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]

[1]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI

[2]Department of Integrative Biology, University of California, Berkeley, 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 New additions include 305 previously unpublished genomes from inbred lines representing six population samples in Egypt, Ethiopia, France, and South Africa, along with another 193 genomes added from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 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.**

Letter

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

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

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

NOTE.—Further details concerning the population samples and individual genomes represented in these data sets are given in Table S1 and Table S2, respectively.

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 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 eight generations. These genomes were added to much smaller samples of genomes originating from a pair of Ethiopian populations (EA, EF), a pair of South African populations (SD, SP), and populations from Egypt (EG) and France (FR). Genomic sequencing was performed using identical methods to those described by (Lack *et al.* 2015). Briefly, for each inbred line, ~30 female flies were used to prepare genomic DNA libraries. Sequencing on a HiSeq 2000 was performed to generate paired end 100 bp paired end reads with ~300 bp inserts.

Drosophila Genome Nexus (DGN) 1.1 also adds 193 genomes from four published studies. The Global Diversity Lines (GDL; Grenier *et al.* 2015) include 85 genomes from Australia, China, the Netherlands, the USA, and Zimbabwe. The 50 genomes published by Bergman and Haddrill (2015) originate from France, Ghana, and the USA. Campo *et al.* (2013) studied 35 genomes from a California population. Kao *et al.* (2015) added 23 genomes originating from 12 New World locations.

The data sets represented in DGN1.1 are summarized in table 1. The 74 population samples they encompass are described in Table S1, and many of these are depicted in figure 1. Characteristics of all 1,121 individual strain genomes are given in supplementary table S2, Supplementary Material online. Instead of just three geographic population samples with 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 3 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

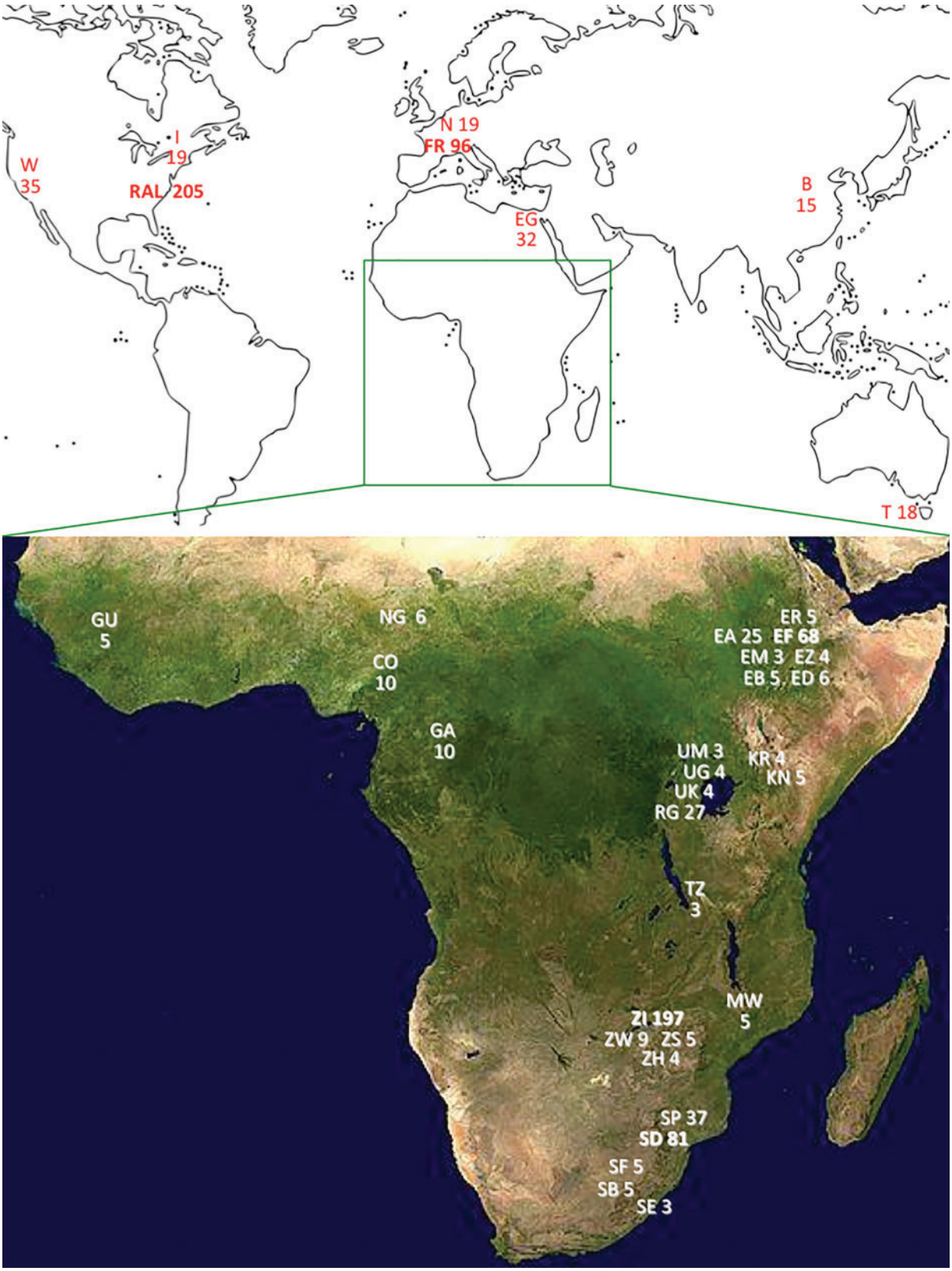

Fig. 1. Geographic locations of selected population samples are shown, with the largest samples in bold print. These populations have at least three sequenced genomes with DGN consensus sequences available.

Lack *et al.* (2015). Detailed filtering outcomes for heterozygosity, relatedness IBD, and admixture are provided in sup plementary tables S3–S5, Supplementary Material online, respectively. Users can also deploy a script to extract FastA alignments for specific genomic regions from downloaded data.

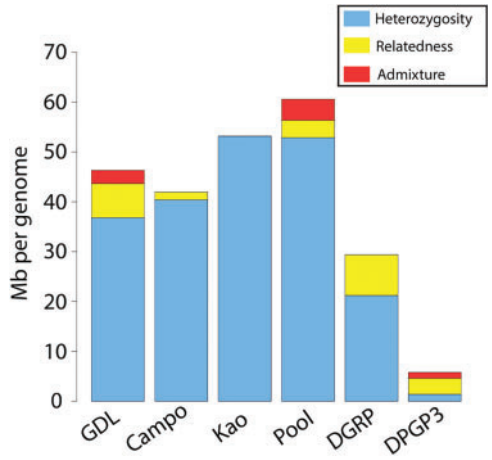

FIG. 2. The extent of genomic data annotated for masking due to heterozygosity, relatedness, and admixture is shown per 119 Mb genome (filtered in that order).

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 ("psuedoheterozygosity"; 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

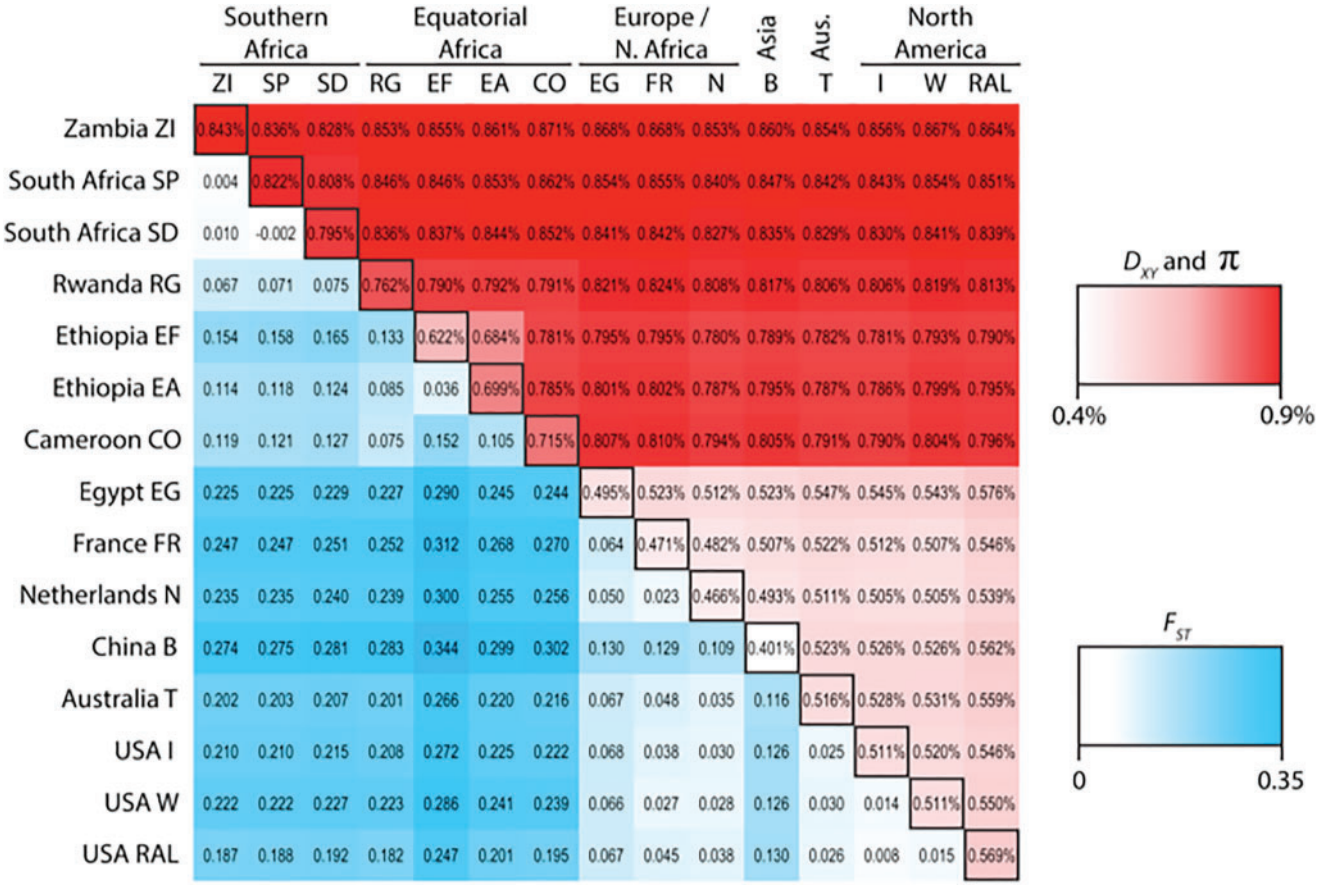

FIG. 3. Average values of nucleotide diversity (π) within populations (on the diagonal), average pairwise distance between populations (D_{xy}, above the diagnonal), and F_{ST} between populations (below the diagonal) are shown. Values are averaged across chromosome arms X, 2L, 2R, 3L, and 3R, each of which was analyzed using inversion-free genomes only.

values (EG, RG, SP, and the DPGP3 Zambia ZI 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

Bergland AO, Tobler R, Gonzalez J, Schmidt P, Petrov D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *Mol. Ecol.* 25:1157–1174.

Bergman CM, Haddrill PR. 2015. Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. *F1000Research* 4:31.

Campo D, Lehmann K, Fjeldsted C, Souaiaia T, Kao J, Nuzhdin SV. 2013. Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. *Mol. Ecol.* 22:5084–5097.

Corbett-Detig RB, Cardeno C, Langley CH. 2012. Sequence-based detection and breakpoint assembly of polymorphic inversions. *Genetics* 192:131–137.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variant discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491–498.

Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG. 2015. Global diversity lines—a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3 (Bethesda)* 5:593–603.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Res.* 23:89–98.

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* 24:1193–1208.

Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D. 2015. Population genomic analysis uncovers African and European admixture in *Drosophila melanogaster* populations from the south-eastern United States and Caribbean Islands. *Mol. Ecol.* 24:1499–1509.

King EG, Macdonald SJ, Long AD. 2012. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics* 191:935–949.

Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L, Ashburner M. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* 22:159–225.

Lack JL, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The *Drosophila* Genome Nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* 199:1229–1241.

Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K. 2011. Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics* 188:239–246.

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192:533–598.

Laurent SJ, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. *Mol. Biol. Evol*. 28:2041–2051.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.

Lunter G, Goodson M. 2010. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 18:821–829.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, McGwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482:173–178.

Pool JE. 2015. Natural selection shapes the mosaic ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* reference genome. *Mol. Biol. Evol*. 32:3236–3251.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population genomics of Sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet*. 8: e1003080.

Stanley CE, Kulathinal RJ. 2016. Genomic signatures of domestication on neurogenetic genes in *Drosophila melanogaster*. *BMC Evol. Biol*. 16:6.

Wright S. 1922. Coefficients of inbreeding and relationship. *Am. Nat*. 56:330–338.