Structural variants exhibit allelic heterogeneity and shape variation in complex traits

Mahul Chakraborty[1*], J.J. Emerson[1], Stuart J. Macdonald[2], Anthony D. Long[1*]

[1]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA 92697

[2]Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045

*To whom correspondence should be addressed: mchakrab@uci.edu, tdlong@uci.edu

Abstract

Despite extensive effort to reveal the genetic basis of complex phenotypic variation, studies typically explain only a fraction of trait heritability. It has been hypothesized that individually rare hidden structural variants (SVs) could account for a significant fraction of variation in complex traits. To investigate this hypothesis, we assembled 14 *Drosophila melanogaster* genomes and systematically identified more than 20,000 euchromatic SVs, of which ~40% are invisible to high specificity short read genotyping approaches. SVs are common in *Drosophila* genes, with almost one third of diploid individuals harboring an SV in genes larger than 5kb, and nearly a quarter harboring multiple SVs in genes larger than 10kb. We show that SV alleles are rarer than amino acid polymorphisms, implying that they are more strongly deleterious. A number of functionally important genes harbor previously hidden structural variants that likely affect complex phenotypes (e.g., *Cyp6g1*, *Drsl5*, *Cyp28d1&2*, *InR*, and *Gss1&2*). Furthermore, SVs are overrepresented in quantitative trait locus candidate genes from eight Drosophila Synthetic Population Resource (DSPR) mapping experiments. We conclude that SVs are pervasive in genomes, are frequently present as heterogeneous allelic series, and can act as rare alleles of large effect.

Introduction

Understanding the molecular basis of heritable variation in complex traits is of central importance to evolution, animal and plant breeding, and medical genetics (Mauricio 2001, Goddard and Hayes 2009, Mackay *et al.* 2009, Stranger *et al.* 2011). Over the last decade, short read (50-150 bp) genomic data appropriate for characterizing SNPs and small indels in non-repetitive genomic regions has accumulated at an exponential rate (Shendure and Ji 2008, Bansal *et al.* 2010). This in turn has catalyzed hundreds of quantitative trait loci (QTL) mapping and genome wide association (GWAS) studies in model organisms, humans, and agriculturally important animals and plants (Varshney *et al.* 2009, Davey *et al.* 2011, Day-Williams and Zeggini 2011). Despite these efforts, for most traits, GWAS hits only explain a small fraction of known trait heritability (Manolio *et al.* 2009, Eichler *et al.* 2010). One hypothesis accounting for hidden genetic variation is that individually rare large mutations (>100 bp) that alter genome structure make significant contributions to complex trait variation (Frazer *et al.* 2009, Eichler *et al.* 2010). These structural variants (SVs) change the genome via duplication, deletion, transposition, and inversion of sequences. This hypothesis is attractive since rare causative variants are difficult to detect with GWAS (Spencer *et al.* 2009). Moreover, genotyping approaches based on short reads or microarrays miss a significant number of SVs (Huddleston and Eichler 2016, Chakraborty *et al.* 2018). Finally, SVs are likely to be more deleterious and deleterious more often than SNPs (Emerson *et al.* 2008, Conrad *et al.* 2010, Cridland *et al.* 2013, Rogers *et al.* 2015).

Accurate and gap-free genomes provide a direct and reliable path to comprehensive identification of SVs (Alkan *et al.* 2011, Huddleston *et al.* 2017, Chakraborty *et al.* 2018). To achieve this goal, we assembled reference-quality genomes for fourteen geographically diverse *Drosophila melanogaster* strains (Fig. 1a) using Single Molecule Real Time sequencing (Chakraborty *et al.* 2016). These assemblies are contiguous and complete (N50 18.9-22.3Mb; BUSCO[22] 99.9-100%)(Table 1, Fig. 1b, supplementary Table 1), making them comparable to the *D. melanogaster* reference genome, arguably the best metazoan genome assembly. Thirteen of the fourteen strains are near isogenic founders of the Drosophila Synthetic Population Resources (DSPR) (King *et al.* 2012), a large set of advanced intercross recombinant inbred lines (RILs) designed to map

quantitative trait loci (QTLs) (Long *et al.* 2014). We also assembled the genome of Oregon-R, an outbred stock widely used as a "wild-type" strain both by Drosophila geneticists and by large scale community projects like modENCODE (The modENCODE Consortium *et al.* 2010, Graveley *et al.* 2011, Schwartz *et al.* 2012).

Materials and Methods

DNA extraction

Genomic DNA was extracted from females following the protocols described in Chakraborty *et al.* 2016 and the genomic DNA was sheared using 10 plunges of a 21-gauge needle, followed by 10 of a 24-gauge needle (Jensen Global, Santa Barbara). SMRTbell template library was prepared following the manufacturer's guidelines and sequenced using P6-C4 chemistry in Pacific Biosciences RSII platform at University of California Irvine Genomics High Throughput Facility. The total number of SMRTcell and base pairs sequenced, and read length metrics for each strain is given in supplementary Table 6.

Genome assembly

The genomes were assembled following the approach described in Chakraborty et al. (2016). For all calculations of sequence coverage, a genome size of 130Mbp is assumed ($G = 130 \times 10^6$ bp). For each strain, we generated a hybrid assembly with DBG2OLC (Ye *et al.* 2016) and longest 30X PacBio reads and a PacBio assembly with canu v1.3 (Koren *et al.* 2017) (supplementary Table 5). The paired end Illumina reads were obtained from King et al. (King *et al.* 2012). The hybrid assemblies were merged with the PacBio only assemblies with *quickmerge* v0.2 (Chakraborty *et al.* 2016, Solares *et al.* 2018) (l = 2Mb, ml = 20000, hco = 5.0, c = 1.5), with the hybrid assembly being used as the query. Because the PacBio assembly sizes were closer to the genome size of *D. melanogaster*, we added the contigs that were present only in the PacBio only assembly but not the hybrid assembly by performing a second round of *quickmerge* (Solares *et al.* 2018). For the second round of *quickmerge* (l = 5Mb, ml = 20000, hco = 5.0, c = 1.5), the PacBio assembly was used as the query and the merged assembly from the first merging round the reference assembly. The resulting merged assembly

was processed with *finisherSC* to remove the redundant sequences and additional gap filling using raw reads (Lam *et al.* 2015). The assemblies were then polished twice with *quiver* (SMRTanalysis v2.3.0p5) and once with *Pilon v1.16* (Walker et al. 2014). For *Pilon*, we used the same Illumina reads as used for hybrid assemblies.

Comparative Scaffolding

We scaffolded the contigs for each assembly based on the scaffolds from the reference assembly (Hoskins *et al.* 2015), following a previously described approach (Chakraborty *et al.* 2018). Briefly, TEs and repeats in the assemblies were masked using RepeatMasker (v4.0.7) and aligned to the repeat-masked chromosome arms (X, 2L, 2R, 3L, 3R, and 4) of the *D. melanogaster* ISO1 assembly using MUMmer (Kurtz *et al.* 2004). After filtering of the alignments due to the repeats (delta-filter -1), contigs were assigned to specific chromosome arms on the basis of the mutually best alignment. The scaffolded contigs were joined by 100 Ns, a convention representing assembly gaps. The unscaffolded sequences were named with a 'U' prefix.

BUSCO analysis

We ran BUSCO (v3.02) (Waterhouse *et al.* 2017) on the Pilon polished pre-scaffolding assemblies to evaluate the completeness of all the assemblies relative to the ISO1 release 6 (r6.13) assembly. We used both the arthropoda and diptera datasets for the BUSCO evaluation. For the arhtropoda database, three orthologs (EOG090X0BNZ, EOG090X0M0J, EOG090X049L) weren't found in any of the 15 strains (ISO1, Oregon-R, and 13 DSPR founders). Further inspection of these orthologs revealed that they are present in ISO1 even though the BUSCO analysis misses them when applied to ISO1 (EOG090X0BNZ is CG3223, EOG090X0M0J is Pa1, and EOG090X049L is CG40178). Consequently, we removed these three genes from consideration as uninformative.

Variant detection

For variant detection, we aligned each DSPR assembly individually to the ISO1 release 6 assembly (release 6.13) (Hoskins *et al.* 2015) using nucmer (nucmer –maxmatch –

noextend) (Kurtz *et al.* 2004). We identified and classified the variants using SVMU 0.2beta (Structural variants from MUMmer) (n = 10) (Chakraborty *et al.* 2018). SVMU classifies the structural differences between two assemblies as insertion, deletion, duplication, and inversion based on whether the DSPR assemblies have longer, shorter, more copy, or inverted sequence, respectively, with respect to the reference genome. The variant calls for individual genomes were combined using bedtools merge (Quinlan 2014) and converted into a vcf file using a custom script (https://github.com/mahulchak/dspr-asm). TE insertions were identified by examining the overlap between RepeatMasker identified TEs and SVMU insertion calls using bedtools, requiring that at least 90% of RepeatMasker TE annotation overlap with svmu insertion annotation. 12.8% SV mutations, for which mutation annotation were complicated by secondary mutations, were flagged as 'complex' (CE=2 in the VCF file). Additionally, 16.3% SVs that were located within 5Kb of a complex SV were often part of a complex event and were also assigned a tag (CE=1) to differentiate them from the unambiguously annotated SVs (CE=0).

Genotype validation

To determine the genotyping error rate, a set of randomly selected 50 simple (CE=0) SVs obtained from SVMU were manually inspected on UCSC genome browser representation of the multiple genome alignment of the 15 genomes (http://goo.gl/LLpoNH). Furthermore, to estimate the genotyping accuracy of the SVs occurring in the vicinity of the complex mutations, where mutation annotation is complicated by alignment ambiguities, we manually inspected 217 SVs occurring within 20Kb of 50 randomly selected complex (CE=2) SVs. Among these, 3/217 and 0/50 SVs were absent in the UCSC browser and therefore they are likely mis-annotated by our pipeline. The mis-annotated SVs (insertion in A1 and tandem array CNV in A7) are located in a complex, repetitive, structurally variable genomic region on chromsome 3L (3L:7669500-7679100) (supplementary Fig.5).

Comparing SV genotypes from de novo assemblies to short read only calls

TE genotypes for the founders (Cridland *et al.* 2013) were downloaded from flyrils.org and the insertion coordinates were lifted over to the current release (release 6) of the

reference genome (Hoskins *et al.* 2015) using UCSC liftover tool (Kent *et al.* 2002). For detection of the duplicates, we have previously found that discordant read pair based method (Pecnv) (Rogers *et al.* 2014) was comparable to split read mapping (Ye *et al.* 2009) and more reliable than methods based on coverage alone (Abyzov *et al.* 2011, Chakraborty *et al.* 2018), so we used Pecnv. Pecnv was run using the settings described before (Chakraborty *et al.* 2018). Because svmu reports tandem duplicate CNVs as insertions (with appropriate CNV tags to separate from TE and other insertions) and Pecnv reports sequence range being duplicated, the SVMU CNV insertion coordinates were extended by 100 bp before comparison (bedtools intersect) between Pecnv output and svmu output was conducted. The non-TE indel genotypes were obtained from Pindel output (the “LI” and “D” events) using the commands described previously[15]. For determining population frequency of indel SVs (e.g. the reference FB element in *InR*), Pindel output based on the alignment bam files were used. We only estimate the false negative rate of short read only callers, but note that these methods also generate false positive SV calls.

Gene expression analysis

The preprocessed expression data for female heads (King *et al.* 2014) and IIS/TOR expression data (Stanley *et al.* 2017) from whole bodies were downloaded from www.flyrils.org. Expression QTL analysis (supplementary Fig. 11) for *Cyp28d1* and *Gss1* using the head expression data were performed using the R package DSPRqtl following the instructions provided in the manual (DSPRscan,model = gene ~ 1,design = “ABcross”). When expression data for multiple isoforms were present, expression data only for the longest transcript that is expressed in the head was used. The genotype values at the eQTL were determined using the function DSPRpeaks included in the DSPRqtl package. No eQTL were found for *InR* so the genotype values for the *InR* expression data were obtained by assigning the founder genotypes to the RILs used in the IIS/TOR expression dataset, using the posterior probabilities of the forward-backward decoding of the HMM for the panel B RILs available on www.flyrils.org. *Drsl5* expression levels in A4 and A3 were obtained from a publicly available RNAseq dataset (Marriage *et al.* 2014).

Comparison of site frequency spectra

The histogram of allele frequencies (site frequency spectrum or SFS) was collated for four categories: synonymous SNPs, non-synonymous SNPs, duplicate CNVs, and TE insertions. The frequencies of SNPs were collected from the VCF file (King *et al.* 2012) using vcftools and bcftools (Danecek *et al.* 2011, Li 2011). The frequencies of SVs were collected from the column 4 of the combined SVMU output for the TE insertions and duplication CNVs from all DSPR strains (https://github.com/mahulchak/dspr-asm). Complex mutations (CE=1 and CE=2) were excluded from the analysis. Let N be the sample size and x_i be the number of sites in frequency class i, where $0 < i < N$. The SFS was "folded", meaning we focused attention on the minor allele frequency (MAF), or y_i = minimum(x_i, N-x_i). Pairwise comparisons between different SFS site categories were conducted using the χ^2 test on allele frequencies and site categories. For allele frequencies, two types of classifications were used: 1) every y_i for $0 < i < N$ (N-1 df); and 2) considering singletons versus the other frequency categories, or y_i for i = 1 versus $2 < i < N$ (1 df).

Candidate genes associated with mapped QTL

The candidate genes from DSPR QTL papers were selected based the following criteria: 1) The gene falls within the QTL peak; 2) additional functional data is cited by the authors of the respective study to highlight the gene; 3) the functional information cited by the authors did not use knowledge about structural variation affecting the candidate locus (supplementary Table 3). The additional data can either be expression data collected by the authors or existing functional data known about the genes. Only 44 candidate genes from 8 studies fulfilled these criteria but 3 among these fell outside the euchromatic boundaries used here (supplementary Table 1). Hence only 41 candidate genes were included in the SV enrichment analysis. Of the 41 candidate genes identified, 10 of them were at a single locus (*GstE1-10*). As a result, we carry out our analysis treating *GstE1-10* as either a single gene or ten different genes (the qualitative outcome is unchanged). To test if candidate genes are longer than average genes, we considered all genes (supplementary table 4) as well as the dataset

excluding the GstE1-10 genes (supplementary Table 4). The lengths of candidate genes were compared against the rest of the genome using a Mann-Whitney U test.

Candidate gene enrichment analysis

To determine if candidate genes are enriched for SVs relative to the rest of the genome, we analyzed the dataset both without merging and with merging the GstE1-10 into a single 13kb SV-burdened locus (supplementary Table 4). A Fisher's Exact Test was applied to the counts in categories of candidate gene vs. rest of the genome and SV-burdened vs. SV-free genes. To account for the lengths of the candidate genes being longer than the rest of the genome, we performed a Monte Carlo resampling of the whole genome according to the histogram of gene sizes in the candidate gene lists (supplementary Table 4). We sampled from the genome by drawing from each gene length bin with a hypergeometric distribution, where n is the number of candidate genes in the candidate bin, K is the number of SV-burdened genes in the genome bin, and N-K is the number of SV-free genes in the genome bin (supplementary Table 4). We then tallied up the number of observed SVs. We repeated this 100,000 times to construct a Monte Carlo distribution of the SV burden expected of genes matching the size distribution observed in the actual candidate genes. This led to simulated size distributions that matched the observed size distributions (every Mann-Whitney U p-value of Monte Carlo sample lengths compared against the observed candidate lengths > 0.1).

Calculating the SV burden in genes in diploid individuals

In order to calculate the distribution of SV burden expected in diploids, the haploid genotypes of each founder was paired with every other founder, for a total of 78 possible pairings. For each of these diploid pairings, the number of unique SV mutations for each gene in the genome was recorded. A mutation is said to affect a gene if it falls within the gene span, which is defined as affecting nucleotides between the start and end coordinates of the gene feature in the *Drosophila melanogaster* release 6.16 gff file (dos Santos *et al.* 2015). The number of SV mutations overlapping a gene in a given diploid combination is considered that gene's multiplicity for that combination. Any gene

with a multiplicity ≥ 1 for a particular diploid comparison is considered SV-burdened for that diploid.

Results and Discussion

De novo assembly reveals a large number of previously hidden functionally important SVs:

Our assemblies are extremely contiguous, with the majority of each chromosome arm represented by a single contig (Fig. 1b). We also close the two remaining gaps in the major chromosome arms of the euchromatic *D. melanogaster* reference genome (Chang and Larracuente 2018) in all our assemblies (supplementary Fig.1-3). We identified SVs by comparing each assembly to the reference ISO1 genome (Chakraborty *et al.* 2018), focusing our attention on large (>100bp) euchromatic SVs (supplementary Table 2), and ignoring heterochromatin regions as they are gene poor (Smith *et al.* 2007) and require specialized approaches and extensive validation (Khost *et al.* 2017). Manual inspection of 267 randomly sampled SVs indicate that mis-annotations are rare (3/267), and occur in ambiguously aligned structurally complex genomic regions (supplementary Fig. 5; see Methods). We discovered 7,347 TE insertions, 1,178 duplication CNVs, 4347 indels, and 62 inversions in the 94.5 Mb of euchromatin spanning the five major chromosome arms across the DSPR founders (Fig. 1c-d). Each founder strain exhibits 637 TE insertions, 134 duplications, 694 indels, and 7 inversions on average (Table 2).

A large fraction of the SVs (36% of non-reference TEs, 26% of deletions, 48% of insertions, 60% of duplication CNVs) present in the assemblies eluded detection using high specificity SV genotyping methods (Chakraborty *et al.* 2018) employing high coverage paired end Illumina reads (supplementary Fig. 6), Some of these novel events are likely to affect phenotypes. For example, extensive evidence links complex SV alleles of the cytochrome P450 gene *Cyp6g1* to varying levels of DDT resistance (Daborn *et al.* 2002, Schmidt *et al.* 2010). Despite extensive study of this locus, we discovered three new SV alleles involving TE insertions that likely have different functional consequences (supplementary Fig. 7a-b). Similarly, we discovered a previously hidden tandem duplication of the antifungal, innate immunity gene *Drsl5*

(Yang et al. 2006) that exhibits >1000 fold higher expression relative to its single copy counterpart in line A4(supplementary Fig. 8a-b). Read pair orientation methods failed to detect this mutation because one allele bears a 5kb spacer sequence derived from the first exon and intron of *Kst* inserted between the gene copies (supplementary Fig. 8a). Another duplicate allele of *Drsl5* contains a *Tirant* LTR retrotransposon inserted into the same spacer sequence (supplementary Fig. 8a). We also easily detect the two SV mutations underlying the *D. melanogaster* recessive visible genes *cinnabar* (Warren et al. 1996) (*cn*) and *speck* (*sp*) present in the ISO1 reference genome (dos Santos *et al.* 2015) (supplementary Fig. 9-10). In the case of *sp* a large insertion in the reference genome is mis-annotated as an intron. For *cn* a large exonic deletion is not identified as such (dos Santos *et al.* 2015). Both alleles are likely knock-outs.

SVs are deleterious:

Most TEs and duplicates are present in only one strain (Fig. 1e), indicating that purifying natural selection has prevented them from rising to higher frequencies. The folded site frequency spectrum (SFS) of the TEs and CNVs exhibit more rare variants than synonymous and non-synonymous SNPs (Fig. 1e; p-value < 1e-10, $\chi 2$ test between frequency classes of SVs and non-synonymous SNPs), suggesting that SVs are on average under purifying selection (Emerson *et al.* 2008, Cridland *et al.* 2013, Rogers *et al.* 2015). Amongst SVs, TEs are more enriched for rare variants than duplicates, indicating that TE insertions are on average more deleterious than CNVs (Fig. 1e; p-value < 1e-10, $\chi 2$ test between frequency classes of TEs and CNVs). Under mutation selection balance models (Pritchard 2001, Thornton *et al.* 2013), rare deleterious variants (minor allele frequency or MAF <1%) are predicted to contribute significantly to complex trait variation (Gibson 2012), yet are unlikely to be tagged by SNPs typically used in GWAS experiments (Manolio *et al.* 2009). Consequently, if individually rare SVs underlie complex trait variation, they will often go undetected in association studies (Spencer *et al.* 2009).

Functional structural variation at mapped QTL:

Segregation of multiple alleles at a causal genes (ie allelic heterogeneity) can mislead discovery of causative loci in GWAS experiments (Thornton *et al.* 2013), though such

genes can be readily identified in multi-parent panels (MPPs) via QTL mapping (Long *et al.* 2014). However, mapping resolution is often poor, thwarting the identification of causative mutations bearing a variant in a candidate gene of obvious functional significance. Yet, putatively causative SVs are often hidden as they disproportionately escape detection by short read sequencing (Chakraborty *et al.* 2018). This limitation can be solved in the DSPR and other MPPs, as *de novo* assemblies of the founders of the MPP allow genotypes at SVs to be imputed for the lines on which phenotypes are measured (King *et al.* 2012).

A nicotine resistance mapping study employing the DSPR identified differentially expressed cytochrome P450 genes *Cyp28d1* and *Cyp28d2* as candidate causative genes at a mapped QTL, but proposed no causative mutations (Marriage *et al.* 2014). A previous *de novo* assembly of one DSPR founder strain assembled a resistant allele possessing tandem copies of the *Cyp28d1* gene separated by an *Accord* LTR retrotransposon fragment (Chakraborty *et al.* 2018) (Fig. 2a; supplementary Fig. 11). Our assemblies of the DSPR founder strains revealed a total of seven structurally distinct alleles in this region, including additional candidate resistant alleles harboring gene duplications (Fig. 2a-b). For example, the resistant strain A2 carries a tandem duplication of a 15Kb segment containing both *Cyp28d* genes. The expression level of *Cyp28d1* in the adult female heads of RILs bearing the A2 genotype is highest among all founder genotypes measured (Fig. 2c). Consistent with this, DSPR RILs bearing the A2 genotype show the highest resistance to nicotine toxicity among the A genotype RILs (Marriage *et al.* 2014) (Fig. 2b). This implies that the extra copies of *Cyp28d1* and/or *Cyp28d2* account for the increased expression and concomitant resistance to nicotine. Similarly, the B4 allele comprises a tandem duplication of a 6 Kb segment, containing one extra copy of *Cyp28d1* and a nearly complete copy of *Cyp28d2* (Fig. 2a; supplementary Fig. 11). RILs carrying the B4 genotype at the *Cyp28d* locus also show high resistance to nicotine, making the duplication a compelling candidate for the causative mutation. On the other hand, in two alleles, TE insertions disrupt *Cyp28d* gene structure and function. For instance, A1 has the same duplication as B4, but a 4.7Kb F element inserted in the 5th exon disrupts the protein coding sequence of the second *Cyp28d1* copy, likely rendering the copy nonfunctional (Fig. 2a; supplementary

Fig. 11). Consistent with the hypothesis that the duplication causes increased nicotine resistance, the A1 genotype is more susceptible to nicotine than B4 (Fig. 2b). All of these SV alleles are singletons, and thus represent a hidden allelic series composed of individually rare alleles.

SVs may also affect genes central to life history traits. Expression levels of the insulin signaling pathway genes show substantial variation in F1 hybrids between DSPR panel B RILs and the A4 founder (Stanley *et al.* 2017). Among these is *Insulin Receptor* (*InR*), which plays key roles in several life history traits related to lifespan and is likely a key molecular mediator of the tradeoff between reproductive success and longevity (Tatar *et al.* 2001, Toivonen and Partridge 2009, Paaby *et al.* 2014). Amino acid polymorphism in *InR* evolves under positive selection and some non-synonymous variants affect fecundity and stress response (Paaby *et al.* 2010, Paaby *et al.* 2014). Expression variation of *InR* also affects body size, lifespan, and fecundity (Brogiolo *et al.* 2001, Rauschenbach *et al.* 2015), suggesting that natural cis-regulatory variation might also be under selection. We discovered a 215 bp fragment of a DOC6 element within a 2nd intron enhancer (Wei *et al.* 2016) (Fig. 3b-c) of *InR* on the AB8 haplotype, and this allele exhibits reduced gene expression relative to reference genotypes (Fig. 3b). This mutation presumably disrupts the enhancer (supplementary Fig. 12), making it a plausible candidate for expression variation in *InR*. Another founder, A6, carries a 1,042 bp insertion of DMRT1A (LINE) in the 2nd intron and a 946 bp insertion of a fragment of PROTOP in the 3rd intron. Both affect known cis-regulatory elements (Wei *et al.* 2016) (Fig. 3b-c). Except for A2 and A6, all strains, including ISO1, harbor an FB-NOF element (FB{}1698) inside the first intron of *InR* (Fig. 3a). Like many genes, the first intron of *InR* possess several transcription factor binding sites (TFBS), including those for factors *Nejire* and *Caudal* (Negre *et al.* 2011) (Fig. 3c). The FB-NOF element is inserted within this dense cluster of TFBS and active enhancer marks (Fig.3c). Furthermore, the FB element is segregating at high frequency in the strains discussed here (13/15), a North American population (Mackay *et al.* 2012) (125/170), and a French population (Pool *et al.* 2012) (4/9), but is rare in populations derived from *D. melanogaster's* ancestral range in Africa (Pool *et al.* 2012, Lack *et al.* 2015) (Cameroon: 0/10, Rwanda: 1/27, Zambia: 10/139) (Fig. 3d). This raises the possibility that the FB

element is more common in temperate cosmopolitan populations, similar to a previously described adaptive amino acid variant in *InR* (Paaby et al. 2010). In total, *InR* harbors a remarkable amount of potentially functional structural diversity; Including these variants described, there are 9 TE insertions and two deletions throughout the gene, many of which impinge on candidate regulatory regions or transcribed portions of the gene (Fig. 3a,3c).

Public resources like modENCODE annotate molecular phenotypes (e.g., RNAseq, ChIPseq, DNAse1HSseq) against reference genomes which are often genetically different than the strains assayed (The modENCODE Consortium *et al.* 2010, Graveley *et al.* 2011, Negre *et al.* 2011, Schwartz *et al.* 2012). Canton-S (our DSPR founder A1) and Oregon-R are strains commonly used in phenotypic assays (The modENCODE Consortium *et al.* 2010, Graveley *et al.* 2011, Schwartz *et al.* 2012), and we observe SVs segregating between these two strains and the reference (Table 2). Interpretation of functional genomics data such as RNA-seq can be misleading when gene copy number varies between strains. We explored the glutathione synthetase region (containing *Gss1* and *Gss2*), which is just one example among hundreds in modENCODE that likely suffer from misleading annotations. A tandem duplication present in ISO1 has created two copies of *Gss1* and *Gss2*, which are associated with toxin metabolism and linked to tolerance to arsenic (Ortiz *et al.* 2009) and ethanol induced oxidative stress (Logan-Garbisch *et al.* 2015). While this duplication segregates at high frequency in DSPR strains (9/13), it is absent in Oregon-R (Fig. 4a) and escapes detection with high specificity short-read methods. As a result, using transcript and ChIP data derived from Oregon-R (as used in modENCODE (Graveley *et al.* 2011, Schwartz *et al.* 2012)) results in misleading annotations of the two copies in ISO1. Indeed, among the eight structurally distinct Gs alleles in our dataset, ISO1 is the sole representative of its allele (Fig. 4a). The two most common Gs alleles include one that contains only a single Gs gene (in four strains, including Oregon-R) and one carrying only a tandem duplication, creating the Gss1/Gss2 pair (in five strains, including Samarkand/AB8) (Fig. 4a). The remaining 6 alleles have SV genotypes represented by only a single individual in the sample. Collectively, this sample represents a haplotype network of structural variation involving 5 TE insertions, one duplication, one insertion comprising TE and

simple repeats, and two non-TE indels. The single copy allele with a 5' insertion of a 14kb repetitive sequence comprising *Nomad* retrotransposon fragments exhibits the highest expression, followed by duplicate alleles, whereas single copy alleles and duplicate alleles with intronic TE insertions generally have the lowest expression levels (Fig. 4b).

Although hypotheses employing SVs to explain missing heritability have been proposed (Manolio *et al.* 2009, Sudmant *et al.* 2015), the systematic under-identification of SVs via short read- and microarray-based genotyping (Alkan *et al.* 2011) limits their explanatory power. Using our comprehensive SV map, we measured the prevalence of SVs at the candidate genes reported in 8 complex trait mapping experiments employing DSPR (supplementary Table 3). We consider only genes in mapped QTLs explicitly cited by the authors of the original QTL studies (supplementary Table 3; see Methods). In total, we identified 31 candidate genes and a single, tandem array of 10 small genes from the same family (GstE1-10) (Kislukhin *et al.* 2013), which we consider as a single additional locus with structural variation. Half of these candidates (16/32) possess at least one SV in one founder strain, whereas only 23.4% (3,252/13,861) of all D. melanogaster genes harbor SVs ($p = 0.001$, Fisher's exact test). Although the candidates we tested (supplementary Table 3) are approximately twice as large as the genome-wide average (supplementary Table 4; $p = 6.5\times10\text{-}5$), this enrichment of SVs at candidate genes is not merely a consequence of them being longer. SVs are enriched in 100,000 Monte Carlo samples matching the candidate gene length distribution ($p = 0.021$; Fig. 5a). These results persist when the GstE genes are considered individually instead of being merged (length p-value = 0.034 and enrichment p-value = $2.9\times10\text{-}3$; supplementary Table 4).

In order to illustrate how common SVs are in the genome we quantified the per gene SV burden per diploid *D. melanogaster* individual (Fig. 5b). Across the genome, SVs appear in 9.3% of genes in diploid individuals. Of those, more than a third (34.4%) involve multiple SV mutations (Fig. 5c). One or more SVs burden more than half of genes in and above the 20-35kb range (Fig 5b). Furthermore, individual genes bearing multiple SVs comprise more than a third of burdened genes between 20kb and 35kb in

length and more than half of larger genes. These observations suggest that the contribution of rare SVs of large effect to complex traits could be pervasive.

Conclusion

Despite claims that a significant proportion of complex trait variation in humans, model organisms, and agriculturally important animals and plants are likely due to rare SVs of large effect (Eichler *et al.* 2010), systematic inquiry of this hypothesis has been impeded by genotyping approaches attuned to SNP detection (Alkan *et al.* 2011). As reference quality *de novo* assemblies of population samples for eukaryotic model systems become increasingly cost-effective, methodical evaluation of the contribution of SVs to the genetic architecture of complex traits becomes feasible. Our comprehensive map of SVs in *Drosophila* provides the means to systematically quantify the contribution of rare SVs to heritable complex trait variation (Fig 2a,3a,4a,5a). The value of comprehensive SV detection is underscored by the presence of SVs in ~50% of the candidate genes underlying mapped Drosophila QTL, and by the observation that a large fraction of Drosophila genes harbor multiple rare SV alleles. The genomes of humans and agriculturally important plants and animals harbor more SVs than Drosophila, and thus are likely more burdened with genic SVs.

The genetic heterogeneity hypothesis posits that a sizable fraction of human complex disease is associated with an allelic series consisting of individually rare causative mutations at several genes of large effect (McClellan and King 2010). Furthermore, models for complex traits under either stabilizing (Turelli 1984, Johnson and Barton 2005) or purifying selection (Pritchard 2001, Thornton *et al.* 2013) with constant mutational input predict the existence of genes segregating several individually rare causative alleles that account for a sizable fraction of complete trait variation. We provide examples of SVs in genes of functional significance, and show that genes harboring SVs are overrepresented in a collection of QTL candidate genes. Hidden SVs are thus examples of collectively common but individually rare deleterious genetic variants predicted under the genetic heterogeneity hypothesis. Future *de novo* assemblies of other genomes, including humans, models, and agriculturally important species, would quantify the generality of observations from *Drosophila*.

References

Abyzov, A., A. E. Urban, M. Snyder and M. Gerstein (2011). "CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." Genome Research **21**(6): 974-984.

Alkan, C., B. P. Coe and E. E. Eichler (2011). "Genome structural variation discovery and genotyping." Nature Reviews Genetics **12**(5): 363-376.

Bansal, V., O. Harismendy, R. Tewhey, S. S. Murray, N. J. Schork, E. J. Topol and K. A. Frazer (2010). "Accurate detection and genotyping of SNPs utilizing population sequencing data." Genome Research **20**(4): 537-545.

Brogiolo, W., H. Stocker, T. Ikeya, F. Rintelen, R. Fernandez and E. Hafen (2001). "An evolutionarily conserved function of the Drosophila insulin receptor and insulin-like peptides in growth control." Current Biology **11**(4): 213-221.

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long and J. J. Emerson (2016). "Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage." Nucleic Acids Research **44**(19): e147.

Chakraborty, M., N. W. VanKuren, R. Zhao, X. Zhang, S. Kalsow and J. J. Emerson (2018). "Hidden genetic variation shapes the structure of functional elements in Drosophila." Nature Genetics **50**(1): 20-25.

Chang, C.-H. and A. M. Larracuente (2018). "Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome." bioRxiv.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. J. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, M. E. Hurles and W. T. C. Control (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-712.

Cridland, J. M., S. J. Macdonald, A. D. Long and K. R. Thornton (2013). "Abundance and distribution of transposable elements in two Drosophila QTL mapping resources." Mol Biol Evol **30**(10): 2311-2327.

Daborn, P. J., J. L. Yen, M. R. Bogwitz, G. Le Goff, E. Feil, S. Jeffers, N. Tijet, T. Perry, D. Heckel, P. Batterham, R. Feyereisen, T. G. Wilson and R. H. ffrench-Constant (2002). "A single P450 allele associated with insecticide resistance in Drosophila." Science **297**(5590): 2253-2256.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. P. A. Grp (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter (2011). "Genome-wide genetic marker discovery and genotyping using next-generation sequencing." Nature Reviews Genetics **12**(7): 499-510.

Day-Williams, A. G. and E. Zeggini (2011). "The effect of next-generation sequencing technology on complex trait research." European Journal of Clinical Investigation **41**(5): 561-567.

dos Santos, G., A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, W. M. Gelbart and C. FlyBase (2015). "FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations." Nucleic Acids Research **43**(Database issue): D690-697.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore and J. H. Nadeau (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nature Reviews Genetics **11**(6): 446-450.

Emerson, J. J., M. Cardoso-Moreira, J. O. Borevitz and M. Long (2008). "Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster." Science **320**(5883): 1629-1631.

Frazer, K. A., S. S. Murray, N. J. Schork and E. J. Topol (2009). "Human genetic variation and its contribution to complex traits." Nature Reviews Genetics **10**(4): 241-251.

Gibson, G. (2012). "Rare and common variants: twenty arguments." Nature Reviews Genetics **13**(2): 135-145.

Goddard, M. E. and B. J. Hayes (2009). "Mapping genes for complex traits in domestic animals and their use in breeding programmes." Nature Reviews Genetics **10**(6): 381-391.

Graveley, B. R., A. N. Brooks, J. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. H. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Y. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. C. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver and S. E. Celniker

(2011). "The developmental transcriptome of Drosophila melanogaster." Nature **471**(7339): 473-479.

Hoskins, R. A., J. W. Carlson, K. H. Wan, S. Park, I. Mendez, S. E. Galle, B. W. Booth, B. D. Pfeiffer, R. A. George, R. Svirskas, M. Krzywinski, J. Schein, M. C. Accardo, E. Damia, G. Messina, M. Mendez-Lago, B. de Pablos, O. V. Demakova, E. N. Andreyeva, L. V. Boldyreva, M. Marra, A. B. Carvalho, P. Dimitri, A. Villasante, I. F. Zhimulev, G. M. Rubin, G. H. Karpen and S. E. Celniker (2015). "The Release 6 reference sequence of the Drosophila melanogaster genome." Genome research **25**(3): 445-458.

Huddleston, J., M. J. P. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C. S. Chin, J. Korlach, R. K. Wilson and E. E. Eichler (2017). "Discovery and genotyping of structural variation from long-read haploid genome sequence data." Genome Research **27**(5): 677-685.

Huddleston, J. and E. E. Eichler (2016). "An Incomplete Understanding of Human Genetic Variation." Genetics **202**(4): 1251-1254.

Johnson, T. and N. Barton (2005). "Theoretical models of selection and mutation on quantitative traits." Philosophical Transactions of the Royal Society B-Biological Sciences **360**(1459): 1411-1425.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome Research **12**(6): 996-1006.

Khost, D. E., D. G. Eickbush and A. M. Larracuente (2017). "Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster." Genome Research **27**(5): 709-721.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman, A. D. Long and S. J. Macdonald (2012). "Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource." Genome Research **22**(8): 1558-1566.

King, E. G., B. J. Sanderson, C. L. McNeil, A. D. Long and S. J. Macdonald (2014). "Genetic dissection of the Drosophila melanogaster female head transcriptome reveals widespread allelic heterogeneity." PLoS Genetics **10**(5): e1004322.

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**(8): 1301-1310.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman and A. M. Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome Research **27**(5): 722-736.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S. L. Salzberg (2004). "Versatile and open software for comparing large genomes." Genome Biology **5**(2): R12.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley and J. E. Pool (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.

Lam, K. K., K. LaButti, A. Khalak and D. Tse (2015). "FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads." Bioinformatics **31**(19): 3207-3209.

Li, H. (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." Bioinformatics **27**(21): 2987-2993.

Logan-Garbisch, T., A. Bortolazzo, P. Luu, A. Ford, D. Do, P. Khodabakhshi and R. L. French (2015). "Developmental Ethanol Exposure Leads to Dysregulation of Lipid Metabolism and Oxidative Stress in Drosophila." G3-Genes Genomes Genetics **5**(1): 49-59.

Long, A. D., S. J. Macdonald and E. G. King (2014). "Dissecting complex traits using the Drosophila Synthetic Population Resource." Trends in Genetics **30**(11): 488-495.

Mackay, T. F., E. A. Stone and J. F. Ayroles (2009). "The genetics of quantitative traits: challenges and prospects." Nature Reviews Genetics **10**(8): 565-577.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. H. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barron, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L. L. Pu, C. Qu, M. Ramia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y. Q. Wu, A. Yamamoto, Y. M. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman and R. A. Gibbs (2012). "The Drosophila melanogaster Genetic Reference Panel." Nature **482**(7384): 173-178.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A.

Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

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.

Mauricio, R. (2001). "Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology." Nature Reviews Genetics **2**(5): 370-381.

McClellan, J. and M. C. King (2010). "Genetic Heterogeneity in Human Disease." Cell **141**(2): 210-217.

Negre, N., C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A. Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis and K. P. White (2011). "A cis-regulatory map of the Drosophila genome." Nature **471**(7339): 527-531.

Ortiz, J. G. M., R. Opoka, D. Kane and I. L. Cartwright (2009). "Investigating Arsenic Susceptibility from a Genetic Perspective in Drosophila Reveals a Key Role for Glutathione Synthetase." Toxicological Sciences **107**(2): 416-426.

Paaby, A. B., A. O. Bergland, E. L. Behrman and P. S. Schmidt (2014). "A highly pleiotropic amino acid polymorphism in the Drosophila insulin receptor contributes to life-history adaptation." Evolution **68**(12): 3395-3409.

Paaby, A. B., M. J. Blacket, A. A. Hoffmann and P. S. Schmidt (2010). "Identification of a candidate adaptive polymorphism for Drosophila life history by parallel independent clines on two continents." Molecular Ecology **19**(4): 760-774.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. J. Emerson, P. Saelao, D. J. Begun and C. H. Langley (2012). "Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture." PLoS Genetics **8**(12): e1003080.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" American Journal of Human Genetics **69**(1): 124-137.

Quinlan, A. R. (2014). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." Current Protocols in Bioinformatics **47**: 11 12 11-34.

Rauschenbach, I. Y., E. K. Karpova, A. A. Alekseev, N. V. Adonyeva, L. V. Shumnaya and N. E. Gruntenko (2015). "Interplay of insulin and dopamine signaling pathways in the control of Drosophila melanogaster fitness." Doklady Biochemistry and Biophysics **461**(1): 135-138.

Rogers, R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto and K. R. Thornton (2014). "Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans." Molecular Biology and Evolution **31**(7): 1750-1766.

Rogers, R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto and K. R. Thornton (2015). "Tandem Duplications and the Limits of Natural Selection in Drosophila yakuba and Drosophila simulans." PLoS One **10**(7): e0132184.

Schmidt, J. M., R. T. Good, B. Appleton, J. Sherrard, G. C. Raymant, M. R. Bogwitz, J. Martin, P. J. Daborn, M. E. Goddard, P. Batterham and C. Robin (2010). "Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus." PLoS Genetics **6**(6): e1000998.

Schwartz, Y. B., D. Linder-Basso, P. V. Kharchenko, M. Y. Tolstorukov, M. Kim, H. B. Li, A. A. Gorchakov, A. Minoda, G. Shanower, A. A. Alekseyenko, N. C. Riddle, Y. L. Jung, T. T. Gu, A. Plachetka, S. C. R. Elgin, M. I. Kuroda, P. J. Park, M. Savitsky, G. H. Karpen and V. Pirrotta (2012). "Nature and function of insulator protein binding sites in the Drosophila genome." Genome Research **22**(11): 2188-2198.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nature Biotechnology **26**(10): 1135-1145.

Smith, C. D., S. Q. Shu, C. J. Mungall and G. H. Karpen (2007). "The Release 5.1 annotation of Drosophila melanogaster heterochromatin." Science **316**(5831): 1586-1591.

Solares, E. A., M. Chakraborty, D. E. Miller, S. Kalsow, K. Hall, A. G. Perera, J. J. Emerson and R. S. Hawley (2018). "Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing." G3-Genes Genomes Genetics.

Spencer, C. C. A., Z. Su, P. Donnelly and J. Marchini (2009). "Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip." PloS Genetics **5**(5).

Stanley, P. D., E. Ng'oma, S. O'Day and E. G. King (2017). "Genetic Dissection of Nutrition-Induced Plasticity in Insulin/Insulin-Like Growth Factor Signaling and Median Life Span in a Drosophila Multiparent Population." Genetics **206**(2): 587-602.

Stranger, B. E., E. A. Stahl and T. Raj (2011). "Progress and promise of genome-wide association studies for human complex trait genetics." Genetics **187**(2): 367-383.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stutz, X. H. Shi, F. P. Casale, J. M. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. C. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. L. Yu, C. S. Zhang, J. Zhang, X. Zheng-Bradley, W. D. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel and G. P. Consortium (2015). "An integrated map of structural variation in 2,504 human genomes." Nature **526**(7571): 75-+.

Tatar, M., A. Kopelman, D. Epstein, M. P. Tu, C. M. Yin and R. S. Garofalo (2001). "A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function." Science **292**(5514): 107-110.

The modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White and M. Kellis (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science **330**(6012): 1787-1797.

Thornton, K. R., A. J. Foran and A. D. Long (2013). "Properties and Modeling of GWAS when Complex Disease Risk Is Due to Non-Complementing, Deleterious Mutations in Genes of Large Effect." Plos Genetics **9**(2).

Toivonen, J. M. and L. Partridge (2009). "Endocrine regulation of aging and reproduction in Drosophila." Molecular and Cellular Endocrinology **299**(1): 39-50.

Turelli, M. (1984). "Heritable Genetic-Variation Via Mutation Selection Balance - Lerch Zeta Meets the Abdominal Bristle." Theoretical Population Biology **25**(2): 138-193.

Varshney, R. K., S. N. Nayak, G. D. May and S. A. Jackson (2009). "Next-generation sequencing technologies and their implications for crop genetics and breeding." Trends in Biotechnology **27**(9): 522-530.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young and A. M. Earl (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement." PLoS One **9**(11): e112963.

Warren, W. D., S. Palmer and A. J. Howells (1996). "Molecular characterization of the cinnabar region of Drosophila melanogaster: identification of the cinnabar transcription unit." Genetica **98**(3): 249-262.

Waterhouse, R. M., M. Seppey, F. A. Simao, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva and E. M. Zdobnov (2017). "BUSCO applications from quality assessments to gene prediction and phylogenomics." Molecular Biology and Evolution.

Wei, Y., R. H. Gokhale, A. Sonnenschein, K. M. Montgomery, A. Ingersoll and D. N. Arnosti (2016). "Complex cis-regulatory landscape of the insulin receptor gene underlies the broad expression of a central signaling regulator." Development **143**(19): 3591-3603.

Yang, W. Y., S. Y. Wen, Y. D. Huang, M. Q. Ye, X. J. Deng, D. Han, Q. Y. Xia and Y. Cao (2006). "Functional divergence of six isoforms of antifungal peptide Drosomycin in Drosophila melanogaster." Gene **379**: 26-32.

Ye, C., C. M. Hill, S. Wu, J. Ruan and Z. S. Ma (2016). "DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies." Scientific Reports **6**: 31900.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler and Z. Ning (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." Bioinformatics **25**(21): 2865-2871.

Data availability

All scaffolded assemblies and the raw sequence data (HDF5 files and their respective metadata) have been deposited in NCBI under the Bioproject accession PRJNA418342. All scripts, raw SV outputs, and processed data are available at https://github.com/mahulchak/dspr-asm.

Author Contribution

All authors conceived of the work and wrote the manuscript. MC assembled the genomes and wrote the variant caller.

Acknowledgements

We wish to acknowledge support from the following grants OD010974 (SJM and ADL), GM115562 (ADL), R01GM123303-1 and University of California, Irvine setup funds (JJE). We thank Luna Thanh Ngo and Daniel Na for help with data management and fly maintenance. This work was made possible, in part, through access to the Genomics High-Throughput Facility Shared Resource of the Cancer Center Support Grant CA-62203 at the University of California, Irvine, and NIH shared-instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01.

Table 1. Summary of assembly metrics. (+BUSCO = Benchmarking Universal Single Copy Orthologs; N50 = sequence length such that 50% of the assembly is contained within sequences of that length or longer)

Strains	Assembly Size (Mb)	Contig N50 (Mb)	# of scaffolds	Complete BUSCO (%) (Arthropoda)
ISO1	139.5	21.4	1856	100
A1	137.6	21.8	76	100
A2	142.4	22.3	193	99.9
A3	133.3	21.6	44	100
A4	139.6	22.4	95	100
A5	138.9	20.9	99	99.9
A6	133.3	21.5	29	100
A7	146.8	21.5	263	100
AB8	137.7	21.7	56	100
B1	135.9	21.8	39	100
B2	137.4	18.9	58	100
B3	136.2	21.4	43	100
B4	136.2	20	65	100
B6	137.4	18.5	61	100
Ore	136.4	21.5	75	100

Table 2. Number of euchromatic SVs in the sequenced DSPR founder strains and Oregon-R.

Strains	TE	Duplication CNV	Indels	Inversion
A1	620	144	584	4
A2	618	123	785	10
A3	580	134	702	11
A4	581	136	683	7
A5	597	122	700	10
A6	760	136	681	10
A7	629	184	916	8
AB8	606	121	660	10
B1	646	129	699	6
B2	687	132	633	7
B3	624	147	720	8
B4	656	120	646	4
B6	683	116	682	4
Ore	518	135	621	14

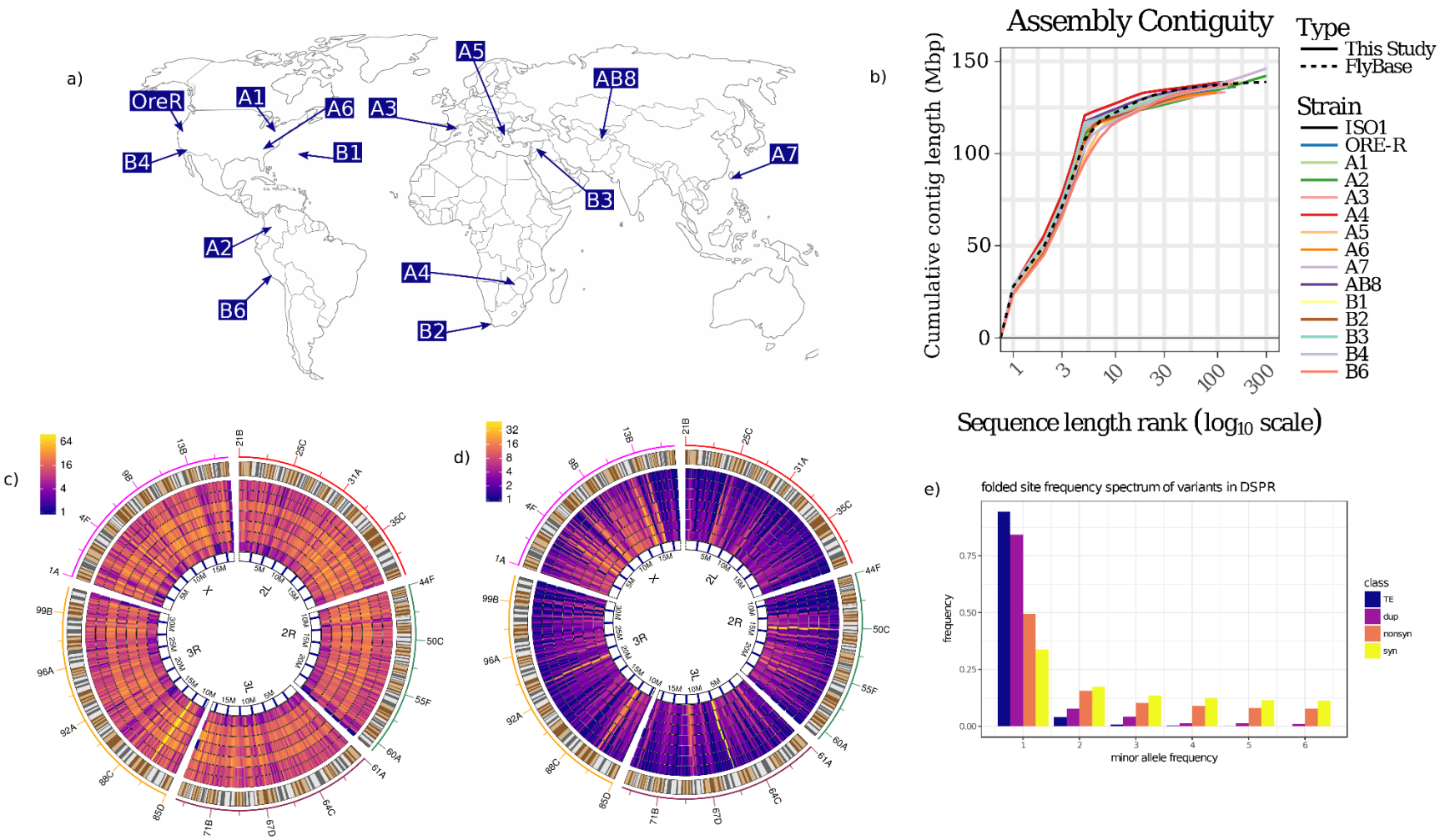

Figure 1. a) Geographic locations of the sequenced strains of *D. melanogaster*. As shown here, the founder strains from DSPR and Oregon-R originate from diverse worldwide populations. b) Cumulative contiguity plot showing comparison of assembly contiguity between the reference strain ISO1 and our 14 assemblies. 1c) Distribution of euchromatic TE insertions across the major chromosome arms. The outermost track represents the chromosome ideogram, showing the

locations of named bands. Each subsequent inner track shows distributions of TE insertions per genomic window of fixed size, ranging from 100-400kb in 50kb increments. Details of the TE rich region (yellow streak) on 3R (12.47-12.5Mb) is shown in supplementary Fig. 4. d) Distribution of duplication CNVs within euchromatin of major chromosome arms. The outermost track represents ideogram as in Fig.1c. Inner tracks represent distributions of duplication CNVs in windows of varying sizes as in Fig1c. Unlike TEs, distribution of duplications are less uniform within and between the chromosomes. e) Counts of minor allele frequency for TE and duplicated sequences.

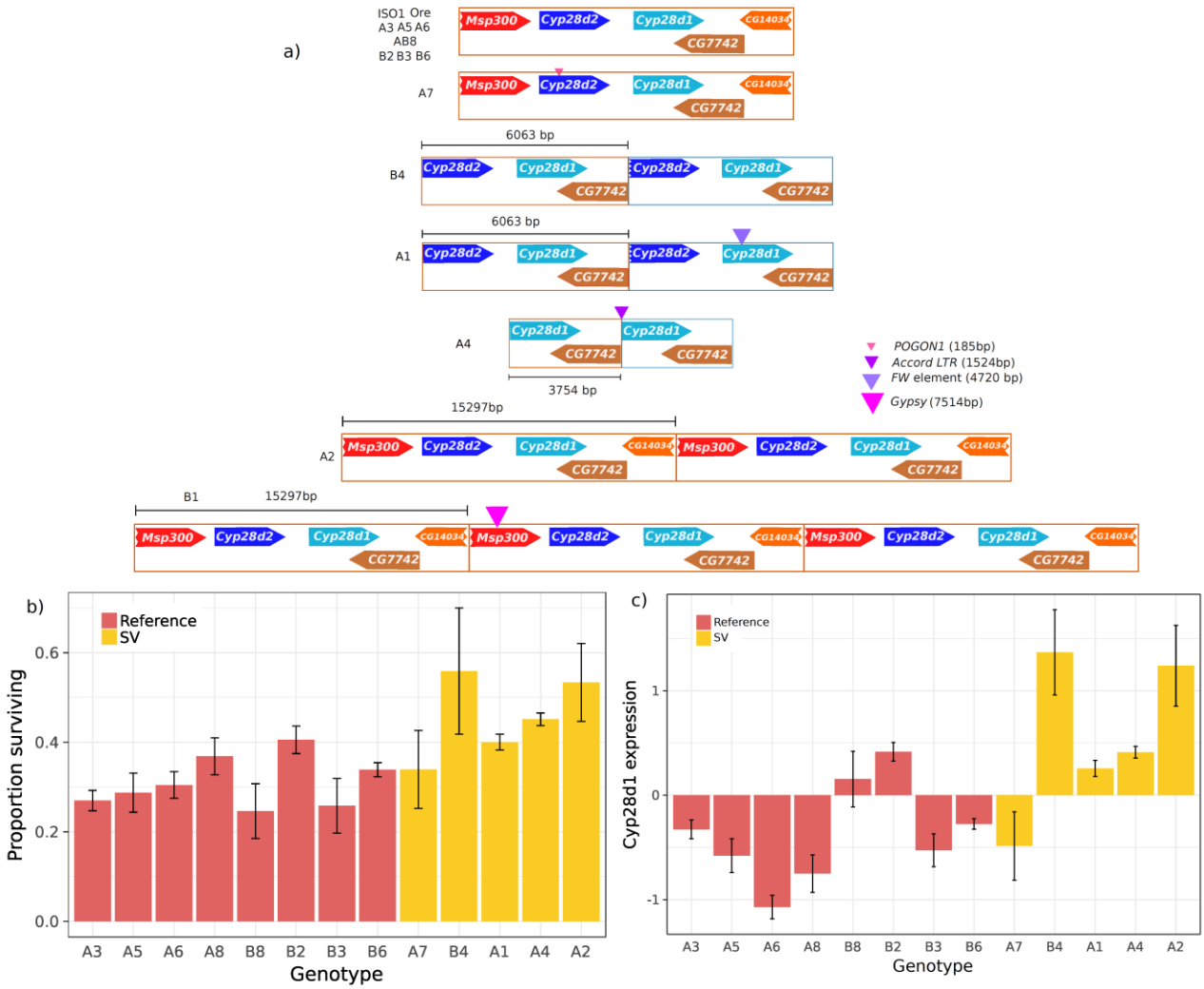

Figure 2. a) Structural alleles at Cyp28d genes underlying a nicotine resistance QTL. The most common allele is the reference (ISO1) allele and contains a single copy of *Cyp28d1* and *Cyp28d2*. All others are private to a single founder strain. A1 and B4 carry the same duplicate but A1 has a FW element inserted into the second *Cyp28d1* copy. A2 and B1 possess two and three copies of a 15 Kb segment ,respectively, that contains both *Cyp28d1* and *Cyp28d2*. b) Nicotine resistance of RILs as a function of founder allele at the QTL harboring *Cyp28d1* and *Cyp28d2*. Genotype ordering matches fig. 2a. A2 and B4 alleles are most resistant to nicotine, followed by A4. A2 possesses a 15 Kb duplication containing full *Cyp28d1* and *Cyp28d2*, whereas B4 contains duplicate of a full *Cyp28d1* and near complete *Cyp28d2*. No RIL homozygous for the B1 allele was present in this sample. c)Normalized *Cyp28d1* expression level in RILs with different founder genotypes at the cis-eQTL for *Cyp28d1* (supplementary Fig. 12). Genotype ordering (L to R) follows genotype ordering (top to bottom) in Fig. 2a. A2 and B4 genotypes, which show highest resistance to nicotine toxicity, also show highest upregulation for *Cyp28d1*. Despite A1 and B4 having the same duplication, the *Cyp28d1* disrupting TE insertion in A1 is likely responsible for lower expression of the gene in A1.

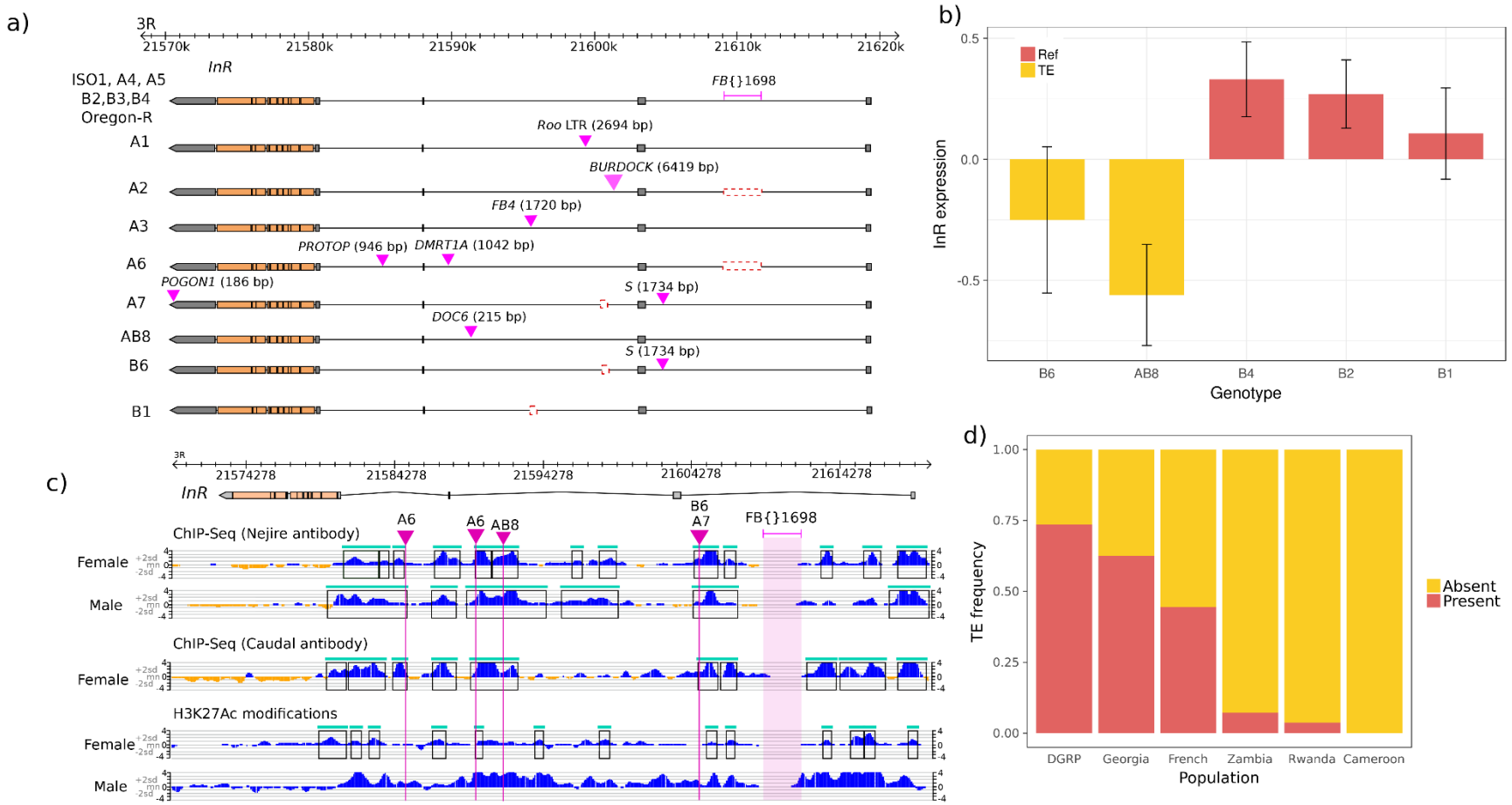

Figure 3. a) Eight structurally different alleles consisting of different intronic TEs at Insulin receptor (*InR*) gene. The reference TE *FB{}1698* is present in all sequenced strains, except A2 and A6. In contrast, most of the other TEs are private to the founder strains carrying them. The *S* element in A7 and B6 and the *Doc6* element in AB8 insert into known cis-regulatory sequence. b) Expression level of *InR* in F1 hybrids between panel B RILs and the founder strain A4. RILs with AB8 and B6 genotype show downregulation of *InR* among the panel B genotypes. Both AB8 and B6 harbors TEs that insert into known cis-regulatory intronic sequence of *InR*. c) Insertion of TE insertions into transcription factor binding sites (TFBS) for Caudal and Nejire. The top and middle panel shows the TF binding peaks detected from ChIP-seq performed

with Nejire and Caudal antibodies, respectively. The bottom panel shows H3K27Ac histone modifications representing a transcriptionally active state. The histone marks largely overlap with the TFBS, supporting the functional significance of the latter. High frequency FB{}1698 is inserted between two TFBS enriched sites. Disruption of TFBS by transposon insertion in B6 and AB8 *InR* alleles likely cause downregulation of the gene as shown in 3b. d) Frequency of FB {} 1698 insertion in *InR* in different cosmopolitan and ancestral populations of *D. melanogaster*. The TE insertion is rare in the ancestral African populations but segregates in intermediate to high frequencies in the derived, cosmopolitan populations.

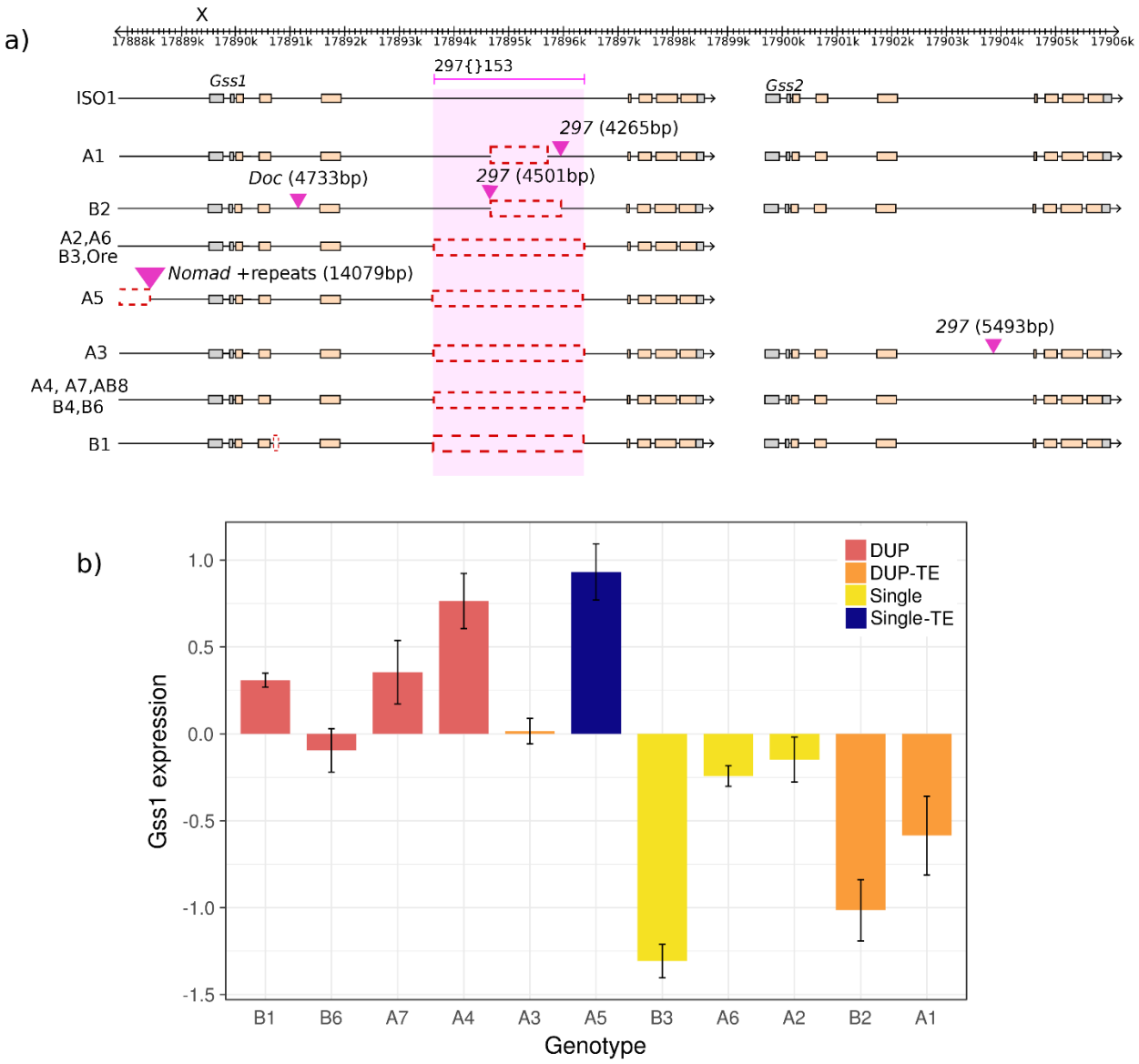

Figure 4. a) Structurally distinct alleles at *Glutathione synthetase* (*Gs*) locus. Nine founder alleles consist of duplication of *Gs*, among which four also carries insertion of the LTR retrotransposon *297.* Among the five single *Gs* copy alleles, A5 possess a 14Kb insertion comprising *Nomad* LTR TE fragments and simple repeats located1064 bp 5' to the transcription start site of *Gss1*. A5 also carries a deletion that removes 3489 bp (X:17,884,993-17,888,482) from the 5' upstream region of *Gss1*. b) Normalized expression level of Gss1 for different founder genotypes. All founder alleles, except A5, possessing single copy *Gs* show varying levels of downregulated transcript levels. We hypothesize that the indel upstream of the *Gss1* transcription start site (TSS) in A5 up-regulates expression.

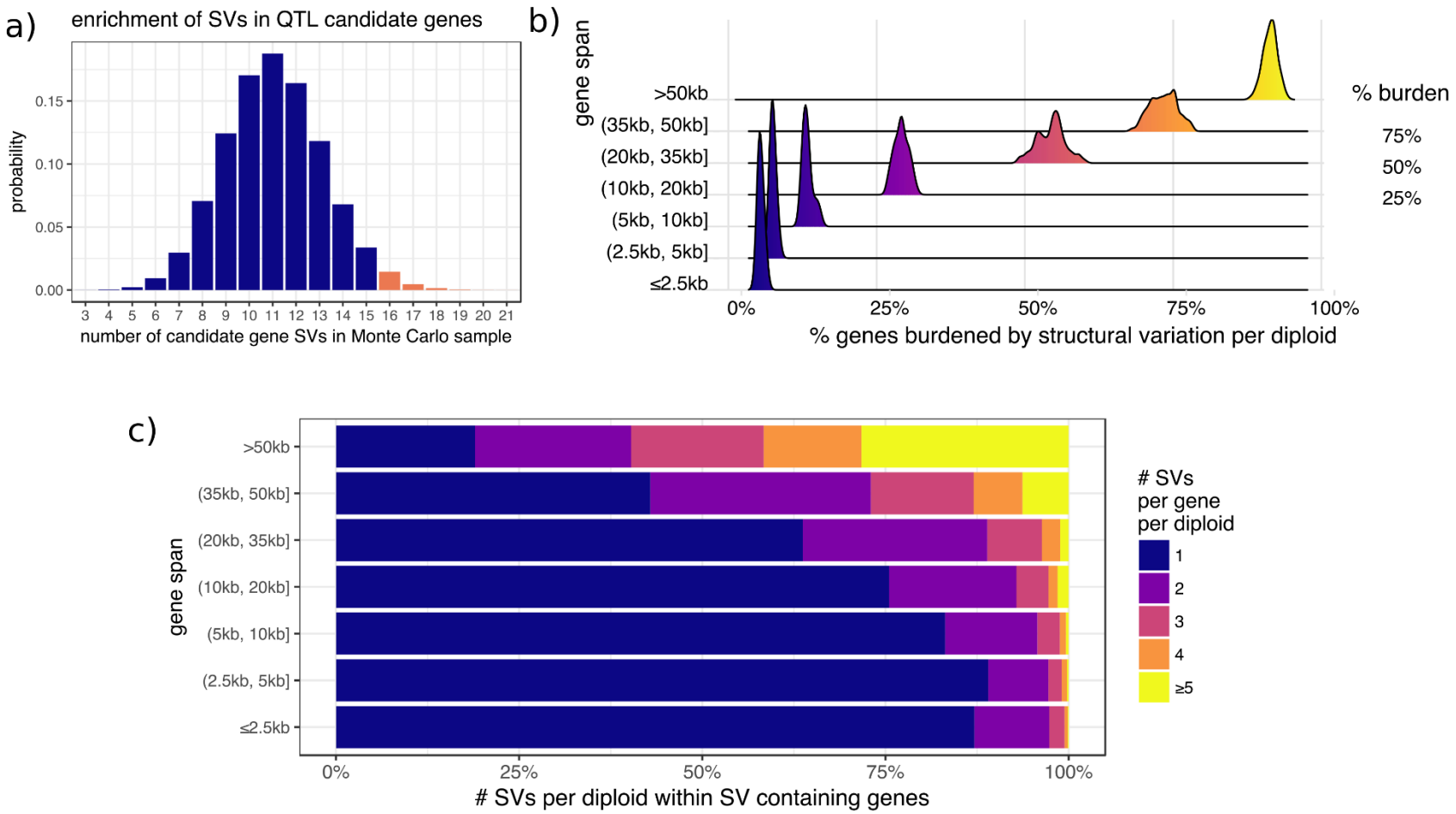

Figure 5. a) Monte Carlo distribution of number of QTL candidate genes possessing at least one SV in a sample of 32 genes. In total, 2% of the samples exhibit at least 16 genes harboring SVs (red bars). Genes are randomly chosen such that the gene length distributions of the Monte Carlo samples are the same as the observed candidate genes. In the empirical dataset, 16 QTL candidate genes possessed one or more SVs. b) Structural variant burden in diploids. The x-axis describes the percentage of genes in a particular length category that carry one or more SVs in diploid individuals. The distributions are derived from the collection of all 78 possible diploids that can be constructed from crosses of the 13

founders reported here. Number of genes in each length bin is in supplementary Table 5. c) Structural variant multiplicity in diploids. The x-axis describes the number of variants per diploid individual in SV carrying genes. The y-axis describes the length of the gene span.