A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus

Justin B. Lack,†,1 Jeremy D. Lange,1 Alison D. Tang,2 Russell B. Corbett-Detig,†,2 and John E. Pool*†,1

1Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI
2Department of Integrative Biology, University of California, Berkeley, CA

†Present address: Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD
‡Present address: Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA

*Corresponding author: E-mail: jpool@wisc.edu.
Associate editor: Michael Rosenberg

Abstract

The Drosophila Genome Nexus is a population genomic resource that provides D. melanogaster genomes from multiple sources. To facilitate comparisons across data sets, genomes are aligned using a common reference alignment pipeline which involves two rounds of mapping. Regions of residual heterozygosity, identity-by-descent, and recent population admixture are annotated to enable data filtering based on the user’s needs. Here, we present a significant expansion of the Drosophila Genome Nexus, which brings the current data object to a total of 1,121 wild-derived genomes. New additions include 305 previously unpublished genomes from inbred lines representing six population samples in Egypt, Ethiopia, France, and South Africa, along with another 193 genomes added from recently-published data sets. We also provide an aligned D. simulans genome to facilitate divergence comparisons. This improved resource will broaden the range of population genomic questions that can be addressed from multi-population allele frequencies and haplotypes in this model species. The larger set of genomes will also enhance the discovery of functionally relevant natural variation that exists within and between populations.

Key words: Drosophila melanogaster, genomic variation, reference alignment, data resource.

Introduction

The genetics model Drosophila melanogaster has played a pivotal role in population genetic research. A growing number of studies have generated population genomic data from this species, but alignment and filtering criteria typically vary among studies, which obscures direct comparisons between these data sets. The Drosophila Genome Nexus (DGN; Lack et al. 2015; http://www.johnpool.net/genomes.html; last accessed September 20, 2016) provides the research community with genomes from multiple published sources that are generated using a common reference alignment pipeline. This more consistent data object is intended to facilitate comparisons of genomic variation between data sets with less potential for methodological bias. The DGN pipeline improved upon typical reference alignment protocols by including a second round of mapping to a modified reference genome that incorporates the variants detected in the first round, a practice that resulted in improved genomic coverage and accuracy (Lack et al. 2015).

Version 1.0 of the DGN included 623 genomes of D. melanogaster from individual wild-derived strains, originating from five data sets (table 1). Phase 2 of the Drosophila Population Genomics Project (DPGP; Pool et al. 2012) included 139 genomes from 22 populations, mainly from Africa. D. melanogaster was known to have originated in sub-Saharan Africa (Lachaise et al. 1988), and this study identified southern-central Africa as the likely ancestral range. It also identified significant recent gene flow re-entering Africa, potentially related to urban adaptation, and powerful effects of inversions on genomic variation (Pool et al. 2012). This geographic sampling across Africa was supplemented by a set of genomes denoted in the first DGN publication as AGES (African Genomes Extended Sequencing; Lack et al. 2015). Phase 3 of DPGP focused on a putative ancestral range population identified in the previous study, and brought this Zambia sample to a total of 197 independent, haploid genomes from a single location (Lack et al. 2015). That study, which also introduced the DGN, confirmed that the focal Zambia sample was maximally diverse among all sampled populations, with minimal presence of non-African admixture (Lack et al. 2015). Most of the DPGP2 genomes and all of the DPGP3 and AGES genomes were sequenced from haploid embryos (Langley et al. 2011). Most other DGN genomes were sequenced from inbred or isofemale lines (Supplementary tables S1 and S2, Supplementary Material online).

Another data source for DGN 1.0 was from the Drosophila Genetic Reference Panel (DGRP), which consists of 205 genomes originating from Raleigh, North Carolina, USA (Mackay et al. 2012; Huang et al. 2014), and has been widely used in genome-wide association studies. These genomes were from strains inbred for 20 generations, resulting in 87% homozygous regions across euchromatic chromosome arms (Lack et al. 2015). North American populations appear...
to have resulted from admixture between European and African gene pools; a recent study that examined population ancestry along DGRP genomes estimated this population to be 20% African, with significant genome-wide evidence for incompatibilities between African and European alleles at unlinked loci (Pool et al. 2015). Beyond the above data sources, DGN 1.0 also included Malawi chromosome extraction line genomes from DPGP Phase 1 (Langley et al. 2012), which are grouped with DPGP2 genomes in the DGN. And it featured source strain genomes from the Drosophila Synthetic Population Resource (DSPR; King et al. 2012), a trait mapping resource that encompasses more than 1,700 recombinant inbred lines.

In the present release, labeled as version 1.1 of the DGN, we add a total of 498 genomes. Of these, 305 are newly published in this study, and were sequenced from strains inbred for at least 15 sequenced genomes (as in DGN 1.0), 14 population samples now fit this criterion, with five of these having more than 60 genomes (fig. 1).

Importantly, the genomic alignments present in the prior DGN release have not been altered in version 1.1. Instead, we have supplemented the existing data resource by aligning and filtering the additional genomes using exactly the same pipeline described for DGN 1.0, again using the Flybase release 5.57 D. melanogaster reference genome (Lack et al. 2015). Beginning with raw sequence read data, mapping is performed using BWA v0.5.9 (Li and Durbin 2010) followed by Stampy v1.0.20 (Lunter and Goodson 2010). GATK (DePristo et al. 2011) is then used to realign indels and generate consensus sequences. Called SNPs and indels are then incorporated into a genome-specific modified reference sequence, and read mapping is performed a second time to reduce mismatches. Genomic coordinates are then shifted back to match the original reference numbering. The “site” and indel variant call files (VCFs) provided by DGN are the direct output of this pipeline.

Drosophila Genome Nexus (DGN) also distributes consensus sequence files that feature additional filtering, and may be more appropriate for most analyses. To reduce the error rate, sites within 30 bp of a called indel are masked to “N”. For genomes that may contain residual heterozygosity, genomic intervals of apparent heterozygosity are fully masked. For fully haploid genomes (Langley et al. 2011), sites with an excess of apparent heterozygosity (e.g., due to technical artifacts or structural variation) are similarly masked as “pseudoheterozygosity”. Following such masking (in addition to removal of non-target chromosome arms from samples such as chromosome extraction line genomes), we find that an average site has homozygous consensus sequence calls from 754 DGN genomes.

We also provide files to enable user-initiated masking for two additional criteria. First, we allow regions of “identity by descent” due to relatedness between genomes in the same population sample to be masked. Second, we allow users to mask from sub-Saharan genomes regions of recent admixture from non-African populations (Pool et al. 2012). Full details on the alignment and filtering processes are given by

---

**Table 1. Genomic Data Sets Present in the Drosophila Genome Nexus Are Summarized.**

<table>
<thead>
<tr>
<th>DGN Set</th>
<th>Data reference</th>
<th>Genomes</th>
<th>Populations</th>
<th>Geographic focus</th>
<th>Genome type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Present in DGN 1.0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGRP</td>
<td>Mackay et al. 2012; Huang et al. 2014</td>
<td>205</td>
<td>1</td>
<td>North America</td>
<td>Inbred</td>
</tr>
<tr>
<td>DPGP3</td>
<td>Lack et al. 2015</td>
<td>197</td>
<td>1</td>
<td>Africa (Zambia)</td>
<td>Haploid</td>
</tr>
<tr>
<td>DPGP2</td>
<td>Pool et al. 2012; Langley et al. 2012</td>
<td>150</td>
<td>22</td>
<td>Mostly Africa</td>
<td>Mostly Haploid</td>
</tr>
<tr>
<td>AGES</td>
<td>Lack et al. 2015</td>
<td>53</td>
<td>12</td>
<td>Africa</td>
<td>Haploid</td>
</tr>
<tr>
<td>DSPR</td>
<td>King et al. 2012</td>
<td>18</td>
<td>18</td>
<td>Worldwide</td>
<td>Inbred</td>
</tr>
<tr>
<td><strong>Added in DGN 1.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POOL</td>
<td>(present study)</td>
<td>305</td>
<td>6</td>
<td>Africa/Europe</td>
<td>Inbred</td>
</tr>
<tr>
<td>CLARK</td>
<td>Grenier et al. 2015</td>
<td>85</td>
<td>7</td>
<td>Worldwide</td>
<td>Inbred</td>
</tr>
<tr>
<td>NLUZHDIN</td>
<td>Campo et al. 2013; Kao et al. 2015</td>
<td>58</td>
<td>13</td>
<td>North America</td>
<td>Inbred</td>
</tr>
<tr>
<td>BERGMAN</td>
<td>Bergman and Haddrill 2015</td>
<td>50</td>
<td>3</td>
<td>Africa/Eur./N. Am.</td>
<td>Isofemale</td>
</tr>
</tbody>
</table>

**Note.**—Further details concerning the population samples and individual genomes represented in these data sets are given in Table S1 and Table S2, respectively.
Lack et al. (2015). Detailed filtering outcomes for heterozygosity, relatedness IBD, and admixture are provided in supplementary tables S3–S5, Supplementary Material online, respectively. Users can also deploy a script to extract FastA alignments for specific genomic regions from downloaded data.

**FIG. 1.** Geographic locations of selected population samples are shown, with the largest samples in bold print. These populations have at least three sequenced genomes with DGN consensus sequences available.
Filtering characteristics of several data sets are depicted in figure 2. Substantial heterozygosity persists in genomes sequenced from inbred lines (GDL, Campo, Kao, Pool, DGRP), in spite of inbreeding efforts that would be expected to reduce heterozygosity to nominal levels under neutral assumptions. Note that in figure 2, "heterozygosity" also includes regions masked due to elevated heterozygous site rates for reasons such as copy number variation or data quality ("pseudoheterozygosity"; Lack et al. 2015). For example, the DGRP data set is estimated to have just 13% genuine heterozygosity (Lack et al. 2015). Previous analysis has shown that most genuine residual heterozygosity is associated with inversions (Grenier et al. 2015; Lack et al. 2015). Inversion genotypes based on prior published calls and the method of Corbett-Detig et al. (2012) are given in supplementary table S2, Supplementary Material online. Genomes from the Bergman and Haddrill (2015) data set, which were sequenced from isofemale lines, were estimated to be 99% heterozygous. DGN provides VCFs but not heterozygosity-filtered consensus sequences for these genomes.

Figure 2 also shows the proportion of data sets that can be masked for relatedness IBD. These IBD tracts can allow the estimation of an average coefficient of relationship (Wright 1922) for each population sample, which may be viewed as the probability that two random genomes are IBD at a given site due to recent relatedness. Focusing on population samples with at least 15 genomes, we estimate that for most population samples, a random pair of individuals has a coefficient of relatedness between 0.001 and 0.005 (supplementary table S4, Supplementary Material online), or roughly the relatedness of fourth cousins. A few populations have lower

**FIG. 2.** The extent of genomic data annotated for masking due to heterozygosity, relatedness, and admixture is shown per 119 Mb genome (filtered in that order).

**FIG. 3.** Average values of nucleotide diversity ($\pi$) within populations (on the diagonal), average pairwise distance between populations ($D_{xy}$ above the diagonal), and $F_{ST}$ between populations (below the diagonal) are shown. Values are averaged across chromosome arms X, 2L, 2R, 3L, and 3R, each of which was analyzed using inversion-free genomes only.
values (EG, RG, SP, and the DPGP3 Zambia Z1 sample). Relatedness in one population (Netherlands N) was an order of magnitude higher than any other; its coefficient of relationship (0.046) exceeded the expectation for second cousins. Thus, it may be important to account for relatedness IBD (e.g., by masking the provided intervals) if analysis will assume that unrelated alleles are being compared.

Pool et al. (2012) found evidence for substantial recent gene flow from non-African populations back into sub-Saharan genomes. Masking admixed genomic regions may allow sub-Saharan genetic diversity to be studied more directly, with fewer departures from typical assumptions of well-mixed populations. Admixture levels are known to vary drastically between sub-Saharan populations, partly as a function of urbanization (Pool et al. 2012). Of the data sets shown in figure 2, “Pool” is mostly comprised of sub-Saharan genomes (62% from Ethiopia or South Africa), whereas one sixth of “GDL” consists of Zimbabwe genomes. “DPGP3” is a sample of 197 genomes from a single Zambia population with very low levels of admixture (Lack et al. 2015).

Among the DGN 1.1 samples, 15 worldwide populations are represented by at least 10 genomes for all three euchromatic chromosomes. A summary of genetic variation within and between these populations is provided in figure 3. As previously indicated, genomic diversity is highest in Zambia and other southern African populations (Pool et al. 2012; Lack et al. 2015), and all sub-Saharan populations are more diverse than all others. Because North American populations have mainly European but partly African ancestry (Kao et al. 2015; Pool 2015; Bergland et al. 2016), they show somewhat higher diversity than European populations. Geographic structure is apparent, especially between sub-Saharan populations and all others, with the latter group showing a common reduced gene pool apparently resulting from a population bottleneck. Additional bottlenecks may have impacted the B population from China (Laurent et al. 2011) and the EF population from the Ethiopian highlands (Pool et al. 2012; Lack et al. 2015), leading to mild population-specific reductions in diversity and increases in genetic differentiation (fig. 3).

In addition to the above-described D. melanogaster genomes, DGN now also distributes an aligned sequence of D. simulans to the same D. melanogaster reference genome. Stanley and Kulathinal (2016) produced this alignment using progressiveMAUVE (Darling et al. 2010) to align the release 2 D. simulans genome (Hu et al. 2013) to the release 5 D. melanogaster reference sequence. We provide sequence text files mirroring our D. melanogaster consensus sequences for D. simulans on the DGN web site (http://www.johnpool.net/genomes.html; last accessed September 20, 2016). Note that for all data hosted by DGN, users should cite the original publications (supplementary table S2, Supplementary Material online) in addition to this alignment resource.

This expansion of the DGN will significantly bolster researchers’ ability to examine genetic variation within and between D. melanogaster populations. Future DGN releases will entail realigning all genomes using updated methods and reference genomes, plus evaluating new formats for providing genomic data. Community input to shape the future of this population genomic resource is welcome.

**Supplementary Material**

Supplementary tables S1–S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The UW-Madison Center for High Throughput Computing provided computational assistance and resources for this work. This research was funded by NIH grants R01 GM111797 to JEP and F32 GM106594 to JBL, and by USDA Hatch grant WIS01900 to JEP.

**References**


3312


