, [‡]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, and [§]Department of Integrative Biology, and **Department of Statistics, University of California, Berkeley, CA 94720

Reference numbers for data available in public repositories...

Running title: *Drosophila* population genomics

Key words: *Drosophila melanogaster*, population genomics, genome assembly

[1]Corresponding authors:

Justin Lack:
5302 Genetics/Biotechnology Building, Laboratory of Genetics, University of Wisconsin-Madison, 425-G Henry Mall, Madison, WI 53706. E-mail: jlack@wisc.edu

Charles Langley:
3342B Storer Hall, Center for Population Biology and Department of Evolution and Ecology, University of California, One Shields Ave., Davis, CA 95616-8554. E-mail: chlangley@ucdavis.edu

John Pool:
5302 Genetics/Biotechnology Building, Laboratory of Genetics, University of Wisconsin-Madison, 425-G Henry Mall, Madison, WI 53706. E-mail: jpool@wisc.edu

Abstract

Hundreds of wild-derived *D. melanogaster* genomes have been published, but rigorous comparisons across data sets are precluded by differences in alignment methodology. The most common approach to reference-based genome assembly is a single round of alignment followed by quality filtering and variant detection. We evaluated variations and extensions of this approach, and settled on an assembly strategy that utilizes two alignment programs and incorporates both SNPs and short indels to construct an updated reference for a second round of mapping prior to final variant detection. Utilizing this approach, we reassembled published *D. melanogaster* population genomic data sets (previous DPGP releases and the DGRP freeze 2.0), and added unpublished genomes from several sub-Saharan populations. Most notably, we present aligned data from phase 3 of the Drosophila Population Genomics Project (DPGP3), which provides 197 genomes from a single ancestral range population of *D. melanogaster* (from Zambia). The large sample size, high genetic diversity, and potentially simpler demographic history of the DPGP3 sample will make this a highly valuable resource for fundamental population genetic research. The complete set of assemblies described here, termed the Drosophila Genome Nexus, presently comprises 605 consistently aligned genomes, and is publicly available in multiple formats with supporting documentation and bioinformatic tools. This resource will greatly facilitate population genomic analysis in this model species by reducing the methodological differences between data sets.

Introduction

Recent advances in next-generation sequencing have led whole genome sequencing to be extended from only a few genetic strains of select model organisms to many genomes from humans and from model and non-model taxa. While all fields of genetic analysis have benefited from these technological advances, population genetics has been especially affected as hundreds or even thousands of whole genome sequences have been generated for some organisms (e.g., The 1000 Genomes Project Consortium 2010, Pool *et al.* 2012; Huang *et al.* 2014; Long *et al.* 2013; Wallberg *et al.* 2014). As a result of these large population genomic databases, we have gained considerable power in detecting and understanding species history (e.g., Li and Durbin 2011), the genome-wide consequences of natural selection (*e.g.*, Comeron 2014), structural variation (*e.g.*, Corbett-Detig and Hartl 2012), and patterns of linkage disequilibrium and recombination (*e.g.*, Chan *et al.* 2012).

While these data sets have undeniable utility, the rapid development and deployment of next-generation technologies has been accompanied by a diversity of opinions on the most appropriate ways to assemble, filter, and curate extremely large, complex data elements. As a result, each population genomic data set is generated using unique combinations of library preparation chemistry and sequencing platform, different short-read aligning programs or pipelines with distinct biases and error rates, a wide range of quality filters and thresholds, and often distinct data formats. Ultimately, this renders population genomic data sets difficult to combine and jointly analyze. For example, it is difficult to understand whether population genetic statistics (*e.g.*, nucleotide diversity) are directly comparable given potential differences in error rate or mapping/coverage biases.

Drosophila melanogaster has played a pivotal role in essentially every field of genetic analysis, from population and evolutionary genetics to the development of fly models for understanding human disease. While *D. melanogaster* likely originated in Sub-Saharan Africa (Lachaise *et al.* 1988; Pool et al. 2012), natural populations now occur in essentially all temperate and tropical localities, and are typically commensal with humans. There are currently multiple independently generated population genomic data sets available that differ in the sequencing platform, assembly pipeline, and the data formats released to the public (Mackay *et al.* 2012; Huang *et al.* 2014; Langley *et al.* 2012; Pool *et al.* 2012). To address at least the last two of these issues, we present the assembly of 605 genomes from natural populations of *D. melanogaster*, all assembled using a common approach. This data set includes the previously published diploid DGRP freeze 2.0 genomes from Raleigh, NC (Huang *et al*. 2014), the DPGP collection of homozygous chromosomes from Malawi (Langley *et al*. 2012), and the haploid DPGP2 (Pool et al. 2012) collections of genomes from Sub-Saharan Africa. In addition, we publish 53 additional haploid embryo and inbred African genomes from Egypt, Ethiopia, Kenya, South Africa, and Uganda, along with the DPGP3 data set of 197 haploid embryo genomes from a single population in Zambia.

This deep genomic sequencing of the Zambia "ZI" sample was motivated by preliminary data suggesting that this population has maximal genetic diversity among known *D. melanogaster* populations, along with minimal levels of admixture from non-sub-Saharan ("cosmopolitan") populations (Pool *et al.* 2012). While the sample size of DPGP3 is comparable to that of DGRP, each data set has particular strengths. DGRP includes a substantial inbreeding effort to facilitate genotype-phenotype comparisons, whereas

DPGP3 uses a haploid method sequencing effort (Langley *et al.* 2011) to generate fully homozygous genomes for population genetic analysis. Because of its location within the sub-Saharan ancestral range of the species, the Zambia sample has not experienced the out-of-Africa bottleneck or New World admixture that are relevant to the DGRP population (Duchen *et al.* 2013). By providing a clear picture of diversity in the ancestral range of *D. melanogaster*, the DPGP3 collection will aid in understanding the histories of other worldwide populations and the species as a whole, as illustrated by studies of sub-Saharan human populations (CITATIONS). While Zambia may not necessarily represent a population at demographic equilibrium, the relative simplicity of its history will also facilitate studies of the effects of natural selection and other processes on genomic diversity.

The *Drosophila* Genome Nexus (DGN) created from these alignments is intended to facilitate population genetic analyses focused on single nucleotide polymorphisms (SNPs). We do not claim our assembly pipeline produces the best possible alignments. It does represent a modest advance over standard methodology, but its primary virtue is to increase the comparability of population genomic datasets. For example, there would be scientific value in comparing the North American DGRP population against separately published genomes from the European and African source populations from which it may derive (Caracristi and Schlötterer 2003; Duchen *et al.* 2013). Detailed population genetic inference is not a focus of the present study, but we present basic summaries of genetic diversity and structure, as well as patterns of admixture into sub-Saharan African populations from cosmopolitan populations.

Materials and Methods

Reassembly of previously published genomes

We obtained the raw sequencing reads from the NCBI short read archive (SRA; http://www.ncbi.nlm.nih.gov/sra) for Illumina data from the DGRP freeze 2.0 (Mackay *et al*. 2012; Huang *et al*. 2014), DPGP (Langley *et al*. 2012), and DPGP2 (Pool *et al*. 2012) collections of genomes (accession numbers given in Table S1). See the above citations for information concerning DNA extraction, library preparation, and sequencing, as these varied considerably among data sets.

Newly sequenced genomes

We present a considerable expansion of the genomic sequences available for *D. melanogaster* with the addition of 246 African lines. Table S1 provides descriptions of fly stocks and their availability, genome and alignment characteristics, and raw data accession numbers. Table S2 gives information about sampling locations. This expansion consists of 193 additional lines (primarily isofemale, some inbred for five generations) collected from Siavonga, Zambia (collectively referred to as the DPGP3 data set), in addition to the 4 Zambia ZI genomes previously published (Pool *et al.* 2012). We also include here 53 additional genomes (referred to here as the African Genomes Extended Sampling, or the AGES data set) from 12 African populations. Isofemale lines for these populations were collected following Pool (2009). For all DPGP3 genomes, genomic library preparation from haploid embryos followed the protocol of Langley et al. (2011). Sequencing for DPGP3 was performed on an Illumina Genome Analyzer IIx (Langley lab, UC Davis). From the AGES

182 data set, all but the Egyptian (EG) and Kenyan (KM) paired-end libraries were prepared
183 from haploid embryos using the same methods as for DPGP3. All AGES genomes were
184 sequenced at the UW-Madison Biotechnology Center on the Illumina HiSeq 2000 platform.
185 For the three EG genomes, inbred lines were established through full-sib mating for 8
186 generations, while the four KM genomes were sequenced directly from isofemale lines that
187 had been maintained in the laboratory for 12 years and therefore passively inbred. For
188 these 7 genomes, we extracted DNA from 30 females, paired-end library preparation and
189 size-selection for 300 bp inserts was conducted using the NEBNext DNA Library Prep
190 Reagent Set (New England BioLabs), and sequencing was conducted at the UW-Madison
191 Biotechnology Center on the Illumina HiSeq 2000 platform.
192
193 ***Genome assembly pipeline***
194 In reference assembly of short-read sequencing data, a major limitation is SNP and
195 indel divergence of sequenced genomes from the reference. This issue could result in
196 alignment biases if reads derived from low diversity regions of the genome can be aligned
197 more confidently, in addition to biases among genomes that vary in their overall
198 divergence from the reference sequence (*e.g.* sub-Saharan vs. cosmopolitan *D.*
199 *melanogaster*). In an attempt to ameliorate such effects, we developed and applied a
200 pipeline that combines two aligners with different degrees of sensitivity to non-reference
201 variation and speed, and utilizes two rounds of mapping (Fig. S1). In brief, we first mapped
202 short read data to the *D. melanogaster* reference genome (release 5.57; http://flybase.org)
203 using BWA v0.5.9 (Li and Durbin 2010) using default settings, followed by mapping of all
204 unmapped reads using Stampy v1.0.20 (Lunter and Goodson 2010). This approach

combines the rapid but strict BWA algorithm to first map the relatively "easy-to-align" reads, with the more sensitive but computationally intensive Stampy algorithm, which more effectively and accurately aligns the relatively divergent reads (Lunter and Goodson 2010). All reads with mapping quality scores below 20 were excluded. Optical duplicates were then removed using Picard v1.79 (http://picard.sourceforge.net/) and assemblies were improved around indels using the GATK v3.2 Indel Realigner (McKenna *et al.* 2010; DePristo *et al.* 2011). The Unified Genotyper (DePristo *et al.* 2011) was then used to call indels and SNPs for each individual genome. Among the indel calling criteria, >50% of the reads at a given position had to support the existence of that indel, with a minimum of 3 reads containing the variant. For SNP calling in this first round, we required a minimum phred-scaled quality value of 31, and that >75% of reads at a given position support the SNP. For the second round of assembly, the SNPs and indels called in the first round were introduced into the *D. melanogaster* reference, and this modified reference was then used for a second round of mapping. Following indel realignment, the Unified Genotyper was then used to call all sites in the modified reference genome. To generate reference-numbered consensus sequences, a custom perl script was used to shift all base coordinates back to those of the original *D. melanogaster* reference. Deletions and all sites within 3 bp of a called indel were coded as "N" (based on the error analysis described in the results section), while insertions do not appear in reference-numbered consensus sequences.

Consensus error rate and sequence generation

To estimate the actual error rate of our assemblies, and to determine the optimal trade-off between error rate and genomic coverage (the number of euchromatic bases with

228 called alleles), we evaluated base-calling accuracy using the previously published
229 resequenced reference genome (y^1 cn^1 bw^1 sp^1; Pool et al. 2012), sequenced on a GAIIx to
230 ~25X average depth with 76 bp paired end reads (Table S1). Variation was simulated via
231 dwgsim (https://github.com/nh13/DWGSIM/wiki); we introduced substitutions randomly
232 across the genome at a rate of 0.012/bp, with an indel rate of 0.002/bp (with a probability
233 of 0.6 of indel extension). This variation was used to produce a modified reference
234 sequence, as in Pool *et al.* (2012). The resequenced reference reads were then mapped to
235 the modified reference using the pipeline described above, as well as several variations, to
236 investigate the performance of our pipeline versus more standard alignment approaches
237 and various degrees of filtering. Analysis of simulated sequence reads via dwgsim gave
238 highly concordant results (not presented).
239
240 ***Heterozygosity filtering***
241 For the DGRP data set, the EG and KM samples from the AGES data set, and the ZK
242 genomes from the DPGP2 data set, libraries were constructed from pools of flies following
243 varying degrees of inbreeding. For these genomes, tracts of heterozygosity can remain (and
244 can even be substantial), presumably due to the presence of multiple recessive lethal and
245 sterile mutations that are segregating in repulsion. These linked lethals may often occur,
246 for example, within large inversions that are polymorphic within a line and suppress
247 crossing over. To allow consistency between haploid and diploid genomes, entire
248 heterozygous regions must be filtered out prior to generating homozygous consensus
249 sequences.

For the samples mentioned above, the Unified Genotyper was run in diploid mode to enable the calling of heterozygous sites. To identify and mask residually heterozygous regions, we scanned the 5 euchromatic arms of each diploid genome for heterozygous calls in 100 kb windows advancing in 5 kb increments. Rather than use a hard boundary for delineating windows of residual heterozygosity, we chose to scale the threshold for a given window to the level of genetic diversity observed in that window within either sub-Saharan or cosmopolitan populations (henceforth referred to as π_{sub} and π_{cos}, respectively), depending on the geographic origin of each individual genome. To determine these thresholds, we estimated nucleotide diversity (π) in 100 kb windows advancing in 5 kb increments for the large Rwandan (RG) sample of 27 haploid embryo genomes and the French sample of 9 haploid embryo genomes to represent sub-Saharan and cosmopolitan diversity, respectively. If the proportion of heterozygous sites in a given window exceeded $\pi/5$, that window was enucleated, and this window was extended in both directions on the chromosome arm until encountering a window with heterozygosity less than $\pi/20$. This procedure was conducted beginning from both ends of each chromosome arm. Tracts of heterozygosity were masked to N, and are provided in Table S3.

For the DGRP data set, a subset of these genomes had elevated baseline levels of heterozygosity for unknown technical reasons. For the majority of these genomes, this constituted <10% of all sites, while 29 genomes had >10% of sites masked for this reason (Table S3). Regions with elevated numbers of putatively heterozygous sites were masked from consensus sequences regardless of whether they reflected true heterozygosity, cryptic structural variation, or technical artefacts. However, we also used a normalization approach to identify the DGRP genome regions that reflect genuine heterozygosity. We

generated normalization factors using the following procedure: (1) for each euchromatic chromosome arm in each genome, first determine the mode of heterozygous calls per site (hets/site) in the same windows above (in bins of 0.00001), only including windows with a hets/site between 0 and that window's $\pi_{cos}/2$ to remove the effects of true heterozygosity in determining the mode of the baseline ("genomic noise factor"); (2) obtain each genome's normalization factor by dividing the above genomic noise factor by the mode of all DGRP genomic noise factors, truncating this normalization factor at 1 (since we are only interested in reducing the influence of non-genuine heterozygosity calls). (3) appropriate for the identification of true heterozygosity, using the criteria described above.

While haploid embryo genomes are not expected to contain any true heterozygosity, repetitive and/or duplicated regions can cause mismapping that results in tracts of "pseudoheterozygosity". To detect these tracts and remove them, we implemented the same threshold approach as outlined above (without normalization, since none of these genomes showed elevated background levels of putative heterozygosity). For these genomes, the Unified Genotyper was run in haploid mode, and so read proportions were analyzed in place of called heterozygous sites. For windows with the proportion of sites with <75% of the reads matching the consensus base above $\pi/5$, that window was enucleated, and this window was extended in both directions until encountering a window below $\pi/20$. This procedure was conducted starting from both ends of the chromosome arms, overlapping windows were merged, and all pseudoheterozygosity tracts are reported in Table S3.

296 ***Chromosomal inversion detection***

297 Chromosomal inversions are known to be common in natural *D. melanogaster*
298 populations (*e.g.*, Krimbas and Powell 1992; Aulard et al. 2002), and can significantly
299 impact the distribution of genetic diversity (*e.g.*, Kirkpatrick and Barton 2006; Hoffman and
300 Rieseberg 2008; Corbett-Detig and Hartl 2012). For the Drosophila Genome Nexus, we
301 compiled known inversions for the previously published genomes and also identified
302 inversions for the newly sequenced genomes. For DGRP genomes, inversions were
303 previously identified cytogenetically (Huang *et al.* 2014). For the DPGP2 dataset, common
304 inversions were previously detected using the approach of Corbett-Detig *et al.* (2012). For
305 the newly sequenced DPGP3 and AGES datasets, inversions were also detected using this
306 method, and we provide the identified inversions for all of the analyzed genomes in Table
307 S5.
308
309 ***Detection of identical-by-descent (IBD) genomic regions***
310 Tracts of IBD may reflect the sampling of related individuals, and can contradict
311 theoretical assumptions and complicate many population genetic analyses. To identify
312 tracts of IBD, we implemented the approach of Pool et al. (2012), but with slight
313 modifications for the diploid genomes and for the large DPGP3 population sample
314 (described below). All possible pairwise comparisons were made for each of the five
315 euchromatic arms of each genome, and pairwise differences per site were calculated in 500
316 kb windows advanced in 100 kb increments. Windows with less than 0.0005 pairwise
317 differences per site were deemed putatively IBD. Some chromosomal intervals (including
318 centromere- and telomere-proximal regions) exhibited large-scale, recurrent IBD between

319 populations suggesting explanations other than close relatedness, and therefore did not
320 contribute to a genome's IBD total unless they extended outside these recurrent IBD
321 regions. Elsewhere, within population IBD (presumably due to very recent common
322 ancestry) was determined to be that which totaled genome-wide >5 Mb for a pairwise
323 comparison of genomes.

324 For the DPGP2 and AGES genomes, we excluded the same recurrent IBD regions as
325 those of Pool et al. (2012). However, for the much larger DGRP and DPGP3 samples of
326 genomes, we visually reexamined these recurrent IBD tracts and generated new regions to
327 be excluded for each of these data sets (provided in Table S4).

328 Due to heterozygosity filtering, some diploid genomes had genomic coverages far
329 less than the typical ~111 Mb. Therefore, the genome-wide threshold for IBD filtering was
330 adjusted to 5% of all called positions rather than 5 Mb. In addition, only 500 kb windows
331 with >100 kb pairwise comparisons were allowed to contribute to the 5% total, minimizing
332 the influence of windows with large numbers of masked sites.

333

334 ***Detection of cosmopolitan admixture***

335 Because admixture from cosmopolitan gene flow into Africa can significantly impact
336 estimates of genetic diversity and violate demographic assumptions of some analyses, it is
337 important to identify instances of cosmopolitan ancestry in the sub-Saharan genomes. We
338 used the HMM approach outlined in detail by Pool et al. (2012), but with updated reference
339 panels. The sub-Saharan reference panel included 27 Rwanda (RG) genomes, but
340 chromosome arms with known inversions were excluded (as identified by Corbett-Detig
341 and Hartl 2012). The cosmopolitan reference panel included 9 France genomes, again

342 excluding inversions, since inverted arms were previously found to have unusually high
343 divergence from standard arms in this population (Pool et al. 2012; Corbett-Detig and Hartl
344 2012). Pairwise distance comparisons indicated that the Egyptian genomes were
345 genetically cosmopolitan. To augment the cosmopolitan sample, we included homozygous
346 regions of standard arrangement Egypt chromosome arms in the cosmopolitan reference
347 panel.
348 Aside from these modifications to the reference panels, we implemented the
349 admixture HMM as described in Pool et al. (2012). Briefly, this HMM works in the following
350 way. For a focal African genome within a particular window, the method compares it
351 against a cosmopolitan reference panel and assesses whether its genetic distance to this
352 reference panel is on the level expected for a sub-Saharan genome, or if instead this
353 resembles a comparison of one cosmopolitan genome against others (indicating
354 admixture). As before, window size for the analysis was based on 1000 non-singleton SNPs
355 among the RG sample, roughly corresponding to a mean window size of 50 kb. The analysis
356 was initially calibrated using the 27 RG genomes to represent the putative non-admixed
357 state, and emission distributions for the non-admixed state were generated as in Pool et al.
358 (2012). A revised sub-Saharan panel was then generated through an iterative analysis of
359 the RG genomes. Following a single round of the method, RG genomes were masked for
360 admixture, and then these masked genomes served as the African panel for a second round.
361 RG genomes were then masked for admixture again, and a third round of the method was
362 applied to the RG genomes to produce a final set of emissions distributions that were used
363 in the analysis of all other African genomes.
364

365 ***Genetic diversity and population structure***

366 For all of the analyses described below, only heterozygosity- and IBD-filtered
367 genomes were utilized. Sub-Saharan genomes were also filtered for cosmopolitan
368 admixture as detailed above. Nucleotide diversity (π) was initially calculated in windows of
369 2,000 non-singleton RG SNPs, corresponding to a median window size of 100 kb for all
370 populations with at least two genomes sampled. For more efficient analysis of the large
371 Raleigh (DGRP) and Siavonga, Zambia (DPGP3) populations, we selected 30 genomes from
372 each of these populations with the highest genomic coverage and with at least 30X average
373 depth. To remove the effects of spurious estimates due to low coverage windows, we
374 excluded windows for a given population if site coverage (the number of sites with alleles
375 called for two or more genomes) was below half the coverage in the large RG sample for
376 that window. To obtain whole-arm and genome-wide estimates, we conducted a weighted
377 average of windows (weighted by the number of sites in each window with data from at
378 least two genomes).

379 To examine patterns of population structure, we calculated D_{xy} and F_{ST} (Hudson *et*
380 *al.* (1992) for all populations with at least two high-coverage genomes (after IBD and
381 admixture filtering), and including the *D. melanogaster* reference genome for *Dxy*. Both
382 analyses were conducted in windows of 2,000 non-singleton RG SNPs, and a weighted
383 average of windows was used to obtain whole-arm and genome-wide estimates. In
384 addition, to lessen the influence of large inversions on estimates of genetic diversity and
385 population structure, we estimated nucleotide diversity and pairwise F_{ST} excluding
386 inverted arms for a subset of populations with larger sample sizes (inversion
387 presence/absence is given in Table S5).

Results

The Drosophila Genome Nexus

The resulting data set, which we have named the Drosophila Genome Nexus (http://www.johnpool.net/genomes.html), consists of 605 sequenced genomes (varying slightly in number among the 5 euchromatic chromosomal arms) from 36 populations from Africa, Europe, and North America. The consensus sequences analyzed below and made available online include only the 5 euchromatic chromosome arms. These consensus sequences have been filtered for heterozygosity, with additional files provided to facilitate masking of IBD and cosmopolitan admixture as well as locus-specific analysis. SNP and indel variant call files (VCFs) are also available online, both for these 5 arms and for other arms (mitochondria, chromosomes 4 and Y, and heterochromatic components of the euchromatic arms). The repetitive nature of non-euchromatic arms may entail much higher error rates; we do not focus on their analysis here.

While the consensus sequences made available online specifically focus only on SNP variants, the provided indel vcf files will also be of considerable utility. For indels, the Unified Genotyper is limited to detecting only those encapsulated entirely within a single read. Therefore, read lengths will limit the size of detected indels. To examine the extent of this effect on indel detection, we examined indel length distribution for two DPGP3 genomes with 76 bp paired-end reads as well as for two DPGP3 genomes with much longer read lengths (146 and 150 bp; Fig. S2), each with similarly low cosmopolitan admixture and high mean depth. For indels approximately 25 bp or shorter, the long and short read

lengths appeared to have no effect on indel length frequencies. However, for longer indels (>25 bp) the gap in detection between the two read lengths gradually increased, illustrating the decreasing ability of the present approach to detect indels as they approach the read length. This potential bias is important to consider when examining the provided indel calls. A more comprehensive analysis of structural variation within and between these populations will be a target of future research.

While there is considerable variation among genomes in terms of average sequencing depth and coverage, the majority of this variance lies in the DGRP and AGES data sets, which range from approximately 12X to more than 100X mean depth, while the remaining genomes are primarily haploid embryo genomes of roughly 30X mean depth or higher (Table S1). In addition, coverage varies considerably among the inbred/isofemale genomes from the AGES and DGRP data sets due to heterozygosity filtering.

Genome assembly pipeline performance

To investigate pipeline performance, basecalling bias, and consensus error rate, we assembled a resequenced *D. melanogaster* reference strain to an artificially mutated reference genome to simulate variation. Overall, adding a second round of mapping that incorporated SNP and indel variants called in the first round of mapping resulted in approximately a 1% increase in sequence coverage (just over 1 million sites added) and a significant improvement in error rate relative to performing only a single round of mapping with only BWA (Fig. 1). This improvement was observed irrespective of the nominal quality value threshold for basecalling (which had only a modest effect on error rates), with error rates for the two round assemblies completely distinct from the

434 distribution of error rates observed for a single round of assembly. We investigated the
435 contribution of Stampy to this improvement and found that, while error rate and coverage
436 both improved, the vast majority of improvement was due to adding the second round of
437 mapping (Fig. 1).

438 We also investigated the impact of filtering around indels, as past analyses have
439 found that positions directly adjacent to indels are difficult for aligners to correctly align
440 and a major contributor to error (Meader *et al.* 2010; Alkan *et al.* 2011). We found similar
441 results, with approximately a five-fold reduction in error rate by masking 3 basepairs on
442 either side of consensus indels. Assessing the possible benefit of masking 5 bases rather
443 than 3, we observed almost no improvement in error rate to justify the nearly 1%
444 reduction in coverage, and therefore used the 3 bp mask. Our use of the GATK Indel
445 Realigner (McKenna *et al*. 2010; DePristo *et al*. 2011), in conjunction with incorporating
446 indels into the reference used in the second round of mapping, may have improved our
447 ability to align around these regions. Finally, to determine the optimal alignment quality
448 value threshold for consensus sequence generation and to estimate the expected error rate
449 of our assemblies, we examined the tradeoff between coverage and error rate at a range of
450 quality values for both the haploid and diploid callers of the Unified Genotyper (DePristo *et*
451 *al.* 2011), and we selected a minimum of Q75 and Q32 for calling a position in haploid or
452 diploid genomes, respectively (Figs. S3 and S4, respectively). These thresholds
453 corresponded to an error rate of roughly 1.36×10^{-5} errors per site.

454 To further examine our two round pipeline performance, we compared sites called
455 only by the two round pipeline versus those called using just a single round of BWA to map,
456 and estimated both error rate and diversity for both classes of sites. In terms of error rate,

457 sites called in both pipelines possessed an error rate of 9.3 x 10^{-6}, just below the genome-
458 wide rate, while sites added only in our two round pipeline had an error rate of 2.9 x 10^{-5},
459 roughly two fold higher than our genome-wide average. This increase in error rate is not
460 surprising given that these sites added by our two round pipeline likely constitute highly
461 diverse, hard-to-align regions relative to those confidently called by both pipelines.
462 To further examine the sites added by our two round pipeline, we identified bases
463 called in both pipelines vs. those called only in our two round pipeline for a single RG
464 genome (RG33), and then calculated nucleotide diversity at each class of sites for the RG
465 sample of 27 genomes. In addition, for two RG genomes sequenced at similar depths (RG33,
466 RG5) we determined both the number of sites added by our two round pipeline and the
467 number of indels (indel "rate") in 100 kb non-overlapping windows. For sites called only
468 by our two round pipeline, nucleotide diversity was over three fold higher than that for
469 sites called in both pipelines (Table 1), and we observed a clear positive relationship
470 between the number of sites added with the new pipeline and the number of indels in that
471 region of the genome (Fig. 2). These lines of evidence support the idea that the sites added
472 by our pipeline are found in high diversity, difficult-to-align genomic regions.
473 To determine whether any particular functional class of sites contributed
474 disproportionately to the sites added by our two round pipeline, we used the *D.*
475 *melanogaster* reference genome annotations (v5.57) to assign each individual site called
476 only by our two round pipeline for RG33 and RG5 to one of 9 site classes: Nonsynonymous,
477 2-, 3-, or 4-fold synonymous, 5' UTR, 3' UTR, intronic, short intron (Halligan and Keightley
478 2006), or intergenic. While all classes of sites contributed to the total sites added by our
479 two round pipeline, only intergenic and intronic sites were positively enriched for both RG

480 genomes (Fig. 3), suggesting our pipeline disproportionately added these two functional
481 classes of sites to the assemblies. However, the representation of each functional category
482 in the "added sites" class is fairly close to null expectations, and we even added 41,368 and
483 36,990 nonsynonymous sites, which we would expect to be the least diverse and therefore
484 easiest to align confidently, to RG33 and RG5, respectively, by applying the full pipeline. We
485 also characterized the tract length and genomic distribution of sites added in the second
486 round of our assembly for both RG5 and RG33. In terms of tract length, the vast majority of
487 bases added occurred in short tracts of 1 to 10 bases (Fig. S5). To examine genomic
488 location of these sites, we calculated the number of sites in 100 kb windows across the 5
489 euchromatic arms of the genome (Fig. S6). While sites were added somewhat uniformly
490 across the genome, repetitive telomeric regions were especially enriched.
491
492
493 ***Impact of sequencing depth on genetic distance***
494 In a previous assembly of the DPGP2 data set, Pool *et al.* (2012) found a positive,
495 non-linear relationship between mean sequencing depth (the average number of reads per
496 basepair) and genetic distance to both the reference and the Siavonga, Zambia (ZI)
497 population. This relationship is especially pronounced below 25X mean depth. Here, we
498 ameliorated that issue by using a consensus caller that is less vulnerable to reference
499 sequence bias and by adding more stringent quality filtering
500 (http://www.dpgp.org/dpgp2/DPGP2.html). To examine the impact of this basecalling bias
501 in our pipeline, we quantified the recall rate of reference and non-reference alleles in the
502 resequenced reference. The recall rate for reference and non-reference alleles was nearly

503 identical (0.958 vs 0.959), suggesting reference bias has a minimal effect. To further
504 examine this relationship for our two round pipeline versus the single round of mapping
505 with only BWA (but including the indel filter), we calculated mean genome-wide distance
506 to the ZI population for each of the genomes in the AGES data set (excluding genomes with
507 whole arms masked due to heterozygosity). For the single round of mapping, a positive
508 relationship between mean depth and distance to ZI was apparent below 20X mean depth,
509 but for our two round pipeline distances remained flat even approaching 10X depth (Fig.
510 4A). When limiting this analysis to sites called in all analyzed genomes, the bias observed
511 below 20X mean depth for single round genomes disappeared, and distance estimates were
512 essentially identical to those of the two round pipeline (though greatly reduced in both
513 cases, reflecting the exclusion of more diverse genomic regions). These results suggest that
514 the depth-related bias observed for the single round alignments (Fig. 4A) was not due to
515 biased consensus-calling (since that would still affect the filtered analysis), but instead
516 stems from differences in genomic coverage between low and high depth genomes. And
517 indeed, we observe that genomic coverage is more dependent on depth in the single round
518 alignments than for the full pipeline (Fig. 4B).
519
520 ***Heterozygosity***
521 Heterozygosity can persist in fly stocks even after many generations of full-sibling
522 mating, probably due to the presence of recessive lethal or infertile mutations, which are
523 commonly found on wild-derived *Drosophila* chromosomes (Greenberg and Crow 1960).
524 Especially when combined with inversion polymorphism (*e.g.* one recessive lethal is fixed
525 on the inversion-bearing chromsomes, and a different recessive lethal is fixed on the

526 standard arms), recombination may be unable to generate reproductively viable
527 homozygous progeny, and residual heterozygosity may extend over much of a chromosome
528 arm.

529 We report heterozygosity tracts in Table S3, including those for the Egypt EG, Kenya
530 KM, and Zimbabwe ZK samples. The largest non-isogeneous sample in our analysis is the
531 205 DGRP genomes originating from Raleigh, North Carolina, USA. In spite of 20
532 generations of full-sib mating for the DGRP lines, considerable residual heterozygosity was
533 maintained within the inbred lines. Overall, 12.6% of the total genomic sequence was
534 masked due to apparent heterozygosity, and for each autosomal arm there were multiple
535 fly lines for which the entire chromosome arm remained heterozygous. Considerably less
536 masking was needed on the X chromosome than on the autosomes, which is expected given
537 the increased efficacy of selection against recessive lethals and steriles in hemizygous
538 males.

539 To examine the role of inversions in maintaining heterozygosity in inbred lines, we
540 obtained inversion genotypes for each euchromatic arm of the DGRP lines from Huang *et al.*
541 (2014). As is evident from the distribution of heterozygosity proportions for inverted vs
542 standard autosomal arms (Fig. 5 and Table S7), more than 80% of the chromosome arms
543 with inversion polymorphism retain over 95% heterozygosity, compared to chromosome
544 arms lacking inversion polymorphism for which more than 80% retained less than 10%
545 heterozygosity. These results support the role of recessive deleterious mutations residing
546 within large inversions driving chromosome arm-wide residual heterozygosity, but fail to
547 explain the remaining residual heterozygosity evident in the standard arm distribution
548 shown in Fig. 5 (6.5% of non-inverted chromosome arms still retained >25%

549 heterozygosity after 20 generations of inbreeding). While it is possible that inversion
550 differences between sequenced and karyotyped sub-lines might exist in some cases,
551 another explanation is that multiple recessive lethals in repulsion on a single chromosome
552 arm might reduce the rate at which viable recombinants arise during inbreeding (Falconer
553 1989).

554 In addition to true heterozygosity, artifactual "heterozygosity"
555 (pseudoheterozygosity) can result from mismapping or other technical issues with genome
556 assembly. For the haploid embryo genomes presented here, these positions constituted a
557 very small proportion of total sites in a given genome (Mean = 0.00798; SD = 0.00501; Min.
558 = 0.00025; Max. = 0.0310).

559

560 ***Identity by descent***

561 Identity-by-descent (IBD) regions passing all filters were flagged and are provided
562 as an optional filter in the Drosophila Genome Nexus release (all IBD tracts are given in
563 Table S6). For the DPGP2 data set, the IBD tracts we identified were essentially identical to
564 those of Pool *et al*. (2012) and therefore are not discussed. For the AGES data set, we
565 detected IBD for a single pair of samples from the SF South African population, but this
566 segment included one long tract encompassing all of Chr3R and half of Chr3L. For the
567 DPGP3 data set, we detected IBD for 20 sample pairs, constituting only 3.2% of all called
568 bases and 0.1% of all pairwise comparisons for those 197 genomes. For the 205 DGRP
569 genomes, IBD appeared to be more widespread, with 9.8% of all called bases flagged for
570 masking. For a case of two IBD genomes, these base counts refer to only the masked
571 individual, and a total of 54 IBD sample pairs were detected.

572

573 ***Cosmopolitan admixture in African genomes***

574 It has previously been noted that the introgression of cosmopolitan alleles into some
575 African populations could have a significant influence on genetic diversity within Africa
576 (Begun and Aquadro 1993; Capy *et al.* 2000; Kauer *et al.* 2003), and cosmopolitan
577 admixture proportions were previously estimated for the DPGP2 data set by Pool *et al.*
578 (2012). We repeated this analysis for all sub-Saharan genomes published here, but with
579 improved reference panels (including more genomes, but excluding inverted arms – see
580 Materials and Methods). All identified cosmopolitan admixture tracts are given in Table S8,
581 and are provided as an optional sequence filter in the Drosophila Genome Nexus release.
582 As in the DPGP2 analysis (Pool *et al.* 2012), admixture varied considerably among
583 populations, from <1% in the Ethiopian EM population to > 80% in the Zambian ZL
584 population (Fig. S7). Within-population variation was also striking, as is evident from
585 individual genome plots of admixture (Fig. 6). One important exception to the high level of
586 inter-individual variation in cosmopolitan admixture proportions was the large DPGP3
587 sample (ZI) from Siavonga, Zambia. We targeted this population sample for large-scale
588 genome sequencing for multiple reasons, including its hypothesized position within the
589 ancestral range of *D. melanogaster* (showing maximal genetic diversity), as well as its
590 relatively low level of cosmopolitan admixture among four genomes surveyed in the DPGP2
591 analysis (Pool *et al.* 2012). Our analysis of the larger DPGP3 data set illustrates that the ZI
592 population does in fact have a very low level of cosmopolitan admixture, with the
593 population average at 1.1% of the genome, the highest individual genome at 26%, and the
594 second highest at 9% (Fig. 6). Looking across the genome, DPGP3 is similar to other sub-

595 Saharan genomes in having the lowest admixture levels on the X chromosome (Fig. S8), but
596 has a pronounced increase in the middle of arm 3R (roughly 7.6 Mb to 15.0 Mb), where up
597 to 13 putatively admixed genomes are found in the maximal window (6.6% of the sample),
598 compared with a genome-wide median of just 2 out of 197 individuals.
599
600 ***Genetic diversity and structure***
601 Although the present study is not primarily focused on population genetic analysis,
602 we present a few simple summaries of the data to guide potential users of these assemblies.
603 First, we estimated nucleotide diversity for all populations with multiple high coverage
604 genomes for all chromosome arms, both including and excluding inverted arms. For the
605 DPGP2 data set, nucleotide diversity was largely consistent with the previous estimates of
606 Pool *et al.* (2012), although estimates for the newly assembled sequences were generally
607 slightly higher (Table S9), perhaps due to the improved coverage of more diverse regions.
608 Nucleotide diversity comparisons among populations revealed similar patterns of
609 past analyses (Table S9). The France, Egypt, and U.S. (DGRP RAL) populations had much
610 lower diversity levels than any sub-Saharan populations (Table S9), particularly on the X.
611 This strong reduction in diversity has been previously documented (Begun and Aquadro
612 1993; Baudry *et al*. 2004) and presumably results from the bottleneck that occurred during
613 expansion out of sub-Saharan Africa. With additional sub-Saharan African genomes, as well
614 as the expansion of the Siavonga, Zambia population to nearly 200 genomes, south-central
615 Africa remains the most diverse portion of the *D. melanogaster* distribution. While
616 Siavonga, Zambia still has the highest nucleotide diversity at 0.854% (Table S9), samples
617 from Zimbabwe and inland South Africa reach 0.814% to 0.850%. The ancestral range of

618 the species may have included much of southern Africa, unless a more recent expansion
619 occurred with very little loss of diversity. Both eastern and western African populations
620 were still reduced in diversity relative to southern Africa (generally 0.73% to 0.80%). Pool
621 *et al.* (2012) reported a further, mild diversity reduction in Ethiopian highland populations
622 was as previously described (Pool *et al.* 2012), but a lowland sample from far western
623 Ethiopia (EA) showed little diversity reduction.

624 To examine the effects of inversions on diversity at the genome-scale, we estimated
625 nucleotide diversity with inverted arms removed (Table 2). Previous analyses revealed that
626 the effects of inversions on nucleotide diversity were not limited to regions surrounding
627 breakpoints, but could affect entire chromosome arms (Corbett-Detig and Hartl 2012, Pool
628 *et al.* 2012). Among the studied sub-Saharan populations, inversions appeared to have
629 effects of both elevating and reducing arm-wide diversity (Table 2). The North American
630 RAL sample showed less diversity elevation from inversions compared with the European
631 FR sample.

632 Estimation of F_{ST} and D_{xy} revealed patterns similar to those of Pool *et al.* (2012) for
633 the DPGP2 populations (Table S9). Within the population groupings identified in that
634 study, population differentiation was particularly low among southern African populations
635 (mean F_{ST} = 0.0092), and somewhat elevated among Ethiopian samples (mean F_{ST} =
636 0.0331) – which as previously observed, showed moderate differentiation from other sub-
637 Saharan samples (Table S9). Examination of F_{ST} restricted to standard chromosome arms
638 indicated mainly small effects of inversions on genetic differentiation: in some cases the
639 addition of inversions increased genetic differentiation (*e.g.* Nigeria NG vs. other sub-
640 Saharan samples), while in other cases inversions decreased genetic differentiation (*e.g.* for

641 comparisons involving the France or U.S. samples). Concordant with previous observations
642 (*e.g.* Caracristi and Schlotterer 2003; Haddrill *et al.* 2005) and the hypothesized admixed
643 origin of New World populations from European and African sources, standard arms from
644 the North American RAL sample had consistently higher diversity than the European FR
645 sample (Table 2), as well as closer relationships to sub-Saharan populations (Table 3).
646
647 **Discussion**
648 We have presented a set of 605 consistently aligned *D. melanogaster* genomes.
649 Although our pipeline primarily makes use of published methods, the resulting alignments
650 are expected to yield a better combination of accuracy and genomic coverage than standard
651 approaches. However, the primary motivation for the Drosophila Genome Nexus is to
652 increase the comparability of population genomic data sets, as well as make available more
653 than 250 additional genomes, including the large Siavonga, Zambia sample.
654 Our effort accounts for one category of potential biases between data sets
655 (differences in alignment methodology and data filtering), but other potential concerns
656 should still be recognized. Although not addressed here, differences in data generation,
657 including (but not limited to) methods of obtaining genomic DNA and sequencing
658 platform/chemistry, may influence the resulting genomic data (Quail *et al.* 2012; Ratan *et*
659 *al.* 2013; Solonenko *et al.* 2013). Our pipeline reduces the population genetic consequences
660 of differences in sequencing depth, but depth still has an important influence on genomic
661 coverage. Mapping success may vary according to a genome's genetic similarity to the
662 reference sequence, which for *D. melanogaster* is expected to have primarily cosmopolitan
663 origin. This genetic similarity to the reference sequence will vary geographically (*e.g.* sub-

664 Saharan genomes being more genetically distant from the reference) and across the
665 genome (especially for admixed populations). Demography may also bias downstream
666 population genetic analyses: for example, recent admixture and identity-by-descent are
667 contrary to the predictions of models that assume random sampling of individuals from
668 large randomly mating populations (thus we provide filters to reduce the effects of these
669 specific issues).

670 It should be emphasized that the present DGN is primarily aimed at SNP-oriented
671 analysis of the five major euchromatic chromosome arms. Aside from inversion-calling and
672 the detection of short indels, we do not address the important topic of structural variation.
673 Furthermore, the challenge of reliably aligning heterochromatin and other repetitive
674 regions (on a population scale) awaits further technological and methodological progress.

675 Thorough population genetic analysis of the DPGP3 Zambia (ZI) population sample
676 will be a topic of future analyses. However, the preliminary statistics reported here
677 support the notion that these genomes will be widely utilize in the field of population
678 genetics. This population continues to present the maximal genetic diversity of any *D.*
679 *melanogaster* population studied to date, offering hope that it may be the least affected by
680 losses of genetic diversity via expansion-related population bottlenecks. Unlike many sub-
681 Saharan populations, it also contains very little cosmopolitan admixture. The availability of
682 nearly 200 genomes from this single sub-Saharan population sample, which may have a
683 relatively simpler demographic history than many *D. melanogaster* populations, will be an
684 asset for studies seeking to understand the genetic, selective and demographic mechanisms
685 that shape genomic polymorphism and divergence in large populations.

686

687 **Acknowledgements**

688 We thank Stephen Richards for assistance with the DGRP data, J. J. Emerson for packaging
689 the admixture HMM method, and Isaac Knoflicek for help with data management and the
690 DGN web site. We also acknowledge the University of Wisconsin Center for High
691 Throughput Computing (CHTC) for computational resources and assistance regarding our
692 alignments. Funding was provided by NIH grant HG02942 to CHL, NIH grant R01
693 GM111797 to JEP, and support to JBL from a Ruth L. Kirschstein National Research Service
694 Award (F32 GM106594) and from the University of Wisconsin-Madison Genome Sciences
695 Training Program (GSTP).

696

697 **Literature Cited**

698 Alkan, C., S. Sajjadian, and E. E. Eichler, 2011 Limitations of next-generation genome
699 assembly. Nature Methods 8: 61–65.

700 Aulard, S., J. R. David, and F. Lemeunier, 2002 Chromosomal inversion polymorphism in
701 Afrotropical populations of *Drosophila melanogaster*. Genetical Research 79: 49–63.

702 Baudry, E., B. Vinier, and M. Veuille, 2004 Non-African populations of *Drosophila*
703 *melanogaster* have a unique origin. Mol. Biol. Evol. 21: 1482–1491.

704 Begun, D. J., and C. F. Aquadro, 2006 African and North American populations of *Drosophila*
705 *melanogaster* are very different at the DNA level. Nature 365: 548–550.

706 Capy, P., M. Veuille, M. Paillette, M. Jallon, J. M. Vouidibio *et al.*, 2000 Sexual isolation of
707 genetically differentiated sympatric populations of *Drosophila melanogaster* in
708 Brazzaville, Congo: the first step towards speciation? Heredity 84: 468–475.

709 Caracristi, G., and C. Schlotterer, 2003 Genetic differentiation between American and
710 European *Drosophila melanogaster* population could be attributed to Admixture of
711 African alleles. Mol. Biol. Evol. 20: 792–799.
712 Chan, A. H., P. A. Jenkins, and Y. S. Song, 2012 Genome-wide fine-scale recombination rate
713 variation in *Drosophila melanogaster*. PLoS Genetics 8: e1003090.
714 Comeron, J. P., 2014 Background selection as baseline for nucleotide variation across the
715 *Drosophila* genome. PLoS Genetics 10: e1004434.
716 Corbett-Detig, R. B., C. Cardeno, and C. H. Langley, 2012 Sequence-based detection and
717 breakpoint assembly of polymorphic inversions. Genetics 192:131–137.
718 Corbett-Detig, R. B., and D. L. Hartl, 2012 Population genomics of inversion polymorphisms
719 in *Drosophila melanogaster*. PLoS Genetics 8: e1003056.
720 DePristo, M., E. Banks, R. Poplin, K. Garimella, J. Maguire *et al.*, 2011 A framework for
721 variant discovery and genotyping using next-generation DNA sequencing data.
722 Nature Genetics 43: 491–498.
723 Duchen, P., D. Zivkovic, S. Hutter, W. Stephan, and S. Laurent, 2013 Demographic inference
724 reveals African and European admixture in the North American *Drosophila*
725 *melanogaster* population. Genetics 193: 291–301.
726 Falconer, D. S., 1989 *Introduction to Quantitative Genetics*. Longmans Green/John Wiley &
727 Sons, Harlow, Essex, UK/New York.
728 Greenberg, R., and J. F. Crow, 1960 A comparison of the effect of lethal and detrimental
729 chromosomes from Drosophila populations. Genetics 45: 1153–1168.

730 Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns
731 of nucleotide variability and the demographic and selection history of *Drosophila*
732 *melanogaster* populations. Genome Res. 15: 790–799.
733 Halligan, D. L., and P. D. Keightley, 2006 Ubiquitous selective constraints in the *Drosophila*
734 genome revealed by a genome-wide interspecies comparison. Genome Res. 16: 875–
735 884.
736 Hoffman, AA., and L. H. Rieseberg, 2008 Revisiting the impact of inversion in evolution:
737 from population genetic markers to drivers of adaptive shifts and speciation? Annu.
738 Rev. Ecol. Evol. Syst. 39: 21–42.
739 Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia *et al.*, 2014 Natural variation in
740 genome architecture among 205 *Drosophila melanogaster* genetic reference pane
741 lines. Genome Res. 24: 1193–1208.
742 Hudson, R. R., M. Slatkin, W. P. Madison, 1992 Estimation of levels of gene flow from DNA
743 sequence data. Genetics 132: 583–589.
744 Kauer, M. O., D. Dieringer, and C. Schlotterer, 2003 A microsatellite variability screen for
745 positive selection associated with the 'out of Africa' habitat expansion of *Drosophila*
746 *melanogaster*. Genetics 165: 1137–1148.
747 Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and
748 speciation. Genetics 173:419–434.
749 Krimbas, C. B., and J. R. Powell, 1992 Drosophila Inversion Polymorphism. CRC Press, Boca
750 Raton, FL.

751 Lachaise, D., M. L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, M. Ashburner, 1988 Historical
752 biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22: 159–
753 225.
754 Langley, C. H., M. Crepeau, C. Cardeno, R. Corbett-Detig, and K. Stevens, 2011
755 Circumventing heterozygosity: sequencing the amplified genome of a single haploid
756 *Drosophila melanogaster* embryo. Genetics 188: 239–246.
757 Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider *et al.*, 2012 Genomic
758 variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533–598.
759 Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler
760 transform. Bioinformatics 26: 589–595.
761 Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-
762 genome sequences. Nature 475: 493–497.
763 Long, Q., F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow *et al.* 2013 Massive genomic
764 variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nature
765 Genetics 45: 884–890.
766 Lunter, G., and M. Goodson, 2010 Stampy: a statistical algorithm for sensitive and fast
767 mapping of Illumina sequence reads. Genome Res. 18: 821–829.
768 Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The
769 *Drosophila melanogaster* Genetic Reference Panel. Nature 482: 173–178.
770 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome
771 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
772 sequencing data. Genome Res. 20: 1297–1301.

773 Meader, S., L. W. Hillier, D. Locke, C. P. Ponting, and G. Lunter, 2010 Genome assembly
774 quality: assessment of and improvement using the neutral indel model. Genome
775 Research 20: 675–684.
776 Pool, J. E., 2009 Notes regarding the collection of African *Drosophila melanogaster*. Dros.
777 Inf. Serv. 92: 130–134.
778 Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno *et al.*, 2012
779 Population genomics of Sub-Saharan *Drosophila melanogaster*: African diversity and
780 non-African admixture. PLoS Genetics 8: e1003080.
781 Quail, M. A., M Smith, P. Coupland, T. D. Otto, S. R. Harris *et al.*, 2012 A tale of three next
782 generation sequencing platforms: comparison of Ion Torrent, Pacific Bioscience and
783 Illumina MiSeq sequencers. BMC Genomics 13: 341.
784 Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri *et al.*, 2013 Comparison of sequencing
785 platforms for single nucleotide variant calls in a human sample. PLoS One 8: e55089.
786 Solonenko, S. A., J. C. Ignacio-Espinoza, A. Alberti, C. Cruaud, Steven Hallam *et al.*, 2013
787 Sequencing plastform and library preparation choices impact viral metagenomes.
788 BMC Genomics 14: 320.
789 The 1000 Genomes Project Consortium, 2010 A map of human genome variation from
790 population-scale sequencing. Nature 467:1061–1073.
791 Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M. *et al.*, 2014 A worldwide survey of
792 genome sequence variation provides insight into the evolutionary history of the
793 honeybee *Apis mellifera*. Nature Genet., accepted.

Figure Legends

Figure 1. Comparison of genomic coverage and error rate for several genome assembly pipeline variations, based on resequencing of the *D. melanogaster* reference strain (Pool *et al*. 2012). All quality values from Q10 to Q100 are shown; many gave very similar results.

Figure 2. Relationship between the number of indels and the number of sites called by our two round pipeline but not called in a single-round pipeline for two RG genomes (RG5 and RG33). Site counts (y axis) and indel counts (x axis) were determined in 100 kb windows across each genome.

Figure 3. Enrichment of each of 9 annotation classes in the sites added by our two round pipeline, but not called by a single round pipeline, relative to genome-wide frequencies. We examined two RG genomes (RG5 and RG33) with approximately 30X mean depth and comparable coverage.

Figure 4. Mean sequencing depth vs genetic distance (A) from the Zambia population and depth vs coverage (B) for the POOL dataset genomes with high coverage on all chromosome arms (listed in Table S1). Circles indicate comparisons utilizing all windows with called sites, while triangles indicate comparisons including only sites called for all of the POOL and ZI genomes. Comparisons illustrate the effect of depth on genetic distance (A) and coverage (B) for genomes assembled using a single-round pipeline (red), vs our two round pipeline (blue). The two round pipeline appears to alleviate the potential downward bias present in the single-round pipeline for depths below approximately 20X,

817 and the greater impact of depth on coverage for the single-round pipeline (B) suggests the
818 sites added by the two round pipeline are driving the differences in distance to ZI.
819
820 **Figure 5**. A histogram of the proportions of each autosomal chromosome arm called
821 heterozygous from the 205 DGRP genomes. Based on the cytological analysis of Huang *et al.*
822 (2014), red arms were reported to be free of inversion polymorphism, while blue arms
823 contained polymorphic inversions. The greatly increased heterozygosity of the latter
824 category illustrates the effects of inversion polymorphism on inbreeding efficacy.
825
826 **Figure 6**. Heterogeneity in estimated cosmopolitan admixture proportions among
827 individuals for each Sub-Saharan African population.

828 **Table 1**. Chromosome arm nucleotide diversity (π) for the RG population based on sites
829 called in both rounds of our pipeline, and for sites called only by adding the second round
830 of mapping.

	RG Nucleotide Diversity	
Chr	**Both Rounds**	**2nd Round Only**
X	0.0083	0.0277
2L	0.0086	0.0259
2R	0.0077	0.0252
3L	0.0079	0.0248
3R	0.0065	0.0260

831

832

Table 2. Chromosomal arm nucleotide diversity (π) for populations with inversion polymorphism. Nucleotide diversity estimates include both the total data set for a given population (Total) and excluding arms carrying inversions (Standard). "Non" denotes populations with less than two standard chromosome arms; "no inv." denotes populations without inversion polymorphism. Values in bold indicate a difference ≥5%.

	X		2L		2R		3L		3R	
Pop	Standard	Total	Standard	Total	Standard	Total	Standard	Total	Standard	Total
CO	0.0075	0.0075	0.0082	0.0083	0.0073	no inv.	0.0076	no inv.	Non	0.0058
EA	0.0071	0.0071	0.0087	no inv.	0.0074	no inv.	0.0075	no inv.	0.0060	no inv.
EB	0.0064	no inv.	0.0074	no inv.	0.0066	no inv.	0.0067	no inv.	**0.0054**	**0.0060**
EG	0.0034	0.0034	Non	0.0066	0.0053	no inv.	Non	Non	Non	0.0062
FR	0.0036	0.0036	**0.0055**	**0.0061**	0.0051	no inv.	**0.0054**	**0.0063**	**0.0045**	**0.0058**
GA	0.0077	0.0076	**0.0087**	**0.0082**	0.0077	no inv.	0.0080	no inv.	0.0066	0.0068
GU	0.0076	0.0076	0.0083	0.0084	0.0075	no inv.	0.0077	no inv.	Non	0.0066
KN	0.0087	0.0086	**0.0061**	**0.0078**	0.0077	no inv.	0.0082	no inv.	0.0070	0.0073
KR	0.0082	0.0085	Non	0.0068	0.0080	0.0081	0.0085	no inv.	0.0067	no inv.
NG	0.0076	0.0076	**0.0109**	**0.0077**	0.0075	0.0074	**0.0085**	**0.0073**	**0.0054**	**0.0065**
RAL	0.0041	0.0041	0.0068	0.0070	0.0062	0.0064	0.0064	0.0065	0.0052	0.0052
RG	0.0080	no inv.	0.0085	0.0094	**0.0076**	**0.0081**	0.0078	no inv.	0.0063	0.0065
SB	0.0088	no inv.	**0.0106**	**0.0094**	0.0082	no inv.	0.0083	0.0089	0.0072	no inv.
SD	0.0086	no inv.	0.0096	0.0097	0.0078	0.0078	0.0080	no inv.	0.0065	no inv.
SE	0.0083	0.0086	Non	0.0075	0.0079	0.0076	0.0081	no inv.	0.0072	no inv.
SF	0.0086	no inv.	**0.0103**	**0.0094**	**0.0090**	**0.0085**	0.0081	no inv.	**0.0048**	**0.0065**
SP	0.0090	no inv.	0.0099	0.0100	0.0083	0.0082	0.0083	no inv.	0.0065	no inv.
TZ	Non	0.0058	Non	0.0062	0.0077	no inv.	0.0075	no inv.	**0.0056**	**0.0066**
UK	0.0081	0.0081	0.0085	0.0086	0.0077	0.0078	0.0077	no inv.	**0.0060**	**0.0065**
UM	0.0080	0.0081	Non	0.0084	0.0078	no inv.	0.0068	no inv.	**0.0075**	**0.0066**
ZI	0.0089	0.0089	0.0099	0.0097	0.0083	0.0082	0.0084	0.0087	0.0076	0.0076
ZS	**0.0089**	**0.0082**	0.0099	0.0098	0.0083	0.0080	0.0082	no inv.	0.0074	0.0074

Table 3. Pairwise population F_{ST} for select populations averaged across chromosome arms. Comparisons utilizing the total data set for each population is above the diagonal, and comparisons using only arms without inversions are shown below the diagonal.

	FR	GA	NG	RAL	RG	SP	ZI
FR	0.0000	0.1898	0.2263	0.0376	0.2173	0.2213	0.2152
GA	0.2270	0.0000	0.0263	0.1626	0.0515	0.0955	0.0874
NG	0.2630	0.0133	0.0000	0.1954	0.0719	0.1128	0.1067
RAL	0.0444	0.1672	0.2000	0.0000	0.1879	0.1967	0.1911
RG	0.2545	0.0515	0.0631	0.1965	0.0000	0.0694	0.0595
SP	0.2565	0.0962	0.1034	0.2079	0.0768	0.0000	0.0130
ZI	0.2508	0.0939	0.1015	0.2025	0.0662	0.0127	0.0000

Supporting Material:

Figure S1. Graphical depiction of the two round assembly pipeline.

Figure S2. Length distributions for called indels for 46 bp (blue) and 150 bp (pink) read lengths. The inset zooms in on the frequencies for lengths ≥40 bp.

Figure S3. Evaluation of the tradeoff between genomic coverage and error rate for the haploid caller of the Unified Genotyper; quality values ranged from 10 to 100. Resequenced genomes from the reference strain (y^1 cn^1 bw^1 sp^1) were modified to simulate realistic levels of variation. We chose a cutoff of Q75 (red) to maximize coverage and minimize error.

Figure S4. Evaluation of the tradeoff between genomic coverage and error rate for the diploid caller of the Unified Genotyper; quality values ranged from 10 to 100. Resequenced genomes from the reference strain (y^1 cn^1 bw^1 sp^1) were modified to simulate realistic levels of variation. We chose a cutoff of Q32 (red) to maximize coverage and minimize error.

Figure S5. Histogram of sequence tract lengths for sites added by our two round pipeline for the Rwandan genomes RG5 and RG33.

Figure S6. Number of sites added by our two round assembly pipeline in 100 kb windows across the 5 euchromatic chromosome arms for the Rwandan genomes RG5 and RG33.

Figure S7. Variation among African populations in estimated cosmopolitan admixture proportions.

Figure S8. Numbers of sub-Saharan genomes inferred to have cosmopolitan ancestry in each genomic window: (A) For the DPGP3 Zambia ZI sample, and (B) across all other sub-Saharan populations. Windows are depicted for arms X (green), 2L (blue), 2R (purple), 3L (red), and 3R (orange).

Table S1. Individual sequenced genomes included in this data release, including fly stock ID, genomic library ID and type/source for each library, NIH SRA access numbers, focal chromosome arms, read length, coverage and mean depth of focal chromosome arms, and the data set from which the original sequenced reads originated.

Table S2. Population samples from which the sequenced genomes originated. The number of sequenced individuals for each focal chromosome arm is given.

Table S3. Coordinates of residual heterozygosity tracts and pseudoheterozygosity tracts filtered from genomes, and proportions of true heterozygosity and total masked heterozygosity for every genome in the data set. The distinction between the masked proportion and true heterozygosity proportion is due to the presence of artefactual heterozygosity (pseudoheterozygosity) resulting from mismapping or technical issues with individual libraries.

Table S4. Recurrent identity-by-descent (IBD) tracts for the each data set. Only IBD tracts outside of these regions were allowed to contribute to individual totals.

Table S5. Inversions detected from fly stocks via cytology (DGRP; Huang et al. 2014) or from genomes via bioinformatics (Corbett-Detig and Hartl 2012). Note that in the case of the haploid embryo genomes, live stocks may harbor undetected inversion polymorphism. "INV/ST" indicates known polymorphism. "INV/?" indicates that inverted reads were detected, but the genome was heterozygous in this region. Blank cells indicate inversions that were untested or unreported for this genome/stock.

Table S6. Regions of IBD masked from the analyzed genomes, including both individual genomes identified for each tract. See the methods for a detailed description of IBD detection and filtering criteria, and Table S4 for the excluded recurrent IBD regions.

Table S7. Inversion polymorphism and proportion heterozygosity on each focal chromosome arm for each DGRP genome, illustrating the role of inversions in maintaining heterozygosity in spite of considerable inbreeding effort. "Pseudoheterozygosity" corresponds to the proportion of a chromosome arm prior to normalization, and "Corrected heterozygosity" corresponds to the proportion of a chromosome arm following normalization.

Table S8. Regions of cosmopolitan admixture masked in Sub-Saharan African genomes.

Table S9. Genome-wide genetic differentiation and nucleotide diversity for populations with multiple high-coverage focal chromosomes, averaged across the five focal chromosome arms. Values below the diagonal are F_{ST}, values above the diagonal are D_{xy}, and bold values on the diagonal are nucleotide diversity. Distance from the *D. melanogaster* reference genome is given in the bottom row.

Table S10. Individual chromosome arm measures of genetic differentiation and nucleotide diversity for populations with at least two high-coverage sequences. Values below the diagonal are F_{ST}, values above the diagonal are D_{xy}, and bold values on the diagonal are nucleotide diversity. Distance (D_{xy}) from the *D. melanogaster* reference genome is given in the bottom row.

Figure 1

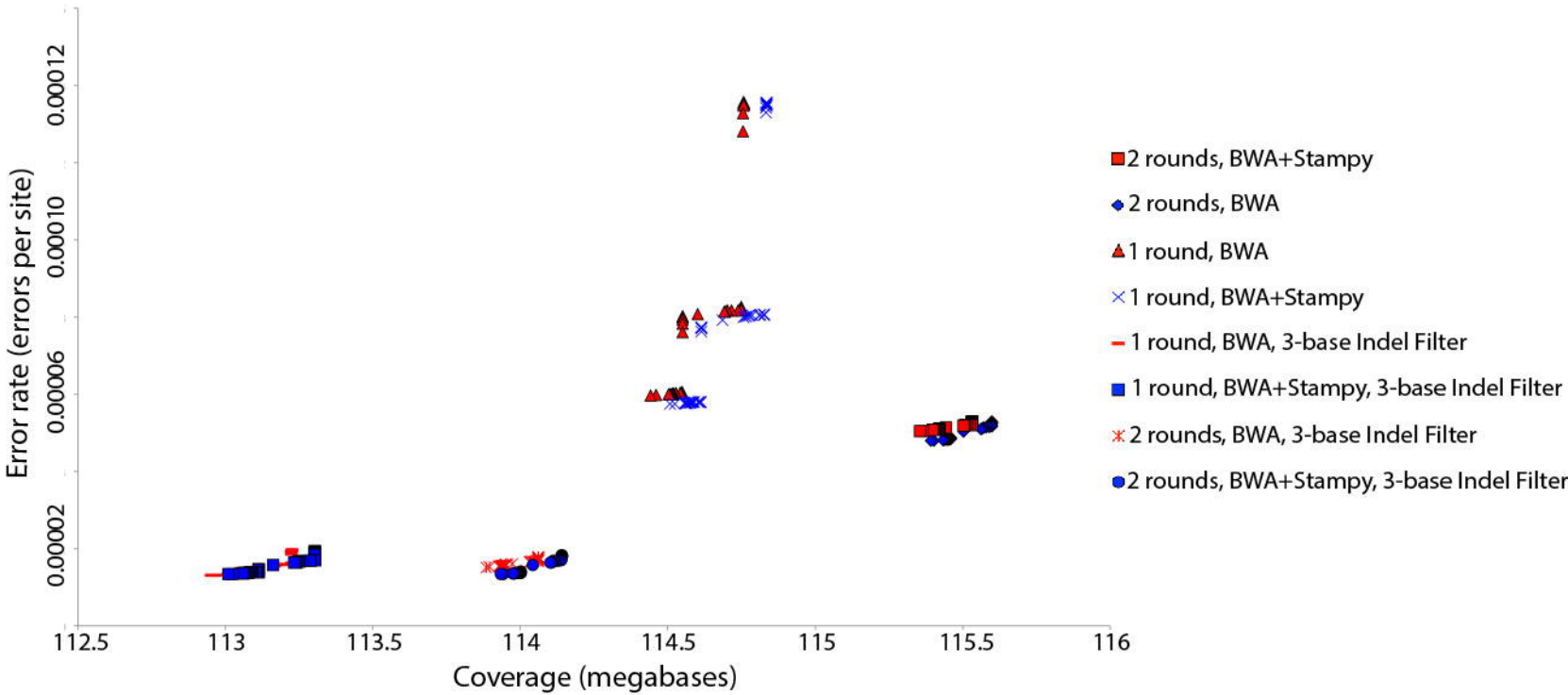

Error rate (errors per site)
0.00012
0.00010
0.00006
0.00002
112.5
113
113.5
114
114.5
115
115.5
116
Coverage (megabases)
2 rounds, BWA+Stampy
2 rounds, BWA
1 round, BWA
1 round, BWA+Stampy
1 round, BWA, 3-base Indel Filter
1 round, BWA+Stampy, 3-base Indel Filter
2 rounds, BWA, 3-base Indel Filter
2 rounds, BWA+Stampy, 3-base Indel Filter

Figure 2

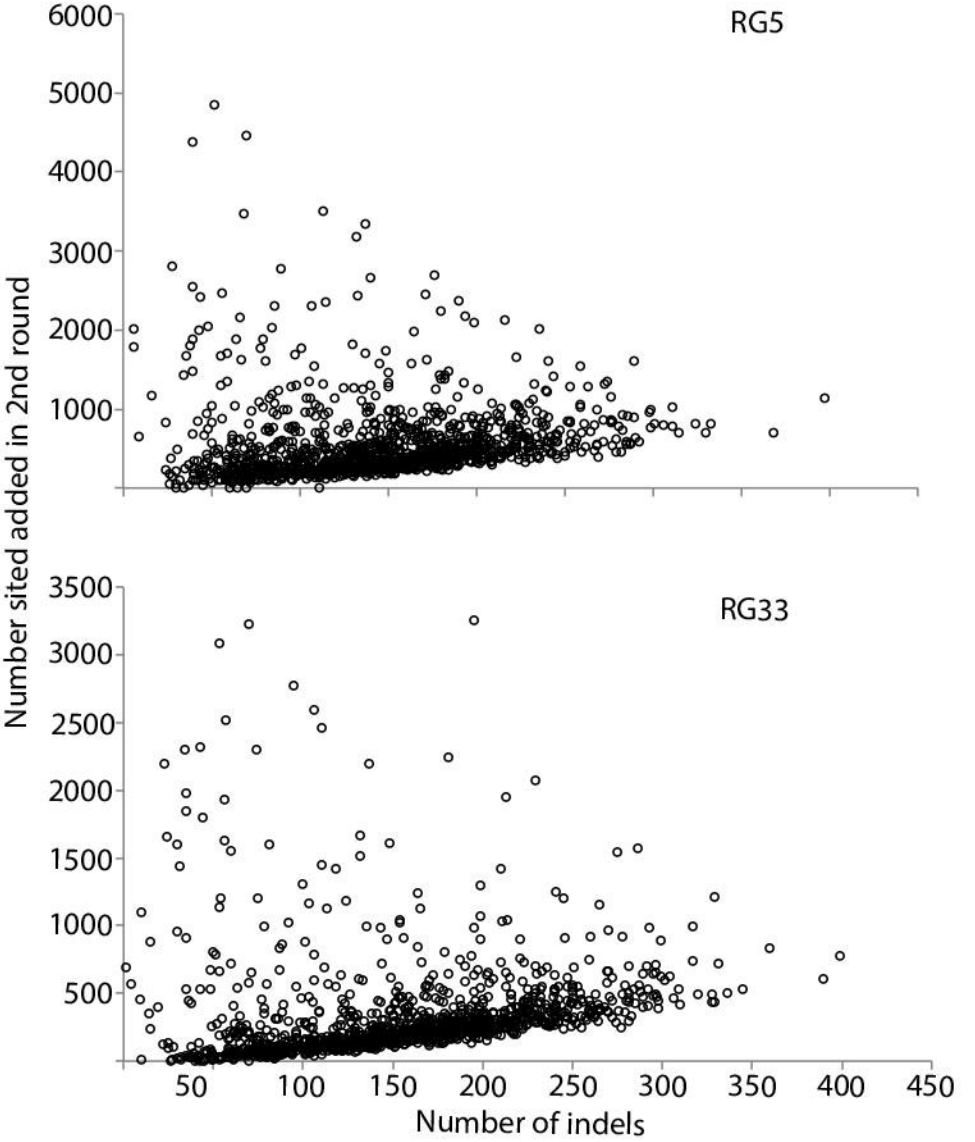

RG5
RG33
Number sited added in 2nd round
Number of indels
6000
5000
4000
3000
2000
1000
3500
3000
2500
2000
1500
1000
500
50
100
150
200
250
300
350
400
450

Figure 3

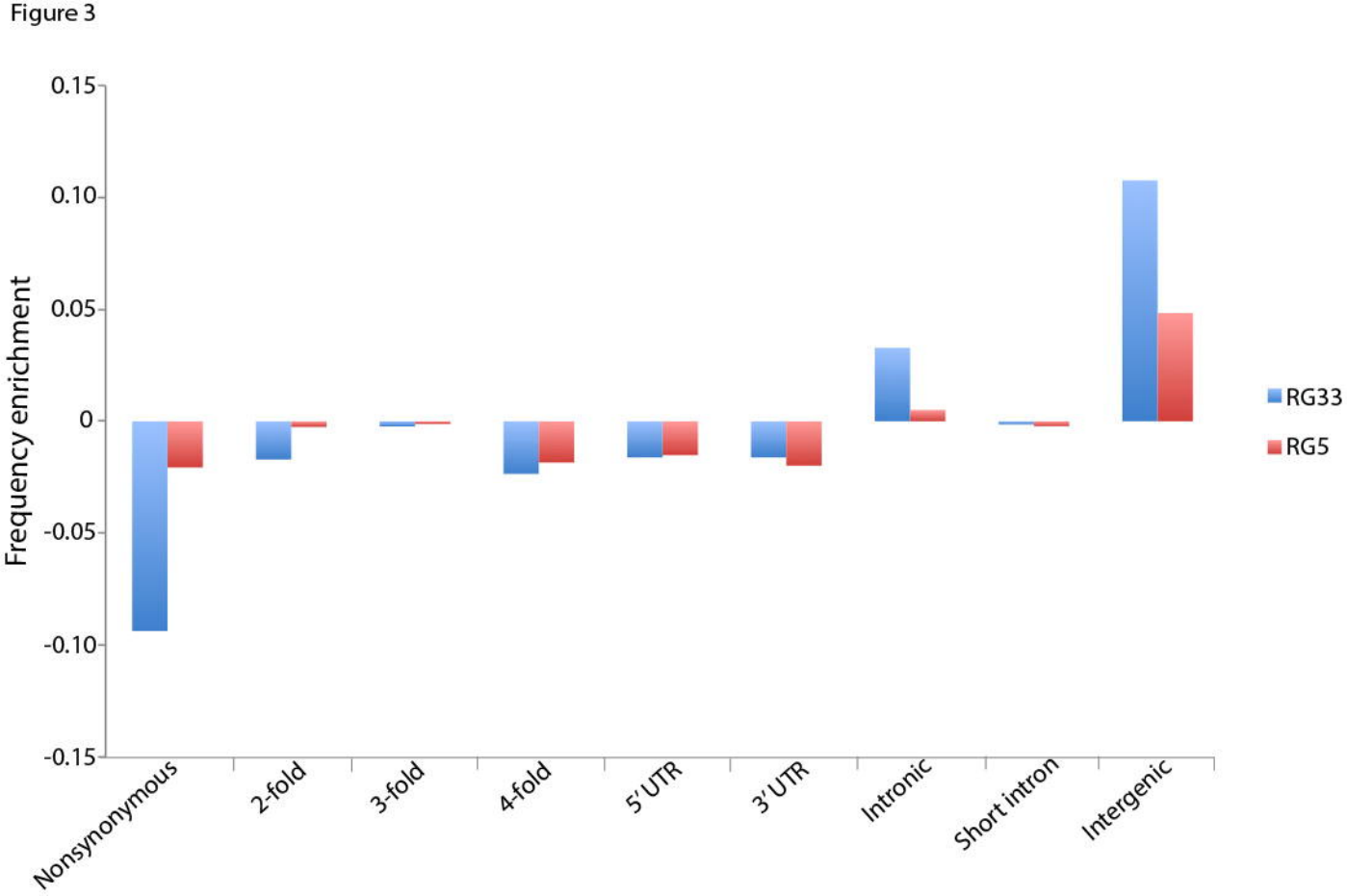

0.15
0.10
0.05
0
-0.05
-0.10
-0.15
Frequency enrichment
RG33
RG5
Nonsynonymous
2-fold
3-fold
4-fold
5' UTR
3' UTR
Intronic
Short intron
Intergenic

Figure 4

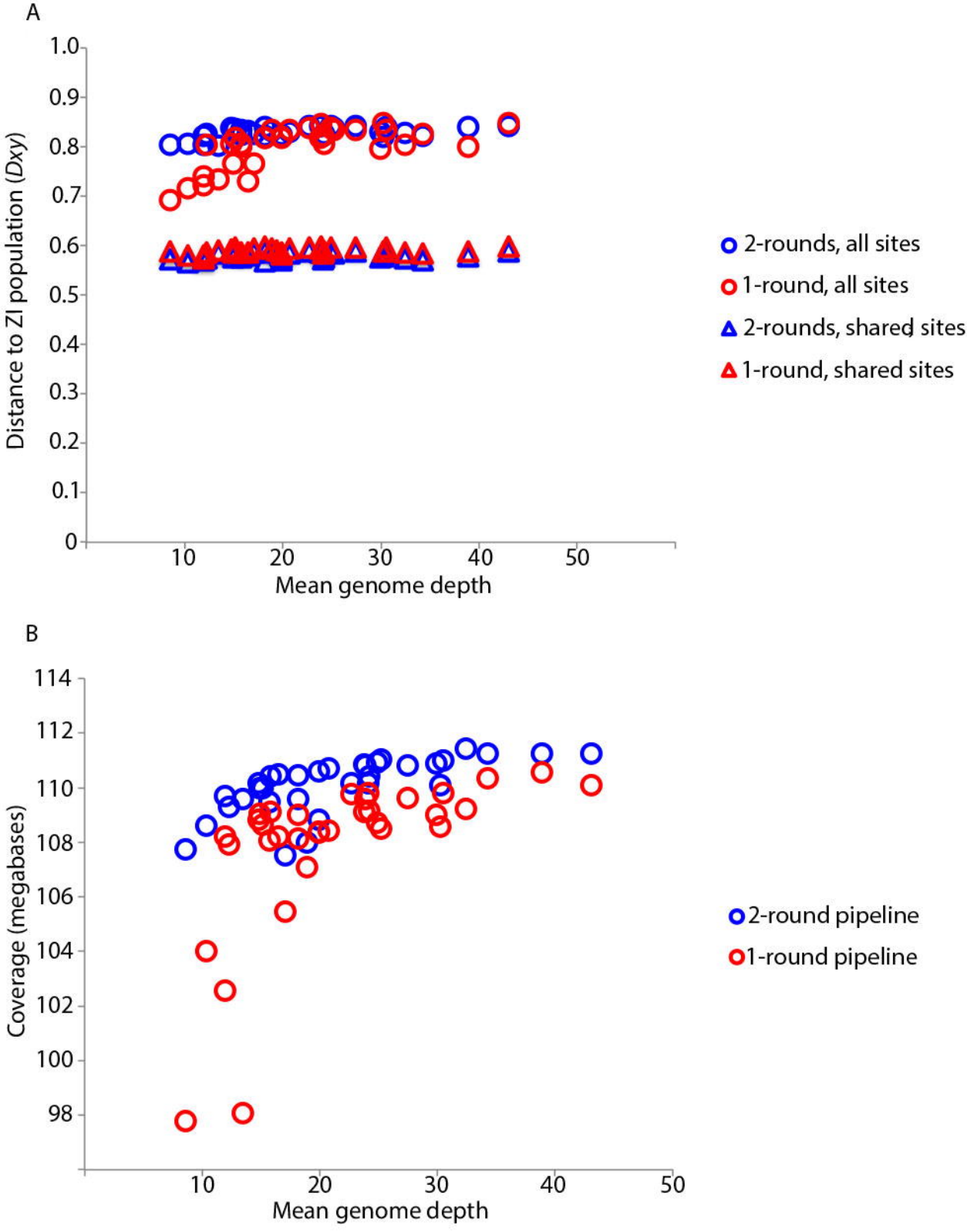

A
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Distance to ZI population (Dxy)
10
20
30
40
50
Mean genome depth
2-rounds, all sites
1-round, all sites
2-rounds, shared sites
1-round, shared sites
B
114
112
110
108
106
104
102
100
98
Coverage (megabases)
10
20
30
40
50
Mean genome depth
2-round pipeline
1-round pipeline

Figure 5

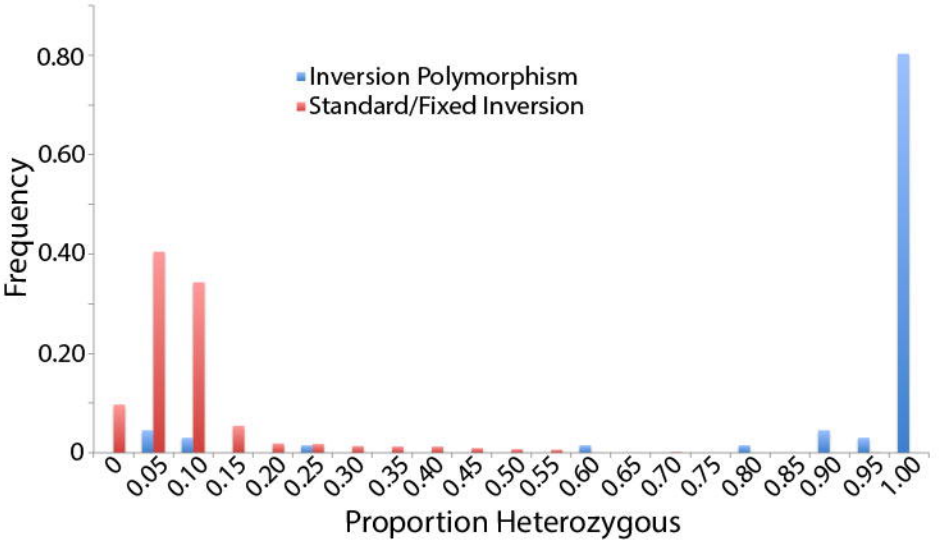
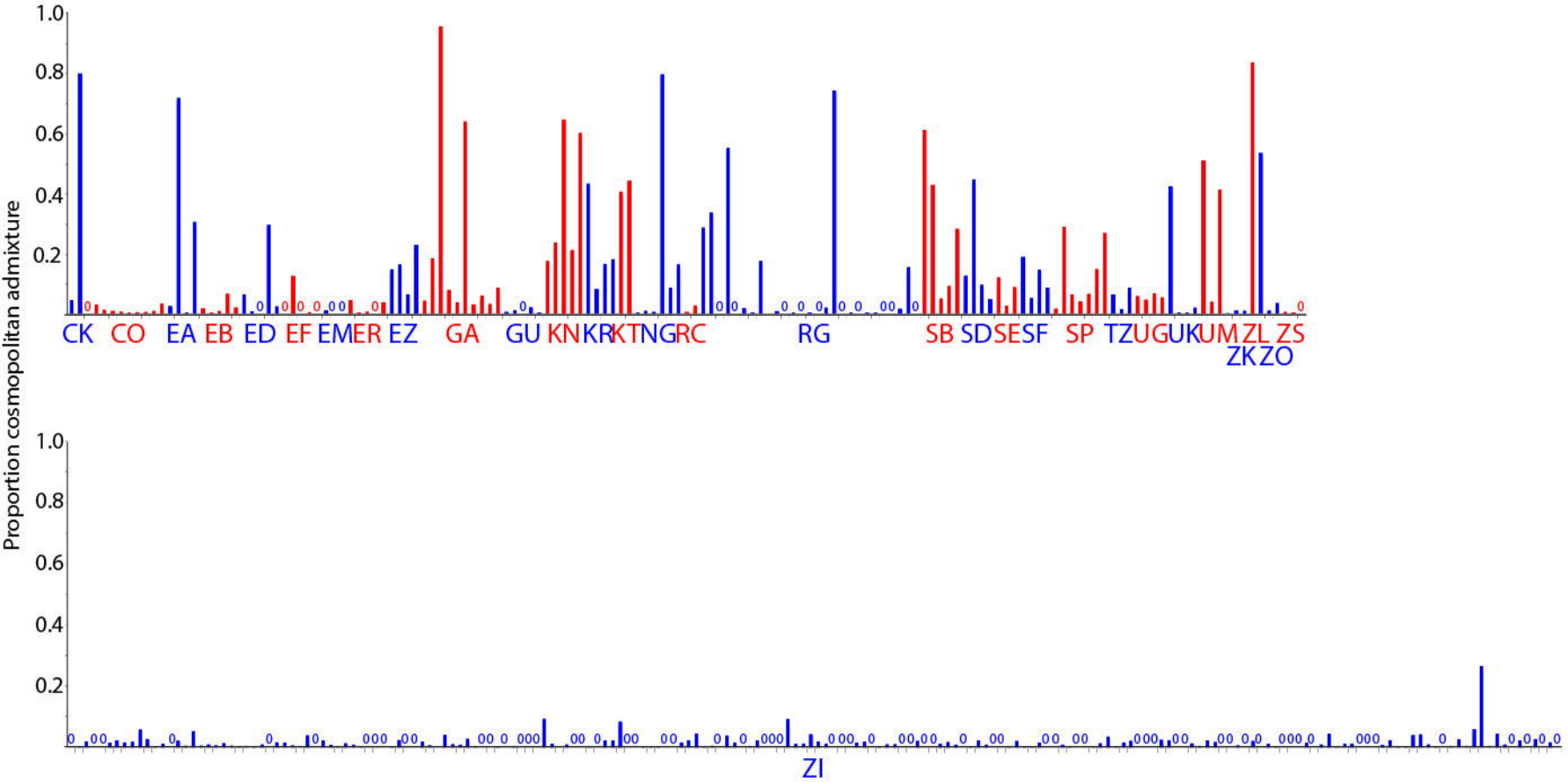

Inversion Polymorphism
Standard/Fixed Inversion
Frequency
0.80
0.60
0.40
0.20
0
0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Proportion Heterozygous

Figure 6


Proportion cosmopolitan admixture
1.0
0.8
0.6
0.4
0.2
CK CO EA EB ED EF EM ER EZ GA GU KN KR KT NG RC RG SB SD SE SF SP TZ UG UK UM ZL ZS ZK ZO
ZI