A Genomic Reference Panel for *Drosophila serrata*

Adam J. Reddiex, Scott L. Allen, and Stephen F. Chenoweth[1]
School of Biological Sciences, The University of Queensland, QLD 4072, Australia
ORCID ID: 0000-0002-8303-9159 (S.F.C.)

ABSTRACT Here we describe a collection of re-sequenced inbred lines of *Drosophila serrata*, sampled from a natural population situated deep within the species endemic distribution in Brisbane, Australia. *D. serrata* is a member of the speciose *montium* group whose members inhabit much of south east Asia and has been well studied for aspects of climatic adaptation, sexual selection, sexual dimorphism, and mate recognition. We sequenced 110 lines that were inbred via 17-20 generations of full-sib mating at an average coverage of 23.5x with paired-end Illumina reads. 15,228,692 biallelic SNPs passed quality control after being called using the Joint Genotyper for Inbred Lines (JGIL). Inbreeding was highly effective and the average levels of residual heterozygosity (0.86%) were well below theoretical expectations. As expected, linkage disequilibrium decayed rapidly, with r^2 dropping below 0.1 within 100 base pairs. With the exception of four closely related pairs of lines which may have been due to technical errors, there was no statistical support for population substructure. Consistent with other endemic populations of other *Drosophila* species, preliminary population genetic analyses revealed high nucleotide diversity and, on average, negative Tajima's D values. A preliminary GWAS was performed on a cuticular hydrocarbon trait, 2-Me-C$_{28}$ revealing 4 SNPs passing Bonferroni significance residing in or near genes. One gene *Cht9* may be involved in the transport of CHCs from the site of production (oenocytes) to the cuticle. Our panel will facilitate broader population genomic and quantitative genetic studies of this species and serve as an important complement to existing *D. melanogaster* panels that can be used to test for the conservation of genetic architectures across the *Drosophila* genus.

The availability of whole genome sequence data for *Drosophila* species has greatly facilitated advances in the fields of genetics and evolutionary biology. For example, the sequencing of 12 *Drosophila* genomes (Clark *et al.* 2007) was instrumental to new discoveries in comparative genomics (Stark *et al.* 2007; Sturgill *et al.* 2007; Zhang *et al.* 2007). The advent of affordable genome sequencing has also allowed population geneticists to characterize genomic variation within and among natural populations, improving our understanding of the complex evolutionary histories of cosmopolitan species such as *D. melanogaster* and *D. simulans* (Begun *et al.* 2007; Lack *et al.* 2015; Langley *et al.* 2012; Pool *et al.* 2012). Most recently, multiple panels of re-sequenced inbred *D. melanogaster* lines

have become available, facilitating the molecular dissection of complex trait variation (Grenier *et al.* 2015; Huang *et al.* 2014; King *et al.* 2012; Mackay *et al.* 2012). With these populations of reproducible genotypes, researchers have used genome-wide association analysis to identify genetic variants underlying variation in a broad range of traits including physiological traits (Burke *et al.* 2014; Dembeck *et al.* 2015; Gerken *et al.* 2015; Unckless *et al.* 2015; Weber *et al.* 2012), behaviors (Shorter *et al.* 2015), recombination rates (Hunter *et al.* 2016), disease susceptibility (Magwire *et al.* 2012), and traits related to human health (Harbison *et al.* 2013; He *et al.* 2014; King *et al.* 2014; Kislukhin *et al.* 2013; Marriage *et al.* 2014).

Just as comparative genomic and population genetic studies of adaptation (*e.g.*, Machado *et al.* 2016; Zhao *et al.* 2015) have been enhanced through the availability of multi-species genome resources, quantitative genetics may also benefit from the availability of multispecies genome panels . The development of panels of re-sequenced lines for *Drosophila* species beyond *D. melanogaster* may support broader lines of inquiry such as the conservation of genetic architectures among related taxa (Yassin *et al.* 2016). To this end, we have developed a new genomic resource for *D. serrata*, a member of the *montium* group of species. The *montium* group has long been regarded as a subgroup

within the melanogaster species group (Lemeunier *et al.* 1986), but has more recently been considered as a species group of its own (Da Lage *et al.* 2007; Yassin 2013). Although *montium* contains 98 species (Brake and Bachli 2008) and represents a significant fraction of all known *Drosophila* species, there have been very few genomic investigations of its members. Recently, genomic tools have been developed for *D. serrata* including an expressed sequence tag (EST) library (Frentiu *et al.* 2009), a physical linkage map (Stocker *et al.* 2012), and transcriptome-wide gene expression datasets (Allen *et al.* 2013; Allen *et al.* 2017a; McGuigan *et al.* 2014). Additionally, an assembled and annotated genome sequence (Allen *et al.* 2017b) make *D. serrata* only the second species in the *montium* group with a sequenced genome after *D. kikkawai* (NCBI *Drosophila kikkawai* Annotation Release 101). Coupled to this, *D. serrata* is one member of the *montium* group that has been extensively studied in the field of evolutionary genetics.

Populations of *D. serrata* have been recorded from as far north as Rabaul, Papua New Guinea (4.4°N) (Ayala 1965) to as far south as Woolongong, Australia (34.3°S) (Jenkins and Hoffmann 1999). This broad latitudinal range has made *D. serrata* an ideal model for population studies addressing the evolution of species borders (Blows and Hoffmann 1993; Hallas *et al.* 2002; Magiafoglou *et al.* 2002; van Heerwaarden *et al.* 2009) and adaptation along latitudinal clines (Allen *et al.* 2017a; Frentiu and Chenoweth 2010; Kellermann *et al.* 2009). *D. serrata* has also emerged as a powerful model for the application of quantitative genetic designs to investigate sexual selection (Gosden and Chenoweth 2014; Hine *et al.* 2002; McGuigan *et al.* 2011).

Here, we the report development of a panel of 110 re-sequenced inbred *D. serrata* lines that we have called the *Drosophila serrata* Genome Reference Panel (DsGRP). Similar to the DGRP (Mackay *et al.* 2012), flies were sampled from a single large natural population with the exception that *D. serrata* was sampled from its endemic distribution. In this initial description, we estimate the degree of heterozygosity remaining in the lines after inbreeding, show the degree to which lines are genetically related to one another, estimate genome-wide levels of nucleotide diversity, and describe patterns of linkage disequilibrium. We also demonstrate how this panel of flies can be used to genetically dissect trait variation by performing a genome-wide association analysis on variation in a cuticular hydrocarbon (CHC) trait.

METHODS

Collection and inbreeding
Drosophila serrata were collected from a wild population located at Bowman Park, Brisbane Australia (Latitude: -27.45922, Longitude: 152.97768) during October 2011. We established each line from a single, gravid female before applying 20 generations of inbreeding. Inbreeding was carried out each generation by pairing virgin brothers and sisters. 100 inbred lines, out of the initial 239 iso-female lines established, survived the full 20 generations inbreeding and a further 10 lines were established after 17 generations of inbreeding.

Sequencing
We sequenced the genomes of 110 inbred lines using 100 base-pair paired-end reads with a 500 base-pair insert on an Illumina Hisequation 2000 sequencing machine. Sequencing and library preparation were carried out by the Beijing Genomics Institute. DNA from each line was isolated from a pool of at least 30 virgin female flies using a standard phenol-chloroform extraction method.

Quality control and SNP calling
We received reads from the Beijing Genomics Institute for which approximately 95% of the bases from each line had a base quality score greater than or equal to 20 (Illumina GA Pipeline v1.5). Read quality was also assessed using FastQC v0.11.2 before being mapped to the *Drosophila serrata* reference genome (Allen *et al.* 2017b) using BWA-mem v0.7.10 (Li 2013) and were realigned around indels using the GATK IndelRealigner v3.2-2 (McKenna *et al.* 2010). Genotypes for every line were inferred simultaneously using the Joint Genotyper for Inbred Lines (JGIL) v1.6 (Stone 2012). This probabilistic model was especially designed for genotyping large panels of inbred lines or strains and is considered to have high accuracy (Mackay *et al.* 2012; Stone 2012). Genotype calls with a probability lower than 99% were treated as missing genotypes.

Residual heterozygosity
The residual heterozygosity per line was estimated as the genome-wide proportion of sites that remained heterozygous after 17-20 generations of inbreeding, more specifically, we summed all of the genotype calls that were heterozygous and expressed this statistic as a percentage of all genotyped sites. In addition, for each site in the genome that differed among the inbred lines, we calculated the percentage of lines that were heterozygous for that site. Site filtering based on minor allele frequency and coverage was not performed for this analysis.

Pairwise relatedness between lines
Pairwise relatedness between lines (j and k) was estimated using the–make-grm-inbred command of GCTA v1.24.2 (Yang *et al.* 2011), which applies the expression:

$$A_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(\chi_{ij} - 2p_i)(\chi_{ik} - 2p_i)}{2p_i(1 - p_i)} \qquad [1]$$

where χ_{ij} is the number of copies of the reference allele for the ith SNP for individual j and p is the population allele frequency. N is the total number SNPs. Only biallelic SNPs with a read depth between 5 and 60 and a minor allele frequency above 5% were used to estimate relatedness.

Estimating population substructure
To test whether the sample of lines exhibited any underlying substructure, we used the approach of Bryc *et al.* (2013) which is founded on random matrix theory. Importantly, while the genomic relatedness matrix calculated in GCTA, $\mathbf{A} = \mathbf{WW'}/N$, is normalized with element $W_{ij} = (\chi_{ij} - 2p_i)/\sqrt{2p_i(1 - p_i)}$, where χ_{ij} is the number copies of the minor allele carried by individual j (Yang *et al.* 2011) and p_i is the population allele frequency, the genomic relatedness matrix used by the approach of Bryc *et al.* (2013) is not. For this approach, the genomic relatedness matrix, $\mathbf{X} = \mathbf{CC'}$, where $\mathbf{C}$ is an M x N rectangular matrix with M corresponding to the number of individuals used to estimate $\mathbf{X}$ and N is the number of SNPs, here the element $C_{ij} = \chi_{ij}$, the number of copies of the minor allele carried by individual j. $\mathbf{X}$ was scaled by equation 2.8 in Bryc *et al.* (2013) using values of M= 110 and N = 3,709,328. Only SNPs without missing data were used for this analysis. The number of sub-populations was determined by the number of eigenvalues larger than that expected for a random relatedness matrix given the significance threshold $t = (1 + F)/2$ (Bryc *et al.* 2013). Here, t corresponds to a value of 0.993 with our expected inbreeding coefficient (F) of 0.986 (Falconer and Mackay 1996) after 20 generations of full-sib mating. Upon seeing that a small number of lines were unusually highly
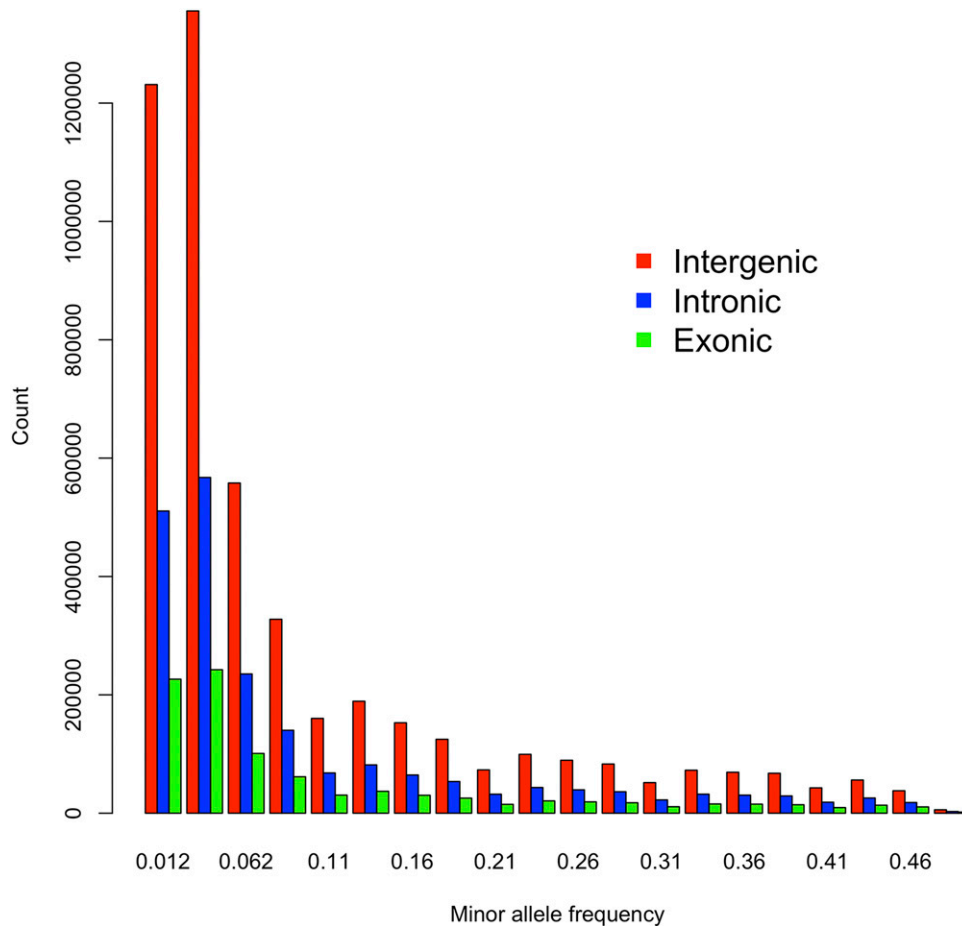
Figure 1 The allele frequency spectrum of SNPs annotated as intergenic (red), intronic (blue), or exonic (green). Singletons (MAF = 0.009) are not shown. We identified 3,748,429 singletons from intergenic regions, 1,535,651 from intronic regions, and 754,075 from exonic regions of the genome.

related (see results), we repeated this analysis after removing four lines (line IDs: 29, 134, 159, 206) to verify that the significant results were driven only by these "outliers".

Linkage disequilibrium

We estimated linkage disequilibrium as the square of the inter-variant allele count correlation (r^2) using PLINK v1.9 (Chang *et al*. 2015). We estimated r^2 in non-overlapping 500 base-pair windows across the entire genome. The analysis was performed on biallelic SNPs that had an average read depth of between 5 and 60 and a minimum minor allele frequency of 5%. The four highly related lines ($r >= 0.1$) were removed prior to the analysis.

Nucleotide diversity and neutrality

We estimated nucleotide diversity (π) and Tajima's D (Tajima 1989) statistic in 50 kilobase non-overlapping sliding windows and took the mean for each of the major chromosome arms (2L, 2R, 3L, 3R, and X) using vcftools v0.1.15 (Danecek *et al*. 2011). The SNP data used for this analysis had an average read depth across all lines of between 5 and 60 but no threshold on minor allele frequency was applied. Again, the four highly related lines were removed prior to this analysis.

Genome-wide association of female CHC expression

As proof-of-concept we performed a GWAS on a single cuticular hydrocarbon, 2-methyloctacosane, (2-Me-C_{28}). CHCs are waxy substances

that are secreted on the cuticle with 2-Me-C_{28} being one of a suite of CHCs that have been extensively studied in this species due to their role in species recognition, mate choice, and desiccation resistance. For each of 94 lines, we extracted CHCs from four virgin females, across two replicate vials using individual whole-body washes in $100\mu l$ of the solvent hexane. We used a standard gas chromatography method to quantify the amount of 2-Me-C_{28} (Blows and Allan 1998). To maintain the trait scale used in previous studies, we transformed the amount of 2-Me-C_{28} into a log-contrast value following Aitchison (1986), using an additional trait, the CHC 9-hexacosane, 9-$C_{26:1}$, as the divisor. This transformation turns the expression of CHCs into a proportional measure and provides an internal control for other sources of variation including body size and condition.

Our GWAS contrasts with other analyses performed on the DGRP in that we model trait variation at the individual, rather than line mean level. We applied a single marker mixed effects association analysis, where the following model was fit for every biallelic SNP that had mean coverage between 5 and 60 and sample MAF of 5%:

$$y_{i,j} = \mu + \text{snp} + G_i + \varepsilon_{i,j} \quad (i = 1, \ldots, 94 \; j = 1, \ldots, 4) \qquad [2]$$

Here, CHC expression (y) of replicate individual j on genotype i is modeled as a function of the mean term (μ), the additive fixed effect of the candidate SNP, the polygenic random effect that is captured by the genomic relatedness matrix (G) where G has a $\sim N(0,\sigma_A^2 A)$ distribution, and the residual error (ε). This model was specifically

designed for populations where identical genotypes can be measured independently in multiple organisms such as inbred lines (Kruijer *et al.* 2015) and was fit using AsReml-R v3 (VSN International) for a total of 3,318,503 SNPs. There are a couple of differences between the approach outlined above and other mixed modeling approaches to GWAS implemented in programs such as GEMMA (Zhou and Stephens 2012), FaST-LMM (Lippert *et al.* 2011), and GCTA (Yang *et al.* 2011). First, the use of individual-level opposed to line mean level observations, allows for estimation of the genomic heritability (Kruijer *et al.* 2015). Although mapping power is unlikely to be significantly boosted through the use of individual level data *per se* when working with inbred lines (Kruijer *et al.* 2015), a second aspect to our approach does potentially increase power. Our, albeit slower, approach re-estimates the polygenic variance component when each SNP tested and results in an exact calculation of Wald's test statistic (Zhou and Stephens 2012). This contrasts with the approach used by many mixed model GWAS programs where, to save computation time, this variance component is estimated once in a null model with no fixed effect of SNP and then held constant for each SNP tested. Such an approach produces an approximate value of the test statistic which can result in power loss under some circumstances (Zhou and Stephens 2012). As these circumstances are difficult to predict beforehand, we chose to re-estimate the polygenic random effect despite the computational cost.

To increase computational speed, we nested this linear model within an R loop that allows the access of multiple cores using the "foreach" and "doMC" packages (Revolution Analytics and Steve Weston 2015). Significant SNPs were identified as those with p-values that passed Bonferroni multiple test correction –log10(p) > 7.8, we also report SNPs with –log10(p) > 5 for comparison to other *Drosophila* GWAS where this arbitrary threshold value is used (Mackay *et al.* 2012). We took statistically significant SNPs and annotated them to the current version of the *D. serrata* genome (Allen *et al.* 2017b). If a significant SNP was located within a gene, we blasted the *D. serrata* gene sequence to the

D. melanogaster genome to determine gene orthology using Flybase (Attrill *et al.* 2016).

Data availability

Raw reads for all sequenced lines are available from the NCBI short read archive under Bioproject ID: PRJNA419238. The genomic relatedness matrices used in the population structure analysis are provided in File S1 (grm_full_Bryc.txt, grm_reduced_Bryc.txt, grm_full_gcta.txt, and grm_reduced_gcta.txt). We have provided the R code, Bryc.R, that implements the test for large eigenvalues (Bryc *et al.* 2013). The SNP list used to analyze the data in this study is available from Dryad https://doi.org/10.5061/dryad.vh0h80g and also from www.chenowethlab.org/resources.html. The CHC phenotype file is provided as the supplementary file pheno.txt. The linear model used to fit the GWAS in ASREML/R model is provided in the file asreml_gwas.R.

RESULTS AND DISCUSSION

Identification of SNPs

We established a panel of 110 inbred lines of *Drosophila serrata* from wild females caught from a single population in Brisbane Australia and sequenced their genomes. 100 base-pair paired-end reads were mapped to the *Drosophila serrata* reference genome (Allen *et al.* 2017b) with a mean coverage of 23.5 $\pm$ 0.5 reads per line. Using the Joint Genotyper for Inbred Lines (Stone 2012), we identified 15,228,692 biallelic single nucleotide polymorphisms (SNPs) applying a 99% probability threshold. 13,959,239 of these SNPs had a median coverage between 5 and 60 in which over 80% of the lines were genotyped for that variant. Most SNPs segregate at low frequencies (Figure 1) with 6,090,058 instances of singletons, where a SNP was present in only one line. The majority (62%) of the SNPs were annotated to intergenic regions of the genome while 12% of SNPs annotated to exonic regions and 26% were found to be in introns. A total of 3,709,329 SNPs met the minimum allele frequency threshold (MAF) of 5% to be used for genome-wide association analysis and the estimation of relatedness between lines.

Residual heterozygosity

Despite the application of inbreeding for many generations, inbred *Drosophila* lines often contain regions of residual heterozygosity (King *et al.* 2012; Lack *et al.* 2015; Mackay *et al.* 2012; Nuzhdin *et al.* 1997). After 17-20 generations of inbreeding, residual heterozygosity in our lines was very low and we observed only a small proportion of segregating sites within lines, suggesting that inbreeding had successfully fixed variation across these genomes. Of the 110 inbred lines, 104 had fewer than 2% segregating SNPs and 82 lines had less than 1% segregating SNPs (Figure 2). Across lines, the average proportion of segregating SNPs was 0.86% $\pm$ 0.11%, less than the theoretical expectation of 1.4% for lines that have experienced 20 generations of full-sib mating which corresponds to an expected inbreeding coefficient of F = 0.986 (Falconer and Mackay 1996). This slightly lower than expected level of residual heterozygosity may simply reflect sampling variation, be due to SNPs on the X chromosome, or may indicate purging of partially deleterious alleles during the inbreeding process (García-Dorado 2008; Garcia-Dorado 2012). There was no detectable difference in the fraction of heterozygous SNPs between the lines inbred for 17 and 20 generations (ANOVA: $F_{1,108}$ = 0.0944, P = 0.76). This result suggests that 17 generations of inbreeding may be sufficient for future line development.

Although there are several mechanisms that can inhibit the fixation of an allele within an inbred line, when short-read re-sequencing technology is used for genotyping, loci can falsely appear to be
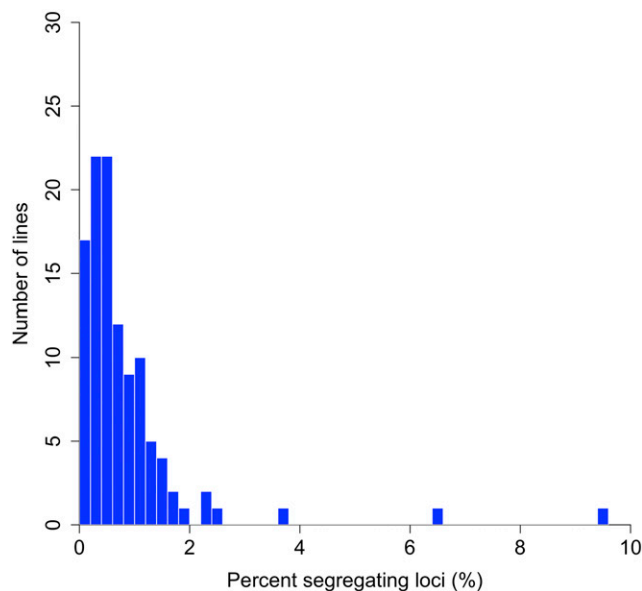

Figure 2 The distribution of residual heterozygosity as measured by the percentage of total genotyped biallelic SNPs (loci) that were called as heterozygous within each inbred line by the Joint Genotyper of Inbred Lines (JGIL).
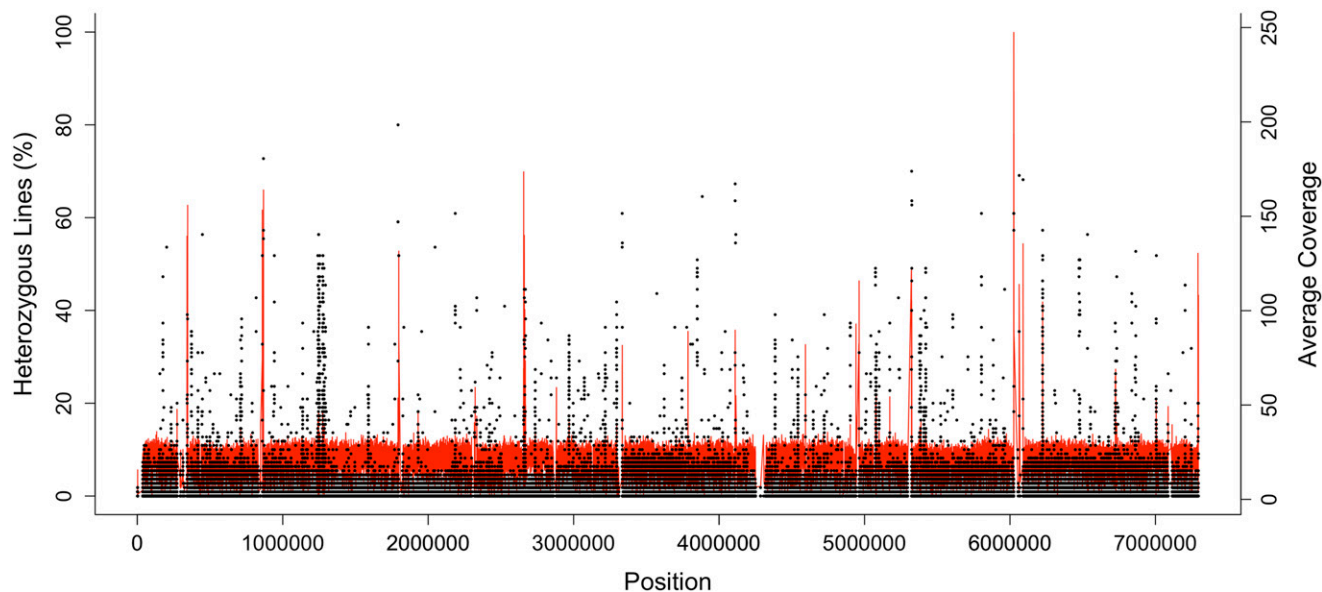
Figure 3 The proportion of lines that are heterozygous at every site along the largest scaffold (scf7180000003208) of the reference genome (black) overlayed with the average read depth at each site shown in red. Peaks in red lines represent potential gene duplications that have collapsed to the same region during assembly.

segregating due to the presence of paralogous genes or other repetitive DNA sequences (Treangen and Salzberg 2012). If paralogous genes are not represented in the reference genome, DNA sequences from the original gene and a divergent duplicate gene are mapped to the same region of the genome, causing the appearance of segregating loci in the population. When this phenomenon occurs, it is expected that regions of the genome with high "apparent heterozygosity" will associate with a higher read depth than the genome-wide average. Across the genome we found a weak, positive correlation between the level of residual heterozygosity and read depth (Spearman's $\rho = 0.036$, $P = 2.2 \times 10^{-16}$). Plots of these two factors however, clearly show an alignment of regions with both high levels of residual heterozygosity and read depth, suggesting that the *D. serrata* reference genome could be missing some duplications (Figure 3). Alternatively, this result could be due to copy number variation among the re-sequenced lines and/or between the reference genome and the 110 lines. We hope that further work and improvement of our genome for this species will elucidate these small regions of residual heterozygosity.

Relatedness between lines

Population structure and cryptic relatedness are well known to confound genetic association studies, potentially generating false positive genotype-phenotype associations (Kittles *et al.* 2002; Knowler *et al.* 1988). Conceptually, these confounding factors can be described as the unobserved pedigree of the sampled individuals caused by distant relationships (Astle and Balding 2009). The sources of these relationships are varied but include population admixture, inadvertent sampling of close relatives, and the presence of shared chromosomal inversions. Fortunately, the unobserved pedigree can be estimated using marker based approaches. Here we used such an approach to estimate the pairwise relatedness between all lines in the form of a genomic relatedness matrix.

The structure of the genomic relatedness matrix shows that the majority of the DsGRP lines are unrelated as would be expected of a sample from a large, randomly mating population (Figure 4). We found

a pair of lines that were 100% related to one another, most likely due to contamination either during the inbreeding process or DNA extraction and subsequent library preparation. Generally however, this panel of flies exhibits a lower level of relatedness compared to the DGRP, where 2.7% of the 20,910 possible pairs of lines had estimates of pairwise relatedness over 0.05 (Huang *et al.* 2014) compared to 0.08% of 5,995 pairs of lines reported here. The discrepancy between the levels of relatedness between the DsGRP and DGRP is potentially due to the different demographic histories of the founding populations that generate population structure. North American populations of *D. melanogaster* have relatively complex demographic histories with admixture of African and European ancestors and instances of secondary contact (Pool *et al.* 2012) compared to the endemic population of *D. serrata*.

Another likely explanation for the increased levels of relatedness in the DGRP is the presence of common segregating inversions. While chromosomal inversions are known to segregate in *D. serrata*, their frequency and number tends to increase in populations approaching the equator (Stocker *et al.* 2004). Therefore, it may be the case that founding the DsGRP from the higher latitude of Brisbane has resulted in sampling relatively few inversions. As of yet, these lines have not been karyotyped; however the low levels of relatedness and the lack of any bimodal distribution for residual heterozyosity, such as the one found in the DGRP (Huang *et al.* 2014), where a portion of the lines had high levels of segregating SNP loci (15–20%), suggests that segregating inversions are negligible in this population.

We performed an eigendecomposition of the genomic relatedness matrix to test for the presence of population structure using the approach outlined in Bryc *et al.* (2013). This analysis revealed two large eigenvalues ($\Lambda_1 = 20.08$ and $\Lambda_2 = 1.12$) that were greater than that expected for a random relatedness matrix of equal size (Threshold = 0.993) (Figure 5). There is therefore evidence that the full set of 110 of lines contain substructure in the form of two subpopulations. We reasoned that the second large eigenvalue was likely caused by the four pairs of lines that were highly related to each other ($A_{jk} = 0.29$, 0.38, 0.39, and 1.04; Figure 4). To test this, we repeated the analysis after
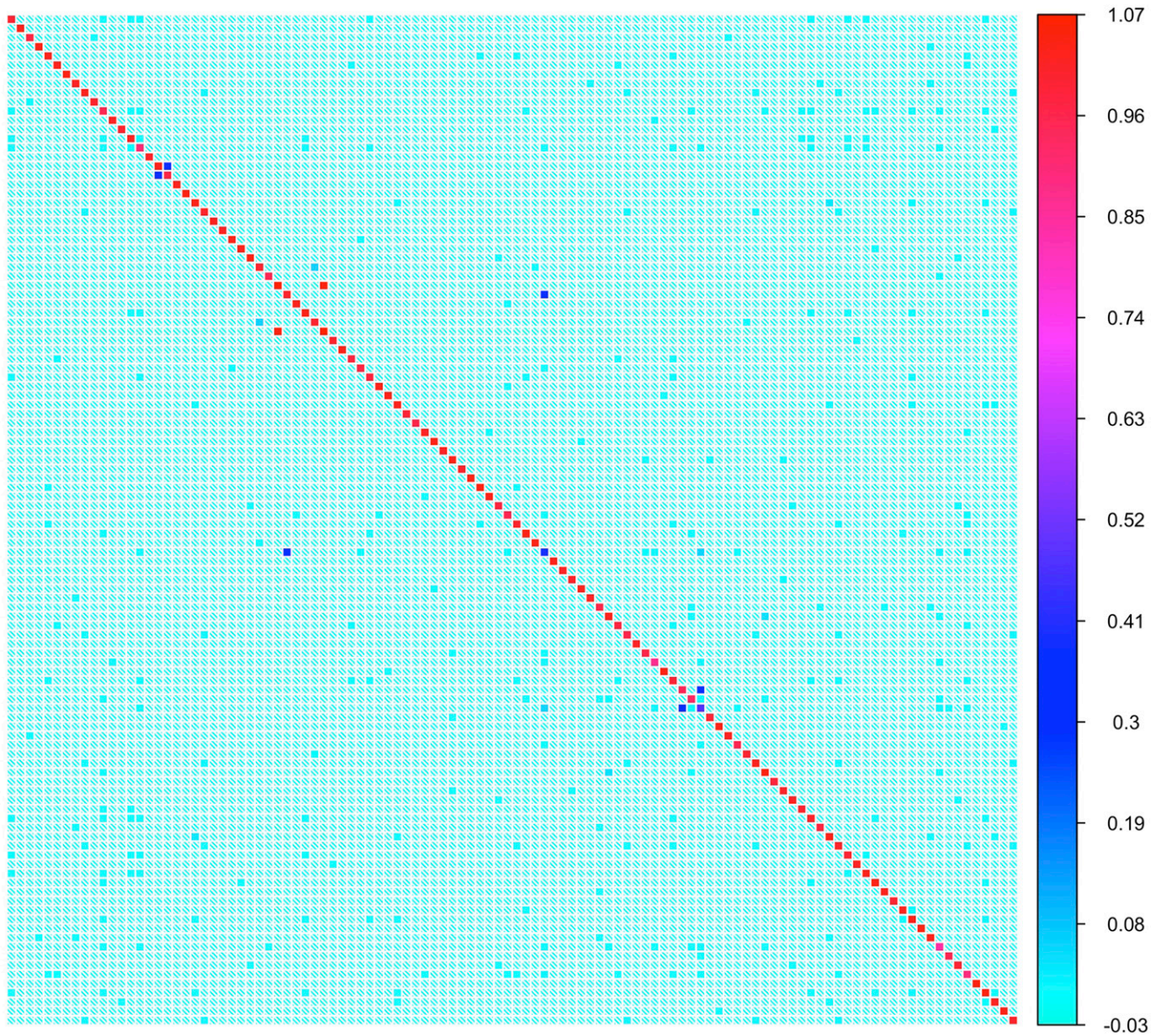
Figure 4 Heat-map of the genomic relatedness matrix for each pair of the 110 inbred DsGRP lines. Each colored squared is an estimate of genomic relatedness between a pair of inbred lines estimated by GCTA (Yang *et al.* 2011). The shade of color represents the degree of relatedness, with light blue showing low levels of relatedness and red high levels of relatedness.

randomly removing one line from each of the four pairs of closely related lines. Confirming the prediction, there was only one significantly large eigenvalue in this second analysis ($\Lambda_1 = 19.66$). Such a result is expected when the data includes only a single population. To summarize, after the four highly-related lines have been removed, there is no clear evidence for population structure in the DsGRP that would lead to spurious genotype-phenotype associations in genome-wide association analysis.

Linkage disequilibrium

The rapid decay of linkage disequilibrium with genomic distance is a common feature of *Drosophila* species with r^2 dropping below 0.1 within 100 base pairs (Long *et al.* 1998; Mackay *et al.* 2012). This allows for higher resolution mapping compared to other species such as maize

and humans in which the equivalent decay does not occur until after approximately 2000 base pairs (Remington *et al.* 2001) and 50,000 base pairs (Koch *et al.* 2013), respectively. In the DsGRP, linkage disequilibrium decays rapidly with r^2, on average, dropping below 0.1 after 75 base pairs. Surprisingly, we observe faster decay on the X chromosome compared to the autosomes (Figure 6), contrary to Mackay *et al.* (2012), despite the fact that the X chromosome has a smaller effective population size than the autosomes.

Nucleotide diversity and neutrality

For the two cosmopolitan species of *Drosophila* that have been studied extensively, *D. melanogaster* and *D. simulans*, the ancestral populations from Africa consistently exhibit higher levels of polymorphism compared to the derived populations from America and Europe (Andolfatto 2001;
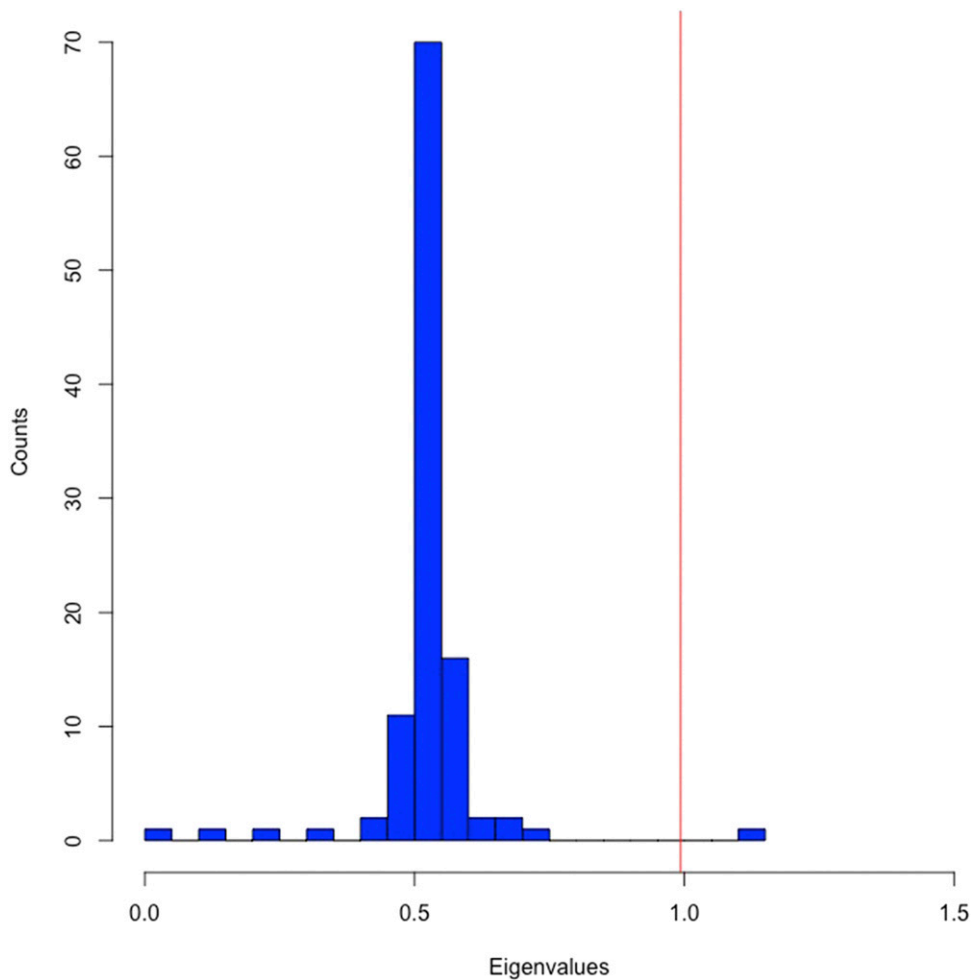
Figure 5 Distribution of eigenvalues from an eigendecomposition of the genomic relatedness matrix for all 110 lines excluding one large eigenvalue where $\Lambda = 20.08$. Eigendecomposition of the genomic relatedness matrix, **X**, which was scaled by equation 2.8 in Bryc *et al.* (2013) using values of N = 3,709,328 and M = 110. Here, N corresponds to the number of SNPs used to estimate **X** and M is the number of lines. Only SNPs with without missing data were used for this analysis. The red vertical line corresponds to the significance threshold, *t*, for declaring an eigenvalue larger than that expected for a random relatedness matrix. $t = (1 + F)/2$ and corresponds to a value of 0.993 with our expected inbreeding coefficient (*F*) of 0.986 after 20 generations of full-sib mating. The two largest eigenvalues were significant, $\Lambda_1 = 20.08$ (not plotted) and $\Lambda_2 = 1.12$.

Baudry *et al.* 2004; Begun and Aquadro 1993; Grenier *et al.* 2015; Lack *et al.* 2015). Presumably, nucleotide diversity is reduced during bottleneck events associated with the colonization of new habitat. The relatively high estimates of nucleotide diversity for *D. mauritiana*, an endemic species from Mauritius, bolster this trend (Garrigan *et al.* 2014). We therefore expected that our population of *D. serrata*, founded from the species' ancestral range, would exhibit relatively high levels of nucleotide diversity. We estimated nucleotide diversity (π) along the major
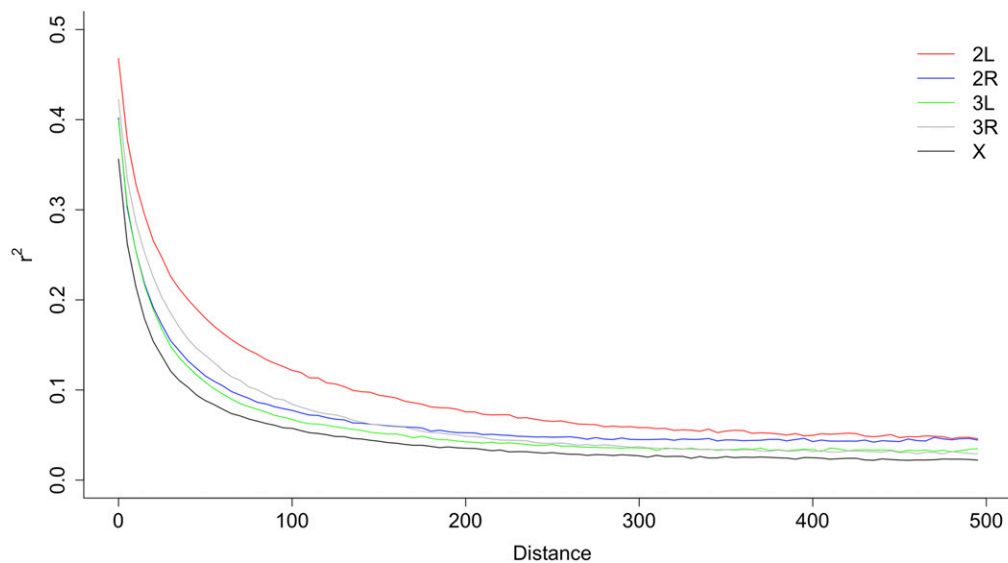

Figure 6 Decay of linkage disequilibrium (r^2) between SNPs with genomic distance (bp) in the DsGRP. Values are averaged across each chromosome.
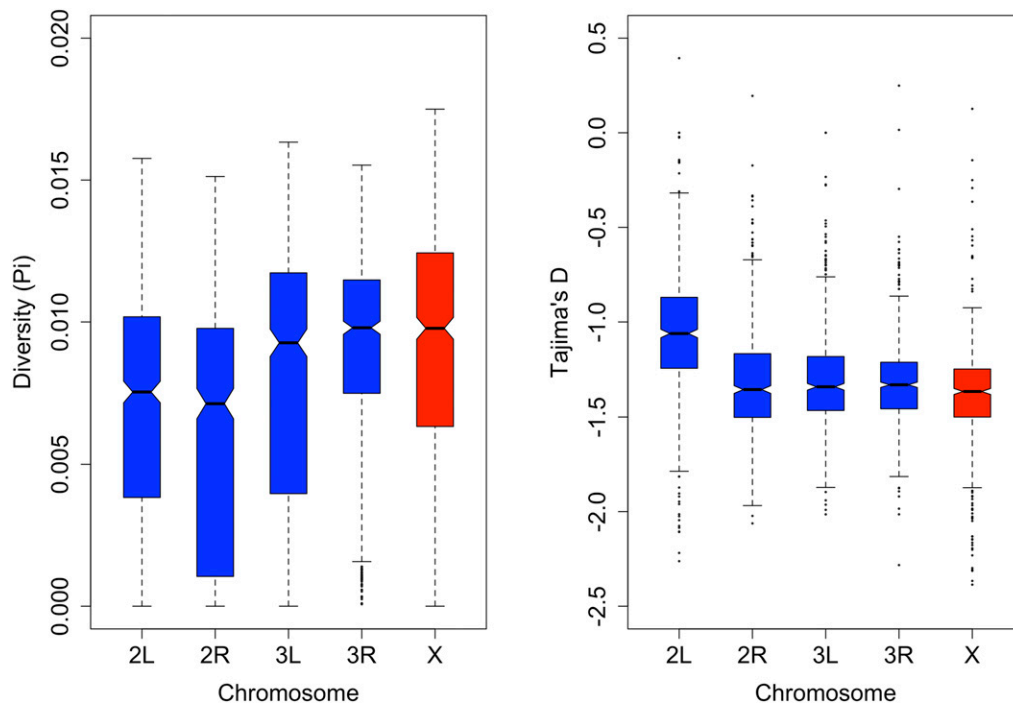
Figure 7 Boxplots of nucleotide diversity and Tajima's D by chromosome arm. Shown are the estimates from 50 kilobase non-overlapping sliding windows with the breakdown of the number of windows per chromosome as follows: 2L = 691, 2R = 664, 3L = 641, 3R = 741, and X = 628.

chromosome arms 2L, 2R, 3L, 3R, and X using a 50 kilobase non-overlapping sliding window approach (Figure 7). Averaged across the genome, we estimated that $\pi = 0.0079$, which is consistent with the pattern seen in other species of relatively increased levels of nucleotide diversity for populations from ancestral ranges compared to more recently established population outside of the ancestral range (Andolfatto 2001; Baudry *et al.* 2004; Begun and Aquadro 1993; Grenier *et al.* 2015; Lack *et al.* 2015).

Using the same sliding window approach, we tested for departures from neutrality using Tajima's D (Figure 7) (Tajima 1989). As with other populations of *Drosophila*, the DGRP (Mackay *et al.* 2012), the Zimbabwean population in the Global Diversity Lines (Grenier *et al.* 2015), and *D. mauritiana* (Garrigan *et al.* 2014), Tajima's D was negative across the entire genome (Tajima's D = -1.27). This is consistent with an abundance of rare alleles and could be indicative of population expansion or the occurrence of selective sweeps, however this statistic cannot distinguish the effects of demography from selection. In the DsGRP, chromosome arm 2L has higher estimates of Tajima's D compared to the other chromosome arms (Figure 7). The causes for this pattern will hopefully be resolved through a more in-depth population genomic analysis, as chromosome 2L may potentially harbor more genomic regions that experience balancing selection, which would increase the average value of Tajima's D.

Genome-wide association analysis of female CHC expression

A major motivation for developing the DsGRP was to begin connecting molecular variation with standing variation for some of the well-studied quantitative traits of *D. serrata*. Here, we applied genome-wide association analysis to the expression of the CHC 2-Me-C$_{28}$ in females and identify new candidate genes that might influence trait variation. Using our mixed model approach, we found 4 SNPs that passed the 0.05 significance threshold after Bonferroni multiple test correction (Figure 8). Two of the SNPs are situated within genes, while the other two lie

within 3kb of genes. We found a further 189 SNPs with p-values lower than a suggestive threshold of 10^{-5}.

We note that a GWAS performed on line mean data using the GCTA program (Yang *et al.* 2011), which like many other mixed model GWAS applications, estimates the polygenic variance only once, detected no SNPs above Bonferroni threshold and only 34 under the arbitrary threshold of 10^{-5}. We also applied our ASREML approach to line means rather than individuals and found that it detected exactly the same 4 SNPs above Bonferroni as our approach in equation 2 (190 SNPs had p-values lower than 10^{-5}). It therefore appears that the increase in detection rate is in this case mainly due to the ASREML model re-estimating the polygenic variance for each SNP tested which results in an exact, rather than approximate, calculation of the test statistic (Zhou and Stephens 2012).

The majority of the literature regarding the expression of CHCs has identified genes that are related to their production within specialized cells, oenocytes. These genes constitute the major biosynthetic pathway known for CHC production and are involved with fatty-acid synthesis, elongation, desaturation, and reduction (Chertemps *et al.* 2007; Chertemps *et al.* 2006; Chung *et al.* 2014; Fang *et al.* 2009; Labeur *et al.* 2002; Marcillac *et al.* 2005; Wicker-Thomas *et al.* 2015). Although none of the genes associated with the statistically significant SNPs found in this study are involved in this biosynthetic pathway, there are other biological processes involved with CHC expression, as measured by hexane washes from the cuticle. How CHCs are transported from the oenocytes to the cuticle is unknown, this study provides a potential candidate gene involved in this process. One of the significant SNPs resides in the gene *Cht9*, a *chitinase* found on chromosome 2R. *Cht9*, along with a number of other *chitinases* and *imaginal-disc-growth-factors* are important for the development of epithelial apical extracellular matrix, which controls the development and maintenance of wound healing, cell signaling, and organ morphogenesis in *Drosophila* (Galko and Krasnow 2004; Turner 2009). Knocking out the expression of *Cht9* with RNAi leads to deformed cuticles, inability to heal wounds, and
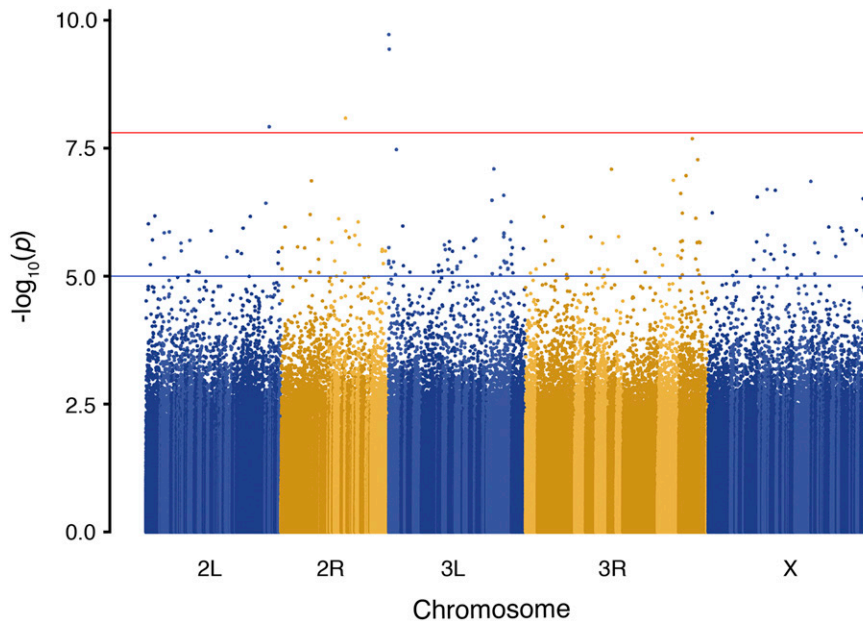
Figure 8 Genome-wide association for 2-Me-C_{28} expression in female *Drosophila serrata*. Red line indicates Bonferroni threshold corresponding to $P = 0.05$ and the blue indicates an arbitrary significance threshold of $P = 10^{-5}$. A total of 3,318,503 biallelic SNPs were analyzed with a minimum minor allele frequency of 0.5. Alternating color shading within chromosomes indicates different contigs in the *D. serrata* genome assembly. Contig order has not yet been established for the genome.

defects in larval and adult molting (Pesch *et al.* 2016), and here, we provide evidence that variation in this gene may also influence other cuticular traits such as CHC abundance. Notwithstanding the small number of lines, the genome-wide association analysis presented here, combined with a previous study that identified the major role of the transcription factor *POU domain motif 3* (*pdm3*) for polymorphic female-limited abdominal pigmentation (Yassin *et al.* 2016), illustrate the potential of the DsGRP to discover novel regions of the genome that underpin the genetic architecture of traits.

Conclusion

We have assembled a new resource for the study of quantitative traits and population genomic variation in a non-model *Drosophila* species within its endemic distribution. These reproducible genotypes sampled from a single population not only provide a rich genomic dataset suitable for population genomic studies, but also provide a critical resource for the discovery of genetic variants underlying ecologically important quantitative traits. We hope that the DsGRP will provide a useful complement to other *Drosophila* resources such as the DGRP (Mackay *et al.* 2012), the DSPR (King *et al.* 2012), and the Drosophila Genome Nexus (Lack *et al.* 2015).

In this first characterization of the DsGRP at the genomic level, we have shown that the inbreeding process has been successful in homogenizing the majority of the genome of each of the lines. Through the estimation of the genomic relatedness matrix we have shown that the DsGRP represents a random sample from a large population that contains very low levels of cryptic relatedness. These characteristics, along with rapid decay of linkage disequilibrium, make the DsGRP an ideal resource for the application of genome-wide association analysis and for generating new multifounder QTL mapping populations that will boost mapping power.

LITERATURE CITED

Aitchison, J., 1986   The statistical analysis of compositional data, Chapman & Hall, London. https://doi.org/10.1007/978-94-009-4109-0

Allen, S. L., R. Bonduriansky, and S. F. Chenoweth, 2013   The genomic distribution of sex-biased genes in *Drosophila serrata*: X chromosome demasculinization, feminization, and hyperexpression in both sexes. Genome Biol. Evol. 5(10): 1986–1994. https://doi.org/10.1093/gbe/evt145

Allen, S. L., R. Bonduriansky, C. M. Sgro, and S. F. Chenoweth, 2017a   Sex-biased transcriptome divergence along a latitudinal gradient. Mol. Ecol. 26(5): 1256–1272. https://doi.org/10.1111/mec.14015

Allen, S. L., E. K. Delaney, A. Kopp, and S. F. Chenoweth. 2017b   Single-molecule sequencing of the *Drosophila serrata* genome. G3-Genes Genomes Genetics 7:781–788. https://doi.org/10.1534/g3.116.037598

Andolfatto, P., 2001   Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18(3): 279–290. https://doi.org/10.1093/oxfordjournals.molbev.a003804

Astle, W., and D. J. Balding, 2009   Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24(4): 451–471. https://doi.org/10.1214/09-STS307

Attrill, H. K., J. L. Falls, G. H. Goodman, G. Millburn, A. J. Antonazzo *et al.*, 2016   FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. Nucleic Acids Res. 44(D1): D786–D792. https://doi.org/10.1093/nar/gkv1046

Ayala, F. J., 1965   Sibling Species of the *Drosophila serrata* group. Evolution 19(4): 538–545. https://doi.org/10.1111/j.1558-5646.1965.tb03329.x

Baudry, E., B. Viginier, and M. Veuille, 2004   Non-African populations of *Drosophila melanogaster* have a unique origin. Mol. Biol. Evol. 21(8): 1482–1491. https://doi.org/10.1093/molbev/msh089

Begun, D. J., and C. F. Aquadro, 1993   African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature 365(6446): 548–550. https://doi.org/10.1038/365548a0

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh *et al.*, 2007   Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5(11): e310. https://doi.org/10.1371/journal.pbio.0050310

Blows, M. W., and R. A. Allan, 1998 Levels of mate recognition within and between two *Drosophila* species and their hybrids. Am. Nat. 152(6): 826–837. https://doi.org/10.1086/286211

Blows, M. W., and A. A. Hoffmann, 1993 The genetics of central and marginal populations of Drosophila serrata. I. Genetic variation for stress resistance and species borders. Evolution 47(4): 1255–1270. https://doi.org/10.1111/j.1558-5646.1993.tb02151.x

Brake, I., and G. Bachli, 2008 Dro- sophilidae (Diptera), (World Catalogue of Insects), Vol. 9. Apollo Books, Stenstrup, Denmark.

Bryc, K., W. Bryc, and J. W. Silverstein, 2013 Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. Theor. Popul. Biol. 89: 34–43. https://doi.org/10.1016/j.tpb.2013.08.004

Burke, M. K., E. G. King, P. Shahrestani, M. R. Rose, and A. D. Long, 2014 Genome-wide association study of extreme longevity in *Drosophila melanogaster*. Genome Biol. Evol. 6(1): 1–11. https://doi.org/10.1093/gbe/evt180

Chang, C. C., C. C. Chow, L. Tellier, S. Vattikuti, S. M. Purcell et al., 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4: 7. https://doi.org/10.1186/s13742-015-0047-8

Chertemps, T. L., C. Duportets, R. Labeur, K. Ueda, K. Takahashi et al., 2007 A female-biased expressed elongase involved in long-chain hydrocarbon biosynthesis and courtship behavior in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 104(11): 4273–4278. https://doi.org/10.1073/pnas.0608142104

Chertemps, T., L. Duportets, C. Labeur, M. Ueyama, and C. Wicker-Thomas, 2006 A female-specific desaturase gene responsible for diene hydrocarbon biosynthesis and courtship behaviour in *Drosophila melanogaster*. Insect Mol. Biol. 15(4): 465–473. https://doi.org/10.1111/j.1365-2583.2006.00658.x

Chung, H., D. W. Loehlin, H. D. Dufour, K. Vaccarro, J. G. Millar et al., 2014 A single gene affects both ecological divergence and mate choice in *Drosophila*. Science 343(6175): 1148–1151. https://doi.org/10.1126/science.1249998

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver et al., 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450(7167): 203–218. https://doi.org/10.1038/nature06341

Da Lage, J. L., G. J. Kergoat, F. Maczkowiak, J. F. Silvain, M. L. Cariou et al., 2007 A phylogeny of *Drosophilidae* using the Amyrel gene: questioning the *Drosophila melanogaster* species group boundaries. J. Zoological Syst. Evol. Res. 45(1): 47–63. https://doi.org/10.1111/j.1439-0469.2006.00389.x

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks et al., 2011 The variant call format and VCFtools. Bioinformatics 27(15): 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dembeck, L. M., W. Huang, M. M. Magwire, F. Lawrence, R. F. Lyman et al., 2015 Genetic architecture of abdominal pigmentation in *Drosophila melanogaster*. PLoS Genet. 11: e1005163. https://doi.org/10.1371/journal.pgen.1005163

Falconer, D. S., and T. F. C. Mackay, 1996 Introduction to Quantitative Genetics, Longman, Burnt Mill, England.

Fang, S., C. T. Ting, C. R. Lee, K. H. Chu, C. C. Wang et al., 2009 Molecular evolution and functional diversification of fatty acid desaturases after recurrent gene duplication in *Drosophila*. Mol. Biol. Evol. 26(7): 1447–1456. https://doi.org/10.1093/molbev/msp057

Frentiu, F. D., M. Adamski, E. A. McGraw, M. W. Blows, and S. F. Chenoweth, 2009 An expressed sequence tag (EST) library for *Drosophila serrata*, a model system for sexual selection and climatic adaptation studies. BMC Genomics 10: 40. https://doi.org/10.1186/1471-2164-10-40

Frentiu, F. D., and S. F. Chenoweth, 2010 Clines in cuticular hydrocarbons in two *Drosophila* species with independent population histories. Evolution 64(6): 1784–1794. https://doi.org/10.1111/j.1558-5646.2009.00936.x

Galko, M. J., and M. A. Krasnow, 2004 Cellular and genetic analysis of wound healing in *Drosophila* larvae. PLoS Biol. 2(8): E239. https://doi.org/10.1371/journal.pbio.0020239

García-Dorado, A., 2008 A simple method to account for natural selection when predicting inbreeding depression. Genetics 180(3): 1559–1566. https://doi.org/10.1534/genetics.108.090597

García-Dorado, A., 2012 Understanding and predicting the fitness decline of shrunk populations: inbreeding, purging, mutation, and standard selection. Genetics 190(4): 1461–1476. https://doi.org/10.1534/genetics.111.135541

Garrigan, D., S. B. Kingan, A. J. Geneva, J. P. Vedanayagam, and D. C. Presgraves, 2014 Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. Genome Biol. Evol. 6(9): 2444–2458. https://doi.org/10.1093/gbe/evu198

Gerken, A. R., O. C. Eller, D. A. Hahn, and T. J. Morgan, 2015 Constraints, independence, and evolution of thermal plasticity: probing genetic architecture of long- and short-term thermal acclimation. Proc. Natl. Acad. Sci. USA 112(14): 4399–4404. https://doi.org/10.1073/pnas.1503456112

Gosden, T. P., and S. F. Chenoweth, 2014 The evolutionary stability of cross-sex, cross-trait genetic covariances. Evolution 68(6): 1687–1697. https://doi.org/10.1111/evo.12398

Grenier, J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed et al. 2015 Global Diversity Lines-A Five-Continent Reference Panel of Sequenced *Drosophila melanogaster* Strains. G3-Genes Genomes Genetics 5:593–603. https://doi.org/10.1534/g3.114.015883

Hallas, R., M. Schiffer, and A. A. Hoffmann, 2002 Clinal variation in *Drosophila serrata* for stress resistance and body size. Genet. Res. 79(02): 141–148. https://doi.org/10.1017/S0016672301005523

Harbison, S. T., L. J. McCoy, and T. F. C. Mackay, 2013 Genome-wide association study of sleep in *Drosophila melanogaster*. BMC Genomics 14: 281. https://doi.org/10.1186/1471-2164-14-281

He, B. Z., M. Z. Ludwig, D. A. Dickerson, L. Barse, B. Arun et al., 2014 Effect of genetic variation in a *Drosophila* model of diabetes-associated misfolded human proinsulin. Genetics 196(2): 557–567. https://doi.org/10.1534/genetics.113.157800

Hine, E., S. Lachish, M. Higgie, and M. W. Blows, 2002 Positive genetic correlation between female preference and offspring fitness. Proc. Biol. Sci. 269(1506): 2215–2219. https://doi.org/10.1098/rspb.2002.2149

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia et al., 2014 Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Res. 24(7): 1193–1208. https://doi.org/10.1101/gr.171546.113

Hunter, C. M., W. Huang, T. F. C. Mackay, and N. D. Singh, 2016 The genetic architecture of natural variation in recombination rate in *Drosophila melanogaster*. PLoS Genet. 12: e1005951. https://doi.org/10.1371/journal.pgen.1005951

Jenkins, N. L., and A. A. Hoffmann, 1999 Limits to the southern border of *Drosophila serrata*: cold resistance, heritable variation, and trade-offs. Evolution 53: 1823–1834. https://doi.org/10.1111/j.1558-5646.1999.tb04565.x

Kellermann, V., B. van Heerwaarden, C. M. Sgro, and A. A. Hoffmann, 2009 Fundamental evolutionary limits in ecological traits drive *Drosophila* species distributions. Science 325(5945): 1244–1246. https://doi.org/10.1126/science.1175443

King, E. G., G. Kislukhin, K. N. Walters, and A. D. Long, 2014 Using *Drosophila melanogaster* To identify chemotherapy toxicity genes. Genetics 198(1): 31–43. https://doi.org/10.1534/genetics.114.161968

King, E. G., S. J. Macdonald, and A. D. Long, 2012 Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. Genetics 191(3): 935–949. https://doi.org/10.1534/genetics.112.138537

Kislukhin, G., E. G. King, K. N. Walters, S. J. Macdonald, and A. D. Long. 2013 The genetic architecture of methotrexate toxicity is similar in *Drosophila melanogaster* and humans. G3-Genes Genomes Genetics 3:1301–1310. https://doi.org/10.1534/g3.113.006619

Kittles, R. A., W. D. Chen, R. K. Panguluri, C. Ahaghotu, A. Jackson et al., 2002 CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population

G3 Genes | Genomes | Genetics

stratification? Hum. Genet. 110(6): 553–560. https://doi.org/10.1007/s00439-002-0731-5

Knowler, W. C., R. C. Williams, D. J. Pettitt, and A. G. Steinberg, 1988   Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians and genetic admixture. Am. J. Hum. Genet. 43: 520–526.

Koch, E., M. Ristroph, and M. Kirkpatrick, 2013   Long range linkage disequilibrium across the human genome. PLoS One 8: e80754. https://doi.org/10.1371/journal.pone.0080754

Kruijer, W., M. P. Boer, M. Malosetti, P. J. Flood, B. Engel et al., 2015   Marker-based estimation of heritability in immortal populations. Genetics 199(2): 379–398. https://doi.org/10.1534/genetics.114.167916

Labeur, C., R. Dallerac, and C. Wicker-Thomas, 2002   Involvement of desat1 gene in the control of Drosophila melanogaster pheromone biosynthesis. Genetica 114(3): 269–274. https://doi.org/10.1023/A:1016223000650

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig et al., 2015   The Drosophila Genome Nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics 199(4): 1229–1241. https://doi.org/10.1534/genetics.115.174664

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider et al., 2012   Genomic variation in natural populations of Drosophila melanogaster. Genetics 192(2): 533–598. https://doi.org/10.1534/genetics.112.142018

Lemeunier, F., J. R. David, L. Tsacas, and M. Ashburner 1986   The melanogaster species group pp. 147-256 in The genetics and biology of Drosophila Vol. 3, edited by M. Ashburner, H. L. Carson, and J. N. Thompson, Academic, London.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson et al., 2011   FaST linear mixed models for genome-wide association studies. Nat. Methods 8(10): 833–835. https://doi.org/10.1038/nmeth.1681

Long, A. D., R. F. Lyman, C. H. Langley, and T. F. C. Mackay, 1998   Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in Drosophila melanogaster. Genetics 149: 999–1017.

Machado, H. E., A. O. Bergland, K. R. O'Brien, E. L. Behrman, P. S. Schmidt, and D. A. Petrov, 2016   Comparative population genomics of latitudinal variation in Drosophila simulans and Drosophila melanogaster. Molecular Ecology 25: 723–740.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles et al., 2012   The Drosophila melanogaster Genetic Reference Panel. Nature 482 (7384): 173–178. https://doi.org/10.1038/nature10811

Magiafoglou, A., M. E. Carew, and A. A. Hoffmann, 2002   Shifting clinal patterns and microsatellite variation in Drosophila serrata populations: a comparison of populations near the southern border of the species range. J. Evol. Biol. 15(5): 763–774. https://doi.org/10.1046/j.1420-9101.2002.00439.x

Magwire, M. M., D. K. Fabian, H. Schweyen, C. Cao, B. Longdon et al., 2012   Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in Drosophila melanogaster. PLoS Genet. 8: e1003057. https://doi.org/10.1371/journal.pgen.1003057

Marcillac, F., F. Bousquet, J. Alabouvette, F. Savarit, and J. F. Ferveur, 2005   A mutation with major effects on Drosophila melanogaster sex pheromones. Genetics 171(4): 1617–1628. https://doi.org/10.1534/genetics.104.033159

Marriage, T. N., E. G. King, A. D. Long, and S. J. Macdonald, 2014   Fine-mapping nicotine resistance loci in Drosophila using a multiparent advanced generation inter-cross population. Genetics 198(1): 45–57. https://doi.org/10.1534/genetics.114.162107

McGuigan, K., J. M. Collet, E. A. McGraw, Y. X. H. Ye, S. L. Allen et al., 2014   The nature and extent of mutational pleiotropy in gene expression of male Drosophila serrata. Genetics 196(3): 911–921. https://doi.org/10.1534/genetics.114.161232

McGuigan, K., D. Petfield, and M. W. Blows, 2011   Reducing mutation load through sexual selection on males. Evolution 65(10): 2816–2829. https://doi.org/10.1111/j.1558-5646.2011.01346.x

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., 2010   The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9): 1297–1303. https://doi.org/10.1101/gr.107524.110

Nuzhdin, S. V., E. G. Pasyukova, C. L. Dilda, Z. B. Zeng, and T. F. C. Mackay, 1997   Sex-specific quantitative trait loci affecting longevity in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA 94(18): 9734–9739. https://doi.org/10.1073/pnas.94.18.9734

Pesch, Y. Y., D. Riedel, K. R. Patil, G. Loch, and M. Behr, 2016   Chitinases and Imaginal disc growth factors organize the extracellular matrix formation at barrier tissues in insects. Sci. Rep. 6: 18340. https://doi.org/10.1038/srep18340

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno et al., 2012   Population genomics of Sub-Saharan Drosophila melanogaster: African Diversity and non-African admixture. PLoS Genet. 8: e1003080. https://doi.org/10.1371/journal.pgen.1003080

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, and S. R. Whitt, J. et al. 2001   Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences of the United States of America 98:11479–11484. https://doi.org/10.1073/pnas.201394398

Shorter, J., C. Couch, W. Huang, M. A. Carbone, J. Peiffer et al., 2015   Genetic architecture of natural variation in Drosophila melanogaster aggressive behavior. Proc. Natl. Acad. Sci. USA 112(27): E3555–E3563. https://doi.org/10.1073/pnas.1510104112

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts et al., 2007   Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450(7167): 219–232. https://doi.org/10.1038/nature06340

Stocker, A. J., B. Foley, and A. Hoffmann, 2004   Inversion frequencies in Drosophila serrata along an eastern Australian transect. Genome 47(6): 1144–1153. https://doi.org/10.1139/g04-078

Stocker, A. J., B. B. Rusuwa, M. J. Blacket, F. D. Frentiu, M. Sullivan et al. 2012   Physical and linkage maps for Drosophila serrata, a model species for studies of clinal adaptation and sexual selection. G3-Genes Genomes Genetics 2:287–297. https://doi.org/10.1534/g3.111.001354

Stone, E. A., 2012   Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. Genome Res. 22(5): 966–974. https://doi.org/10.1101/gr.129122.111

Sturgill, D., Y. Zhang, M. Parisi, and B. Oliver, 2007   Demasculinization of X chromosomes in the Drosophila genus. Nature 450(7167): 238–241. https://doi.org/10.1038/nature06330

Tajima, F., 1989   Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Treangen, T. J., and S. L. Salzberg, 2012   Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13(1): 36–46 (erratum: Nat. Rev. Genet. 13: 146). https://doi.org/10.1038/nrg3117

Turner, J. R., 2009   Intestinal mucosal barrier function in health and disease. Nat. Rev. Immunol. 9(11): 799–809. https://doi.org/10.1038/nri2653

Unckless, R. L., S. M. Rottschaefer, and B. P. Lazzaro. 2015   A genome-wide association study for nutritional indices in Drosophila. G3-Genes Genomes Genetics 5:417–425. https://doi.org/10.1534/g3.114.016477

van Heerwaarden, B., V. Kellermann, M. Schiffer, M. Blacket, C. M. Sgro et al., 2009   Testing evolutionary hypotheses about species borders: patterns of genetic variation towards the southern borders of two rainforest Drosophila and a related habitat generalist. Proc. Biol. Sci. 276(1661): 1517–1526. https://doi.org/10.1098/rspb.2008.1288

Weber, A. L., G. F. Khan, M. M. Magwire, C. L. Tabor, T. F. C. Mackay et al., 2012   Genome-wide association analysis of oxidative stress resistance in Drosophila melanogaster. PLoS One 7: e34745. https://doi.org/10.1371/journal.pone.0034745

Wicker-Thomas, C., D. Garrido, G. Bontonou, L. Napal, N. Mazuras et al., 2015   Flexible origin of hydrocarbon/pheromone precursors in Drosophila melanogaster. J. Lipid Res. 56(11): 2094–2101. https://doi.org/10.1194/jlr.M060368

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88(1): 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yassin, A., 2013 Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Syst. Entomol. 38(2): 349–364. https://doi.org/10.1111/j.1365-3113.2012.00665.x

Yassin, A., E. K. Delaney, A. J. Reddiex, T. D. Seher, H. Bastide *et al.*, 2016 The pdm3 locus is a hotspot for recurrent evolution of female-limited color dimorphism in *Drosophila*. Curr. Biol. 26(18): 2412–2422. https://doi.org/10.1016/j.cub.2016.07.016

Zhang, Y., D. Sturgill, M. Parisi, S. Kumar, and B. Oliver, 2007 Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature 450(7167): 233–237. https://doi.org/10.1038/nature06323

Zhao, L., J. Wit, N. Svetec, and D. J. Begun, 2015 Parallel gene expression differences between low and high latitude populations of *Drosophila melanogaster* and *D. simulans*. Plos Genetics 11.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44(7): 821–824. https://doi.org/10.1038/ng.2310

Communicating editor: R. Kulathinal

G3·Genes | Genomes | Genetics