Structural variants exhibit allelic heterogeneity and shape variation in complex traits

Mahul Chakraborty¹*, J.J. Emerson¹, Stuart J. Macdonald², Anthony D. Long¹*

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

²Department 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 = 130×10^6 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 arthropoda 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 chromosome 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 flyrilis.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 previously15. 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 $y_i = \min(x_i, N-x_i)$. Pairwise comparisons between different SFS site categories were conducted using the $\chi^2$ 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 $\geq 1$ 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 *DrsI5*.
(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, χ2 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, χ2 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...
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{1}698) 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 \textit{InR} (Paaby et al. 2010). In total, \textit{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, DNAseq1HSseq) against reference genomes which are often genetically different than the strains assayed (The modENCODE Consortium \textit{et al.} 2010, Graveley \textit{et al.} 2011, Negre \textit{et al.} 2011, Schwartz \textit{et al.} 2012). Canton-S (our DSPR founder A1) and Oregon-R are strains commonly used in phenotypic assays (The modENCODE Consortium \textit{et al.} 2010, Graveley \textit{et al.} 2011, Schwartz \textit{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 \textit{Gss1} and \textit{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 \textit{Gss1} and \textit{Gss2}, which are associated with toxin metabolism and linked to tolerance to arsenic (Ortiz \textit{et al.} 2009) and ethanol induced oxidative stress (Logan-Garbisch \textit{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 \textit{et al.} 2011, Schwartz \textit{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 = 6.5×10-5), 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 = 2.9×10-3; 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


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)

<table>
<thead>
<tr>
<th>Strains</th>
<th>Assembly Size (Mb)</th>
<th>Contig N50 (Mb)</th>
<th># of scaffolds</th>
<th>Complete BUSCO (%) (Arthropoda)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO1</td>
<td>139.5</td>
<td>21.4</td>
<td>1856</td>
<td>100</td>
</tr>
<tr>
<td>A1</td>
<td>137.6</td>
<td>21.8</td>
<td>76</td>
<td>100</td>
</tr>
<tr>
<td>A2</td>
<td>142.4</td>
<td>22.3</td>
<td>193</td>
<td>99.9</td>
</tr>
<tr>
<td>A3</td>
<td>133.3</td>
<td>21.6</td>
<td>44</td>
<td>100</td>
</tr>
<tr>
<td>A4</td>
<td>139.6</td>
<td>22.4</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>A5</td>
<td>138.9</td>
<td>20.9</td>
<td>99</td>
<td>99.9</td>
</tr>
<tr>
<td>A6</td>
<td>133.3</td>
<td>21.5</td>
<td>29</td>
<td>100</td>
</tr>
<tr>
<td>A7</td>
<td>146.8</td>
<td>21.5</td>
<td>263</td>
<td>100</td>
</tr>
<tr>
<td>AB8</td>
<td>137.7</td>
<td>21.7</td>
<td>56</td>
<td>100</td>
</tr>
<tr>
<td>B1</td>
<td>135.9</td>
<td>21.8</td>
<td>39</td>
<td>100</td>
</tr>
<tr>
<td>B2</td>
<td>137.4</td>
<td>18.9</td>
<td>58</td>
<td>100</td>
</tr>
<tr>
<td>B3</td>
<td>136.2</td>
<td>21.4</td>
<td>43</td>
<td>100</td>
</tr>
<tr>
<td>B4</td>
<td>136.2</td>
<td>20</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>B6</td>
<td>137.4</td>
<td>18.5</td>
<td>61</td>
<td>100</td>
</tr>
<tr>
<td>Ore</td>
<td>136.4</td>
<td>21.5</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 2. Number of euchromatic SVs in the sequenced DSPR founder strains and Oregon-R.

<table>
<thead>
<tr>
<th>Strains</th>
<th>TE</th>
<th>Duplication</th>
<th>CNV</th>
<th>Indels</th>
<th>Inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>620</td>
<td>144</td>
<td>584</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>618</td>
<td>123</td>
<td>785</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>580</td>
<td>134</td>
<td>702</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>581</td>
<td>136</td>
<td>683</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>597</td>
<td>122</td>
<td>700</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>760</td>
<td>136</td>
<td>681</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>629</td>
<td>184</td>
<td>916</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>AB8</td>
<td>606</td>
<td>121</td>
<td>660</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>646</td>
<td>129</td>
<td>699</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>687</td>
<td>132</td>
<td>633</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>624</td>
<td>147</td>
<td>720</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>656</td>
<td>120</td>
<td>646</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>B6</td>
<td>683</td>
<td>116</td>
<td>682</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Ore</td>
<td>518</td>
<td>135</td>
<td>621</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>
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. c) 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 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 located 1064 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.