A Genomic Reference Panel for *Drosophila serrata*

Adam R. Reddiex[1], Scott, L. Allen[1], Stephen F. Chenoweth[1]*

1. School of Biological Sciences, The University of Queensland, QLD 4072, Australia.

* correspondence: s.chenoweth@uq.edu.au

Keywords: *montium, gwas, population genetics, quantitative genetics, multi-parental population*

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-MeC_{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.

Introduction

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 characterise 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), behaviours (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^{o}N) (Ayala 1965) to as far south as Woolongong, Australia (34.3^{o}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 Hiseq 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 call 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:

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

where, χ_{ij} is the number of copies of the reference allele for the i^{th} 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, **A** = **WW'**/N, is normalised 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, **X** = **CC'**, where **C** is an M x N rectangular matrix with M corresponding to the number of individuals used to estimate **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*. **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 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 100µl 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%:

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

Here, CHC expression (y) of replicate individual *j* on genotype *i* is modelled 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 ($\boldsymbol{G}$) where $\boldsymbol{G}$ has a $\sim N(0,\sigma_A^2\boldsymbol{A})$ 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 modelling 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 Statement

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 supplementary files (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. 2011). The SNP list used to analyse the data in this study is available from Dryad (doi:XXXYYY) and also from www.chenowethlab.org/resources). 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 ± 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 (Fig. 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 (Fig. 2). Across lines, the average proportion of segregating SNPs was 0.86% ± 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 (Garcia-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 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 (Fig. 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 (Fig. 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) (Fig. 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; Fig. 4). To test this, we repeated the analysis after 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 summarise, 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 (Fig. 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; 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 colonisation 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 chromosome arms 2L, 2R, 3L, 3R, and X using a 50 kilobase non-overlapping sliding window approach (Table 1, Fig. 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 (Table 1, Fig. 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 (Fig. 7). The causes for this pattern

441 will hopefully be resolved through a more in-depth population genomic analysis,
442 as chromosome 2L may potentially harbour more genomic regions that
443 experience balancing selection, which would increase the average value of
444 Tajima's D.
445
446 *Genome-wide association analysis of female CHC expression*
447 A major motivation for developing the DsGRP was to begin connecting molecular
448 variation with standing variation for some of the well-studied quantitative traits
449 of *D. serrata*. Here, we applied genome-wide association analysis to the
450 expression of the CHC 2-Me-C_{28} in females and identify new candidate genes that
451 might influence trait variation. Using our mixed model approach, we found 4
452 SNPs that passed the 0.05 significance threshold after Bonferroni multiple test
453 correction (Figure 8). Two of the SNPs are situated within genes, while the other
454 two lie within 3kb of genes. We found a further 189 SNPs with p-values lower
455 than a suggestive threshold of 10^{-5}.
456
457 We note that a GWAS performed on line mean data using the GCTA program
458 (Yang et al. 2011), which like many other mixed model GWAS applications,
459 estimates the polygenic variance only once, detected no SNPs above Bonferroni
460 threshold and only 34 under the arbitrary threshold of 10^{-5}. We also applied our
461 ASREML approach to line means rather than individuals and found that it
462 detected exactly the same 4 SNPs above Bonferroni as our approach in eq. 2 (190
463 SNPs had p-values lower than 10^{-5}). It therefore appears that the increase in
464 detection rate is in this case mainly due to the ASREML model re-estimating the
465 polygenic variance for each SNP tested which results in an exact, rather than
466 approximate, calculation of the test statistic (Zhou and Stephens 2012).

467 The majority of the literature regarding the expression of CHCs has identified
468 genes that are related to their production within specialised cells, oenocytes.
469 These genes constitute the major biosynthetic pathway known for CHC
470 production and are involved with fatty-acid synthesis, elongation, desaturation,
471 and reduction (Chertemps et al. 2007; Chertemps et al. 2006; Chung et al. 2014;
472 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 signalling, 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 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 characterisation of the DsGRP at the genomic level, we have shown that the inbreeding process has been successful in homogenising 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.

Acknowledgements

We thank EK Delaney for comments on the manuscript and NC Appleton for assistance in the laboratory. This work was supported by an Australian Postgraduate Award to ARR and funds from the Australian Research Council and The University of Queensland awarded to SFC.

References:

Aitchison, J. 1986, The statistical analysis of compositional data. London, Chapman & Hall.

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 Biology and Evolution 5:1986-1994.

Allen, S. L., R. Bonduriansky, C. M. Sgro, and S. F. Chenoweth. 2017a. Sex-biased transcriptome divergence along a latitudinal gradient. Molecular Ecology 26:1256-1272.

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.

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Molecular Biology and Evolution 18:279-290.

Astle, W., and D. J. Balding. 2009. Population structure and cryptic relatedness in genetic association studies. Statistical Science 24:451-471.

Attrill, H., K. Falls, J. L. Goodman, G. H. Millburn, G. Antonazzo, A. J. Rey, S. J. Marygold et al. 2016. FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. Nucleic Acids Research 44:D786-D792.

Ayala, F. J. 1965. Sibling Species of the *Drosophila serrata* group Evolution 19:538-545.

544 Baudry, E., B. Viginier, and M. Veuille. 2004. Non-African populations of
545 *Drosophila melanogaster* have a unique origin. Molecular Biology and
546 Evolution 21:1482-1491.
547 Begun, D. J., and C. F. Aquadro. 1993. African and North American populations of
548 Drosophila melanogaster are very different at the DNA level. Nature
549 365:548-550.
550 Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh, M. W. Hahn, P. M.
551 Nista et al. 2007. Population genomics: Whole-genome analysis of
552 polymorphism and divergence in *Drosophila simulans*. Plos Biology
553 5:2534-2559.
554 Blows, M. W., and R. A. Allan. 1998. Levels of mate recognition within and
555 between two *Drosophila* species and their hybrids. American Naturalist
556 152:826-837.
557 Blows, M. W., and A. A. Hoffmann. 1993. The genetics of central and marginal
558 populations of Drosophila serrata .1. Genetic variation for stress
559 resistance and species borders. Evolution 47:1255-1270.
560 Brake, I., and G. Bachli. 2008, World Catalogue of Insects, Volume 9: Dro-
561 sophilidae (Diptera). Stenstrup, Denmark, Apollo Books.
562 Bryc, K., W. Bryc, and J. W. Silverstein. 2013. Separation of the largest eigenvalues
563 in eigenanalysis of genotype data from discrete subpopulations.
564 Theoretical Population Biology 89:34-43.
565 Burke, M. K., E. G. King, P. Shahrestani, M. R. Rose, and A. D. Long. 2014. Genome-
566 wide association study of extreme longevity in *Drosophila melanogaster*.
567 Genome Biology and Evolution 6:1-11.
568 Chang, C. C., C. C. Chow, L. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015.
569 Second-generation PLINK: rising to the challenge of larger and richer
570 datasets. Gigascience 4.
571 Chertemps, T., L. Duportets, C. Labeur, R. Ueda, K. Takahashi, K. Saigo, and C.
572 Wicker-Thomas. 2007. A female-biased expressed elongase involved in
573 long-chain hydrocarbon biosynthesis and courtship behavior in
574 *Drosophila melanogaster*. Proceedings of the National Academy of
575 Sciences of the United States of America 104:4273-4278.
576 Chertemps, T., L. Duportets, C. Labeur, M. Ueyama, and C. Wicker-Thomas. 2006.
577 A female-specific desaturase gene responsible for diene hydrocarbon
578 biosynthesis and courtship behaviour in *Drosophila melanogaster*. Insect
579 Molecular Biology 15:465-473.
580 Chung, H., D. W. Loehlin, H. D. Dufour, K. Vaccarro, J. G. Millar, and S. B. Carroll.
581 2014. A single gene affects both ecological divergence and mate choice in
582 *Drosophila*. Science 343:1148-1151.
583 Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C.
584 Kaufman et al. 2007. Evolution of genes and genomes on the *Drosophila*
585 phylogeny. Nature 450:203-218.
586 Da Lage, J. L., G. J. Kergoat, F. Maczkowiak, J. F. Silvain, M. L. Cariou, and D.
587 Lachaise. 2007. A phylogeny of *Drosophilidae* using the Amyrel gene:
588 questioning the *Drosophila melanogaster* species group boundaries.
589 Journal of Zoological Systematics and Evolutionary Research 45:47-63.
590 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E.
591 Handsaker et al. 2011. The variant call format and VCFtools.
592 Bioinformatics 27:2156-2158.

593 Dembeck, L. M., W. Huang, M. M. Magwire, F. Lawrence, R. F. Lyman, and T. F. C.
594 Mackay. 2015. Genetic architecture of abdominal pigmentation in
595 *Drosophila melanogaster*. Plos Genetics 11.
596 Falconer, D. S., and T. F. C. Mackay. 1996, Introduction to Quantitative Genetics.
597 Burnt Mill, England, Longman.
598 Fang, S., C. T. Ting, C. R. Lee, K. H. Chu, C. C. Wang, and S. C. Tsaur. 2009.
599 Molecular evolution and functional diversification of fatty acid
600 desaturases after recurrent gene duplication in *Drosophila*. Molecular
601 Biology and Evolution 26:1447-1456.
602 Frentiu, F. D., M. Adamski, E. A. McGraw, M. W. Blows, and S. F. Chenoweth. 2009.
603 An expressed sequence tag (EST) library for *Drosophila serrata*, a model
604 system for sexual selection and climatic adaptation studies. BMC
605 Genomics 10.
606 Frentiu, F. D., and S. F. Chenoweth. 2010. Clines in cuticular hydrocarbons in two
607 *Drosophila* species with independent population histories. Evolution
608 64:1784-1794.
609 Galko, M. J., and M. A. Krasnow. 2004. Cellular and genetic analysis of wound
610 healing in *Drosophila* larvae. Plos Biology 2:1114-1126.
611 Garcia-Dorado, A. 2008. A simple method to account for natural selection when
612 predicting inbreeding depression. Genetics 180:1559-1566.
613 —. 2012. Understanding and predicting the fitness decline of shrunk
614 populations: inbreeding, purging, mutation, and standard selection.
615 Genetics 190:1461-1476.
616 Garrigan, D., S. B. Kingan, A. J. Geneva, J. P. Vedanayagam, and D. C. Presgraves.
617 2014. Genome diversity and divergence in *Drosophila mauritiana*:
618 multiple signatures of faster X evolution. Genome Biology and Evolution
619 6:2444-2458.
620 Gerken, A. R., O. C. Eller, D. A. Hahn, and T. J. Morgan. 2015. Constraints,
621 independence, and evolution of thermal plasticity: probing genetic
622 architecture of long- and short-term thermal acclimation. Proceedings of
623 the National Academy of Sciences of the United States of America
624 112:4399-4404.
625 Gosden, T. P., and S. F. Chenoweth. 2014. The evolutionary stability of cross-sex,
626 cross-trait genetic covariances. Evolution 68:1687-1697.
627 Grenier, J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett,
628 R. Boughton et al. 2015. Global Diversity Lines-A Five-Continent
629 Reference Panel of Sequenced Drosophila melanogaster Strains. G3-Genes
630 Genomes Genetics 5:593-603.
631 Hallas, R., M. Schiffer, and A. A. Hoffmann. 2002. Clinal variation in *Drosophila*
632 *serrata* for stress resistance and body size. Genetical Research 79:141-
633 148.
634 Harbison, S. T., L. J. McCoy, and T. F. C. Mackay. 2013. Genome-wide association
635 study of sleep in *Drosophila melanogaster*. BMC Genomics 14.
636 He, B. Z., M. Z. Ludwig, D. A. Dickerson, L. Barse, B. Arun, B. J. Vilhjalmsson, S. Y.
637 Park et al. 2014. Effect of genetic variation in a *Drosophila* model of
638 diabetes-associated misfolded human proinsulin. Genetics 196:557-+.
639 Hine, E., S. Lachish, M. Higgie, and M. W. Blows. 2002. Positive genetic correlation
640 between female preference and offspring fitness. Proceedings of the Royal
641 Society B-Biological Sciences 269:2215-2219.

642 Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia, A. M. Tarone, L. Turlapati
643 et al. 2014. Natural variation in genome architecture among 205
644 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Research
645 24:1193-1208.
646 Hunter, C. M., W. Huang, T. F. C. Mackay, and N. D. Singh. 2016. The genetic
647 architecture of natural variation in recombination rate in *Drosophila*
648 *melanogaster*. Plos Genetics 12.
649 Jenkins, N. L., and A. A. Hoffmann. 1999. Limits to the southern border of
650 *Drosophila serrata*: cold resistance, heritable variation, and trade-offs.
651 Evolution 53:1823-1834.
652 Kellermann, V., B. van Heerwaarden, C. M. Sgro, and A. A. Hoffmann. 2009.
653 Fundamental evolutionary limits in ecological traits drive *Drosophila*
654 species distributions. Science 325:1244-1246.
655 King, E. G., G. Kislukhin, K. N. Walters, and A. D. Long. 2014. Using *Drosophila*
656 *melanogaster* To identify chemotherapy toxicity genes. Genetics 198:31-+.
657 King, E. G., S. J. Macdonald, and A. D. Long. 2012. Properties and power of the
658 *Drosophila* Synthetic Population Resource for the routine dissection of
659 complex traits. Genetics 191:935-U522.
660 Kislukhin, G., E. G. King, K. N. Walters, S. J. Macdonald, and A. D. Long. 2013. The
661 genetic architecture of methotrexate toxicity is similar in *Drosophila*
662 *melanogaster* and humans. G3-Genes Genomes Genetics 3:1301-1310.
663 Kittles, R. A., W. D. Chen, R. K. Panguluri, C. Ahaghotu, A. Jackson, C. A.
664 Adebamowo, R. Griffin et al. 2002. CYP3A4-V and prostate cancer in
665 African Americans: causal or confounding association because of
666 population stratification? Human Genetics 110:553-560.
667 Knowler, W. C., R. C. Williams, D. J. Pettitt, and A. G. Steinberg. 1988. Gm3;5,13,14
668 and type 2 diabetes mellitus: an association in American Indians and
669 genetic admixture. American Journal of Human Genetics 43:520-526.
670 Koch, E., M. Ristroph, and M. Kirkpatrick. 2013. Long range linkage
671 disequilibrium across the human genome. Plos One 8.
672 Kruijer, W., M. P. Boer, M. Malosetti, P. J. Flood, B. Engel, R. Kooke, J. J. B.
673 Keurentjes et al. 2015. Marker-based estimation of heritability in
674 immortal populations. Genetics 199:379-393.
675 Labeur, C., R. Dallerac, and C. Wicker-Thomas. 2002. Involvement of desat1 gene
676 in the control of *Drosophila melanogaster* pheromone biosynthesis.
677 Genetica 114:269-274.
678 Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A.
679 Stevens, C. H. Langley et al. 2015. The *Drosophila* Genome Nexus: a
680 population genomic resource of 623 *Drosophila melanogaster* genomes,
681 including 197 from a single ancestral range population. Genetics
682 199:1229-U1553.
683 Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, J. E. Pool, S. A.
684 Langley et al. 2012. Genomic variation in natural populations of
685 *Drosophila melanogaster*. Genetics 192:533-+.
686 Lemeunier, F., J. R. David, L. Tsacas, and M. Ashburner. 1986. The melanogaster
687 species group. The genetics and biology of Drosophila 3:147-256.
688 Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with
689 BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

690 Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman.
691 2011. FaST linear mixed models for genome-wide association studies.
692 Nature Methods 8:833-U894.
693 Long, A. D., R. F. Lyman, C. H. Langley, and T. F. C. Mackay. 1998. Two sites in the
694 Delta gene region contribute to naturally occurring variation in bristle
695 number in *Drosophila melanogaster*. Genetics 149:999-1017.
696 Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. H. Zhu, S.
697 Casillas et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel.
698 Nature 482:173-178.
699 Magiafoglou, A., M. E. Carew, and A. A. Hoffmann. 2002. Shifting clinal patterns
700 and microsatellite variation in *Drosophila serrata* populations: a
701 comparison of populations near the southern border of the species range.
702 Journal of Evolutionary Biology 15:763-774.
703 Magwire, M. M., D. K. Fabian, H. Schweyen, C. Cao, B. Longdon, F. Bayer, and F. M.
704 Jiggins. 2012. Genome-wide association studies reveal a simple genetic
705 basis of resistance to naturally coevolving viruses in *Drosophila*
706 *melanogaster*. Plos Genetics 8.
707 Marcillac, F., F. Bousquet, J. Alabouvette, F. Savarit, and J. F. Ferveur. 2005. A
708 mutation with major effects on *Drosophila melanogaster* sex pheromones.
709 Genetics 171:1617-1628.
710 Marriage, T. N., E. G. King, A. D. Long, and S. J. Macdonald. 2014. Fine-mapping
711 nicotine resistance loci in *Drosophila* using a multiparent advanced
712 generation inter-cross population. Genetics 198:45-+.
713 McGuigan, K., J. M. Collet, E. A. McGraw, Y. X. H. Ye, S. L. Allen, S. F. Chenoweth,
714 and M. W. Blows. 2014. The nature and extent of mutational pleiotropy in
715 gene expression of male *Drosophila serrata*. Genetics 196:911-+.
716 McGuigan, K., D. Petfield, and M. W. Blows. 2011. Reducing mutation load
717 through sexual selection on males. Evolution 65:2816-2829.
718 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K.
719 Garimella et al. 2010. The Genome Analysis Toolkit: a MapReduce
720 framework for analyzing next-generation DNA sequencing data. Genome
721 Research 20:1297-1303.
722 Nuzhdin, S. V., E. G. Pasyukova, C. L. Dilda, Z. B. Zeng, and T. F. C. Mackay. 1997.
723 Sex-specific quantitative trait loci affecting longevity in *Drosophila*
724 *melanogaster*. Proceedings of the National Academy of Sciences of the
725 United States of America 94:9734-9739.
726 Pesch, Y. Y., D. Riedel, K. R. Patil, G. Loch, and M. Behr. 2016. Chitinases and
727 Imaginal disc growth factors organize the extracellular matrix formation
728 at barrier tissues in insects. Scientific Reports 6.
729 Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W.
730 Crepeau, P. Duchen et al. 2012. Population genomics of Sub-Saharan
731 *Drosophila melanogaster*: African Diversity and non-African admixture.
732 Plos Genetics 8.
733 Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J.
734 Doeblay, S. Kresovich et al. 2001. Structure of linkage disequilibrium and
735 phenotypic associations in the maize genome. Proceedings of the National
736 Academy of Sciences of the United States of America 98:11479-11484.
737 Shorter, J., C. Couch, W. Huang, M. A. Carbone, J. Peiffer, R. R. H. Anholt, and T. F. C.
738 Mackay. 2015. Genetic architecture of natural variation in *Drosophila*

melanogaster aggressive behavior. Proceedings of the National Academy of Sciences of the United States of America 112:E3555-E3563.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450:219-232.

Stocker, A. J., B. Foley, and A. Hoffmann. 2004. Inversion frequencies in *Drosophila serrata* along an eastern Australian transect. Genome 47:1144-1153.

Stocker, A. J., B. B. Rusuwa, M. J. Blacket, F. D. Frentiu, M. Sullivan, B. R. Foley, S. Beatson 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.

Stone, E. A. 2012. Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. Genome Research 22:966-974.

Sturgill, D., Y. Zhang, M. Parisi, and B. Oliver. 2007. Demasculinization of X chromosomes in the *Drosophila* genus. Nature 450:238-U233.

Tajima, F. 1989. Statistical method for testing the neitral 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. Nature Reviews Genetics 13:36-46.

Turner, J. R. 2009. Intestinal mucosal barrier function in health and disease. Nature Reviews Immunology 9:799-809.

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.

van Heerwaarden, B., V. Kellermann, M. Schiffer, M. Blacket, C. M. Sgro, and A. A. Hoffmann. 2009. Testing evolutionary hypotheses about species borders: patterns of genetic variation towards the southern borders of two rainforest *Drosophila* and a related habitat generalist. Proceedings of the Royal Society B-Biological Sciences 276:1517-1526.

Weber, A. L., G. F. Khan, M. M. Magwire, C. L. Tabor, T. F. C. Mackay, and R. R. H. Anholt. 2012. Genome-wide association analysis of oxidative stress resistance in *Drosophila melanogaster*. Plos One 7.

Wicker-Thomas, C., D. Garrido, G. Bontonou, L. Napal, N. Mazuras, B. Denis, T. Rubin et al. 2015. Flexible origin of hydrocarbon/pheromone precursors in *Drosophila melanogaster*. Journal of Lipid Research 56:2094-2101.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88:76-82.

Yassin, A. 2013. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Systematic Entomology 38:349-364.

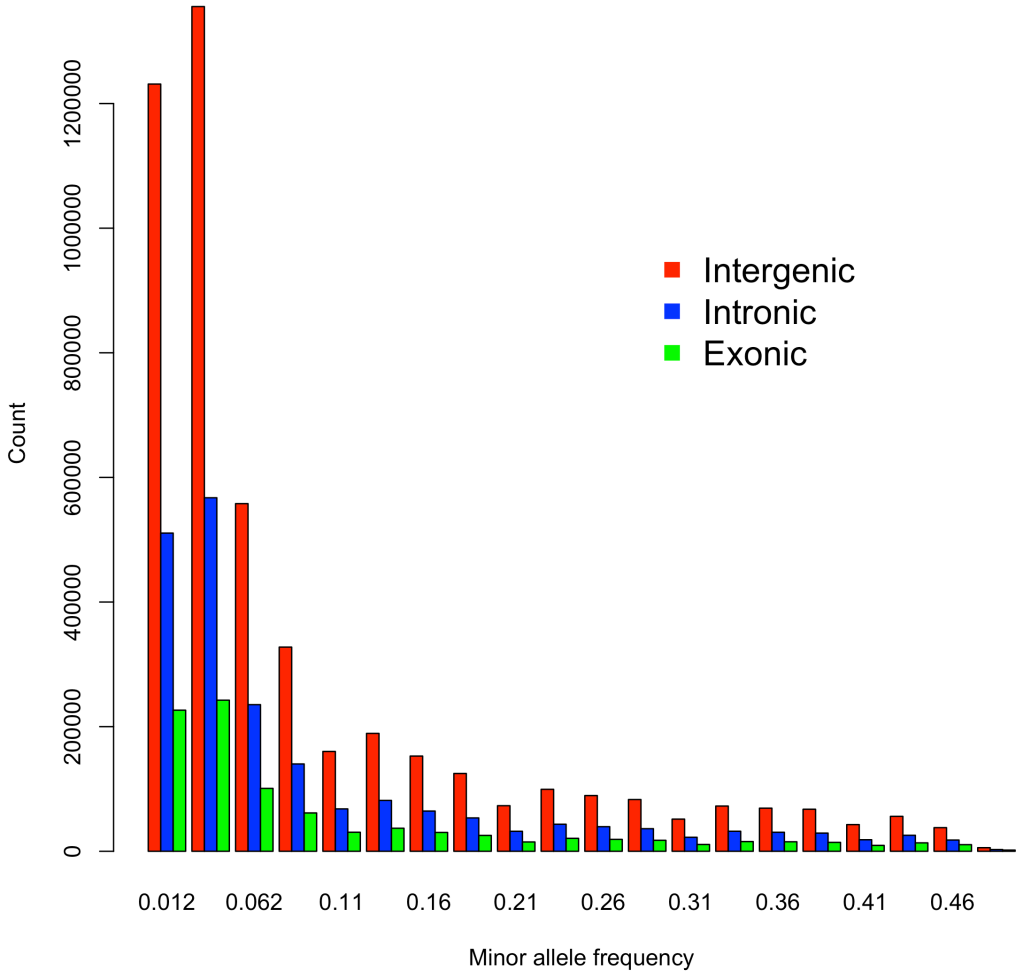
Yassin, A., E. K. Delaney, A. J. Reddiex, T. D. Seher, H. Bastide, N. C. Appleton, J. B. Lack et al. 2016. The pdm3 locus is a hotspot for recurrent evolution of female-limited color dimorphism in *Drosophila*. Current Biology 26:2412-2422.

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 450:233-U232.

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44:821-U136.

821 **Figures and tables:**

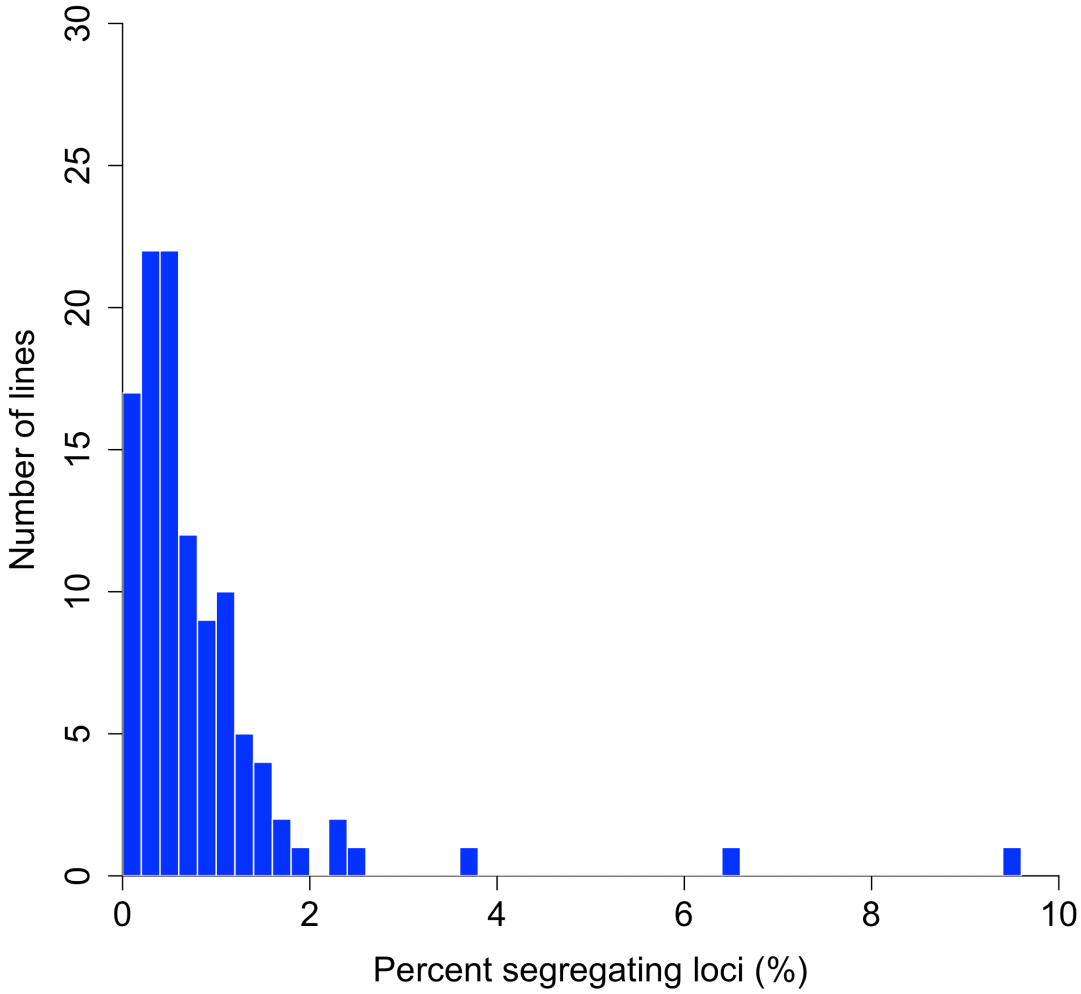
822



823

824 Figure 1: The allele frequency spectrum of SNPs annotated as intergenic (red),
825 intronic (blue), or exonic (green). Singletons (MAF = 0.009) are not shown. We
826 identified 3,748,429 singletons from intergenic regions, 1,535,651 from intronic
827 regions, and 754,075 from exonic regions of the genome.

828



829

830 Figure 2: The distribution of residual heterozygosity as measured by the
831 percentage of total genotyped biallelic SNPs (loci) that were called as
832 heterozygous within each inbred line by the Joint Genotyper of Inbred Lines
833 (JGIL).

834

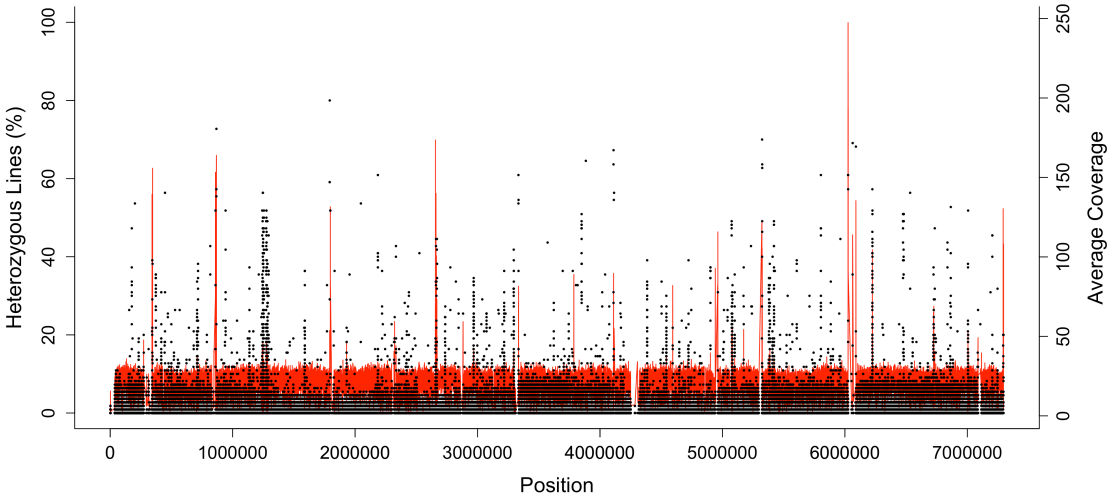
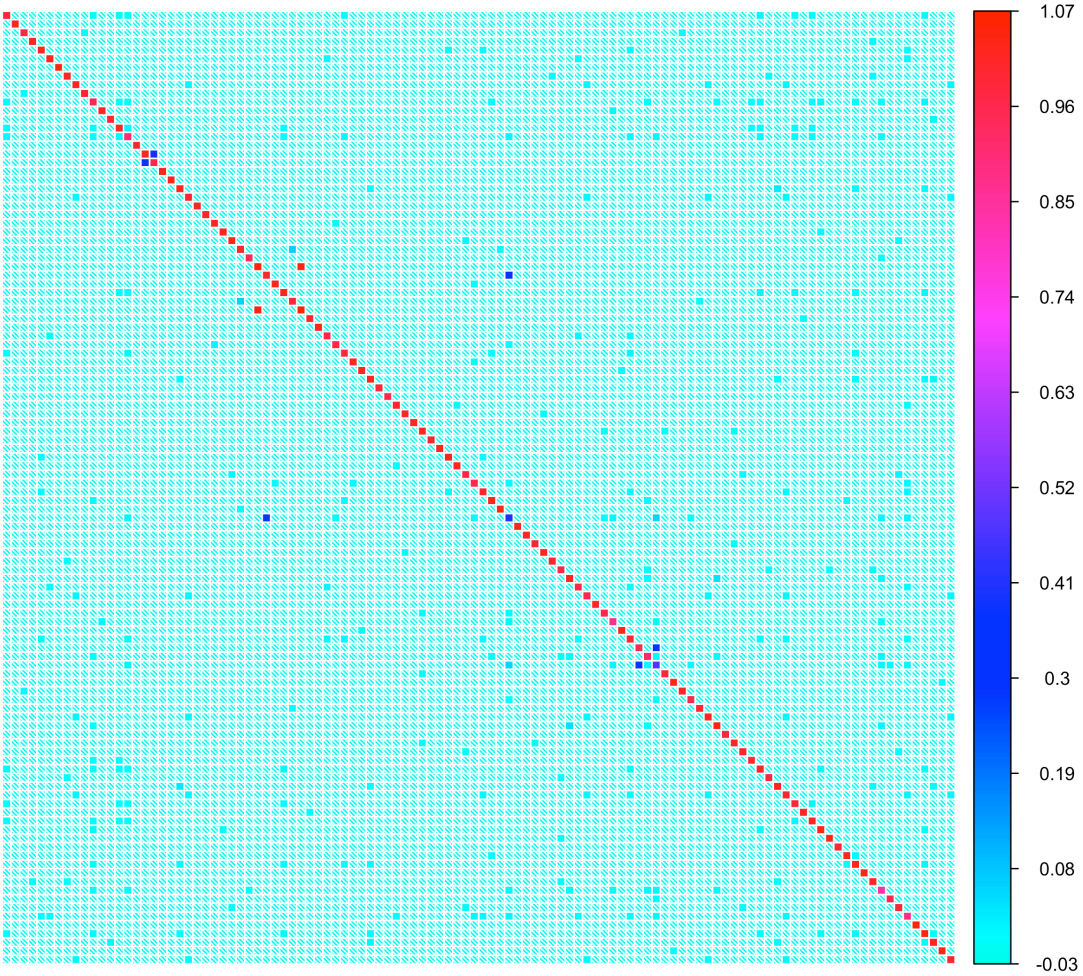


Figure 3: The proportion of lines that are heterozygous at every site along the largest scaffold (scf7180000003208) of the reference genome (black) overlayed with the average read depth at each site shown in red. Peaks in red lines represent potential gene duplications that have collapsed to the same region during assembly.

844



845 Figure 4: Heat-map of the genomic relatedness matrix for each pair of the 110
846 inbred DsGRP lines. Each coloured squared is an estimate of genomic
847 relatedness between a pair of inbred lines estimated by GCTA (Yang et al. 2011).
848 The shade of colour represents the degree of relatedness, with light blue showing
849 low levels of relatedness and red high levels of relatedness.
850

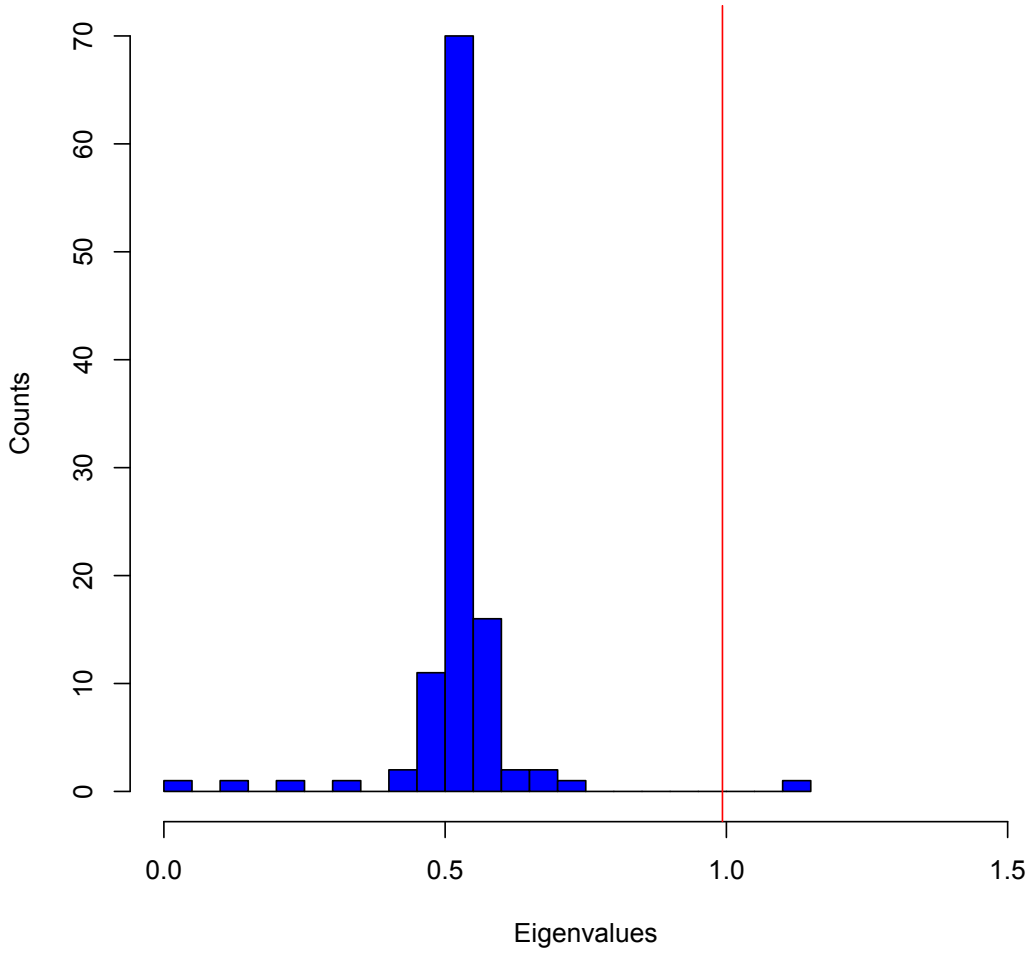
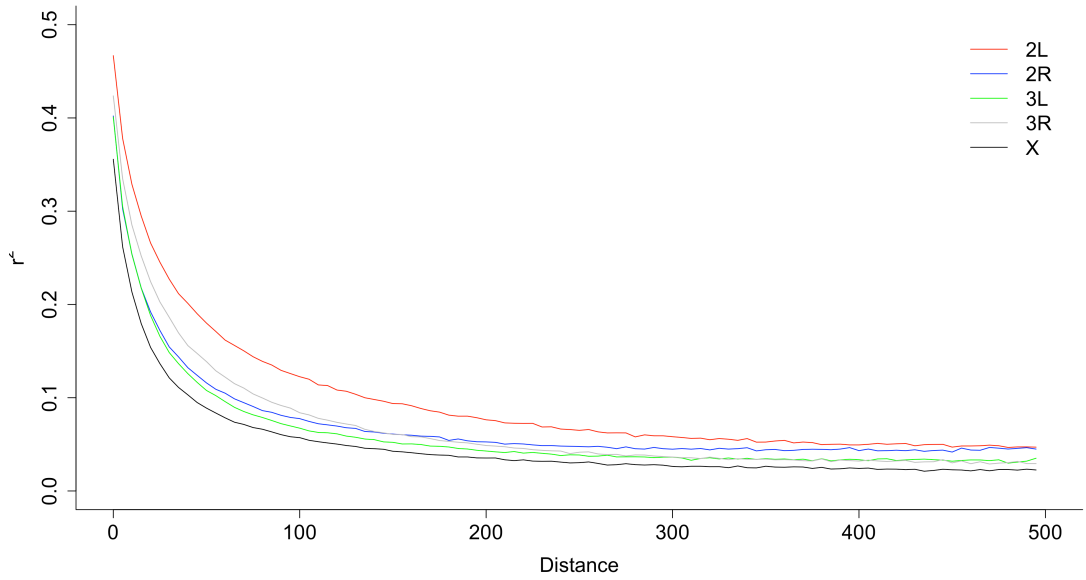


Figure 5: Distribution of eigenvalues from an eigendecomposition of the genomic relatedness matrix for all 110 lines excluding one large eigenvalue where $\Lambda = 20.08$. Eigendecomposition of the genomic relatedness matrix, **X**, which was scaled by equation 2.8 in Bryc et al. (2013) using values 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 $\Lambda_2 = 1.12$.

865



866

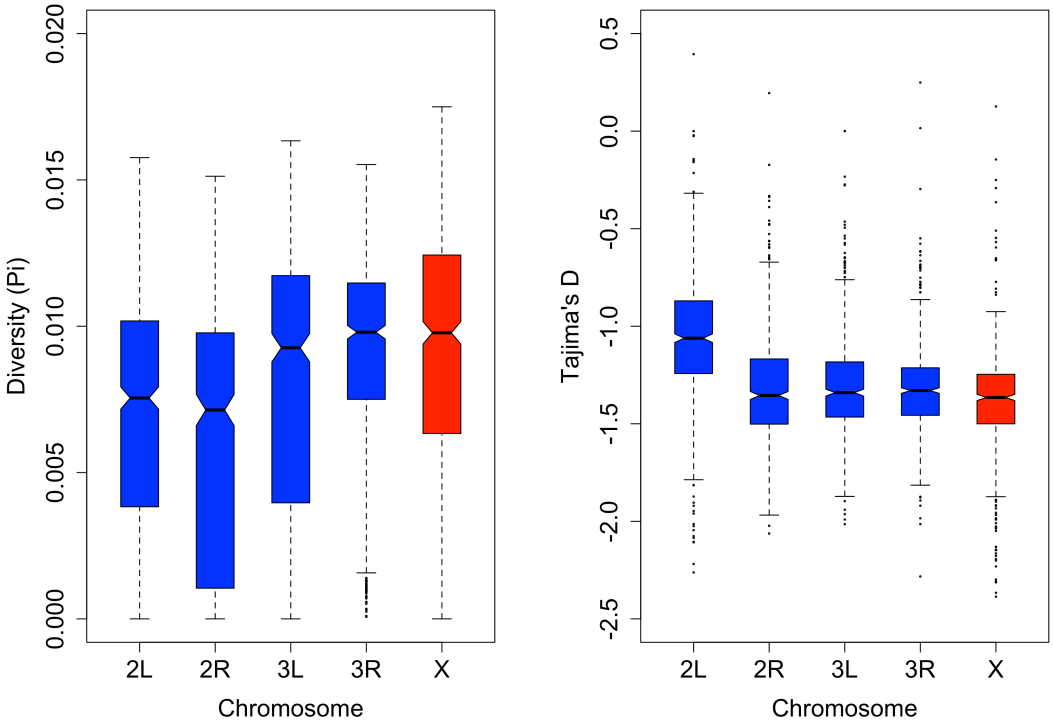
867 Figure 6: Decay of linkage disequilibrium (r^2) between SNPs with genomic

868 distance (bp) in the DsGRP. Values are averaged across each chromosome.

869

870

871

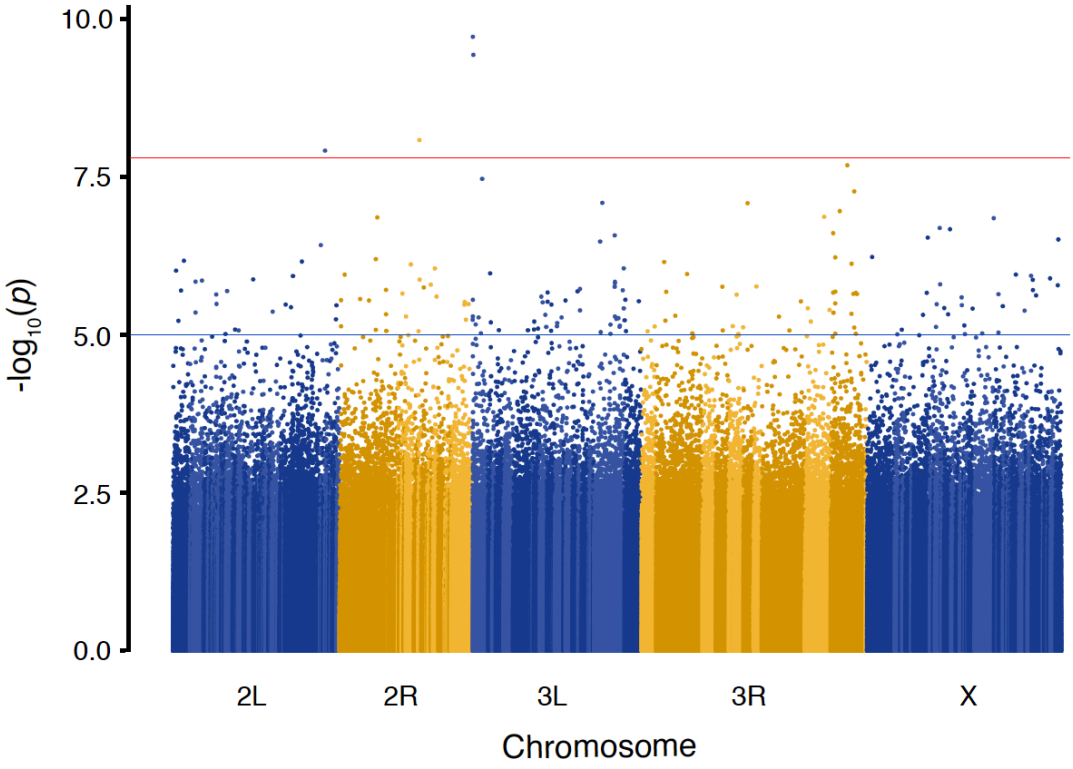


872

873 Figure 7: Boxplots of nucleotide diversity and Tajima's D by chromosome arm.
874 Shown are the estimates from 50 kilobase non-overlapping sliding windows with
875 the breakdown of the number of windows per chromosome as follows: 2L = 691,
876 2R = 664, 3L = 641, 3R = 741, and X = 628.

877

878



879

880

881 Figure 8: Genome-wide association for 2-Me-C_{28} expression in female *Drosophila*
882 *serrata.* Red line indicates Bonferroni threshold corresponding to P = 0.05 and
883 the blue indicates an arbitrary significance threshold of P = 10^{-5}. A total of
884 3,318,503 biallelic SNPs were analyzed with a minimum minor allele frequency
885 of 0.5. Alternating colour shading within chromosomes indicates different
886 contigs in the *D. serrata* genome assembly. Contig order has not yet been
887 established for the genome.

888