Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits

Mahul Chakraborty, J.J. Emerson, Stuart J. Macdonald & Anthony D. Long

It has been hypothesized that individually-rare hidden structural variants (SVs) could account for a significant fraction of variation in complex traits. Here we identified more than 20,000 euchromatic SVs from 14 Drosophila melanogaster genome assemblies, of which ~40% are invisible to high specificity short-read genotyping approaches. SVs are common, with 31.5% of diploid individuals harboring a SV in genes larger than 5kb, and 24% harboring multiple SVs in genes larger than 10kb. SV minor allele frequencies are rarer than amino acid polymorphisms, suggesting that SVs are more deleterious. We show that a number of functionally important genes harbor previously hidden structural variants likely to affect complex phenotypes. Furthermore, SVs are overrepresented in candidate genes associated with quantitative trait loci mapped using the Drosophila Synthetic Population Resource. We conclude that SVs are ubiquitous, frequently constitute a heterogeneous allelic series, and can act as rare alleles of large effect.
Understanding the molecular basis of heritable variation in complex traits is of central importance to evolution, animal and plant breeding, and medical genetics\textsuperscript{1–4}. Over the last decade, short read genomic data (50–150 bp reads) appropriate for characterizing SNPs and small indels in non-repetitive genomic regions has accumulated at an exponential rate\textsuperscript{5,6}. This in turn has catalyzed hundreds of quantitative trait locus (QTL) mapping and genome-wide association (GWAS) studies in model organisms, humans, and agriculturally important animals and plants\textsuperscript{7–22}. Despite these efforts, for most traits, GWAS hits only explain a small fraction of known trait heritability\textsuperscript{10,11,13,14}. One hypothesis accounting for hidden genetic variation is that individually rare hidden mutations that alter genome structure make significant contributions to complex trait variation\textsuperscript{11,12}. 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\textsuperscript{13}. Moreover, genotyping approaches based on short reads or microarrays fail to detect a significant number of SVs\textsuperscript{14,15}. Finally, it is reasonable to assume that SVs are on average likely to be more deleterious and deleterious more often than SNPs\textsuperscript{16–19}.

High quality genomes provide a direct and reliable path to comprehensive identification of SVs\textsuperscript{15,20,21}. To achieve this goal, we assembled reference-quality genomes for fourteen geographically diverse Drosophila melanogaster strains (Fig. 1a) using single molecule real time sequencing\textsuperscript{22}. These assemblies are contiguous and complete (N50 18.9–22.3 Mb; BUSCO\textsuperscript{23} 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)\textsuperscript{24}, a large set of advanced intercross recombinant inbred lines (RILs) designed to map QTLs\textsuperscript{25}. 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 mod-ENCODE\textsuperscript{26–28}. Using these reference quality genome assemblies, we show that SVs are common in D. melanogaster genes, with almost one third of diploid individuals harboring an SV in genes larger than 5 kb, and more than a third of burdened genes carrying multiple SVs. The site frequency spectrum (SFS) of SV alleles relative to amino acid polymorphisms suggests that SVs are under stronger purifying selection, and thus are more likely to impact phenotype than nonsynonymous SNPs. We further show that a number of functionally important genes harboring previously hidden SVs likely to affect complex phenotypes (e.g., Cyp6g1, Drs5, Cyp28d1, Cyp28d2, InR, and Gss1e-2). Finally, we find that SVs are overrepresented in candidate genes associated with mapped QTL. We conclude that SVs are pervasive in genomes, frequently manifest as heterogeneous allelic series affecting the same gene, and exhibit all the properties that make them prime candidates for being rare alleles of large effect.

**Results**

**De novo assembly reveals novel 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\textsuperscript{29} in all our assemblies (Supplementary Figs. 1–3). We identified SVs by comparing each assembly to the reference ISO1 genome\textsuperscript{30}, focusing our attention on large (>100 bp) euchromatic SVs (Supplementary Table 2), and ignoring heterochromatin regions as they are gene poor\textsuperscript{30} and require specialized assembly approaches and extensive validation\textsuperscript{31}. 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 7347 TE insertions, 1178 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). We estimate that 36% of non-reference TEs, 26% of deletions, 48% of insertions, 60% of duplication CNVs are not routinely detected using high coverage paired end Illumina reads and high specificity SV genotyping methods\textsuperscript{32} (Supplementary Fig. 6).

We uncover many examples of previously hidden SVs predicted to affect complex traits. Extensive evidence links complex SV alleles of the cytochrome P450 gene Cyp6g1 to varying levels of DDT resistance\textsuperscript{32,33}. Despite extensive study of this locus, we discovered three 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\textsuperscript{34} that exhibits >1000-fold higher expression relative to its single copy counterpart in line A4 (Supplementary Fig. 8a, b). Read pair orientation and split read methods failed to detect this mutation because one allele bears a 5 kb 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 cinnamon\textsuperscript{35} (cn) and speck (sp) present in the ISO1 reference genome\textsuperscript{36} (Supplementary Figs. 9 and 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\textsuperscript{36}. Both alleles are likely knock-outs.

**SVs are deleterious.** Most TEs and duplicates are present in only one strain (Fig. 1e), with the folded SFS of the TEs and duplicates exhibiting a greater proportion of rare variants than non-synonymous SNPs (nsSNPs) assayed in the same strains (Fig. 1e; p-value < 1 x 10\textsuperscript{-10}, \chi\textsuperscript{2} test between frequency classes of these two types of SVs and non-synonymous SNPs). Since nsSNPs were ascertained via high coverage short reads from virtually isogenic strains\textsuperscript{24}, the low frequency skew of the site frequency spectrum of SVs relative to nsSNPs is unlikely due to SNP miscalcs (see Methods). It is well-known that the SFS is affected by demographic history\textsuperscript{37,38} but selective constraints can be inferred with some confidence by comparing site classes from the same sample\textsuperscript{37,39–41}. The skew toward rare variants we observe in our SVs relative to nsSNPs is strongly indicative of SVs being under stronger purifying selection, consistent with previous work in which SVs were ascertained with higher bias and/or errors\textsuperscript{16,18,19}. Furthermore, TEs are more enriched for rare variants than duplicates, indicating that TE insertions as a class are more deleterious than duplicates (Fig. 1e; p-value < 1 x 10\textsuperscript{-10}, \chi\textsuperscript{2} test between frequency classes of TEs and duplicates). Under mutation selection balance models\textsuperscript{42,43}, rare deleterious variants (minor allele frequency or MAF <1%) are predicted to contribute significantly to the variation in complex traits, yet are unlikely to be tagged by SNPs typically used in GWAS experiments\textsuperscript{10}. Although demography can impact the proportion of variation due to rare deleterious alleles, recent population bottlenecks or growth\textsuperscript{44,45} tend to amplify the contribution of rare alleles to variation in a complex trait.
SVs are common in genes and enriched at mapped QTLs. In order to illustrate how common SV genotypes are in heterozygous individuals, we quantified the per gene SV burden per synthetic diploid *D. melanogaster* individual (Fig. 2a, b; each synthetic diploid is one of 78 possible pairings of the thirteen assembled DSPR founders). On average, SVs appear in 9.3% of genes in diploid individuals (1285/13761). Of those, more than a third of burdened genes in diploids (443/1285) bear multiple SVs. Furthermore, individual genes carrying multiple SVs comprise more than a third of burdened genes between 20 and 35 kb in length and more than half
of larger genes (Fig. 2b). Thus, although generally having rare minor allele frequencies, SVs are ubiquitous in the functional elements of D. melanogaster genome.

Although hypotheses employing SVs to explain missing heritability have been proposed\cite{16,46}, the systematic under-identification of SVs via short read- and microarray-based genotyping\cite{21} limits their explanatory power. Using our comprehensive SV map, we measured the prevalence of SVs at the candidate genes reported in eight complex trait mapping experiments employing DSPR (Supplementary Data 1). We consider only genes in mapped QTLs explicitly cited by the authors of the original QTL studies (Supplementary Data 1; see Methods). In total, we identified 31 candidate genes of which 15 (48.4%) possess at least one SV in one founder strain, whereas only 23.4% (3237/13,830) of other D. melanogaster genes harbor SVs (p = 0.0023; Fisher’s exact test). The 31 candidate genes from QTL mapping work (Supplementary Data 1) are more than twice as large as an average Drosophila gene (12.2 vs 5.4 kb, Wilcoxon rank sum test; p = 1.4 × 10^{-4}; Supplementary Data 2). We then tested whether QTL candidate genes are enriched for SVs independent of gene size. To control for the observed elevated size distribution of QTL candidate genes, we randomly drew 100,000 Monte Carlo gene samples matching the candidate gene length distribution in order to generate null distribution for the number of SVs we expect to observe. In the Monte Carlo sample, 10.4 genes are expected to harbor SVs by chance whereas we observe 15, an enrichment of 45% (p = 0.026; Fig. 2c). Similarly, enrichment is also observed when burden is measured as density (# burdened genes/bp) in comparison of QTL candidate genes to the remaining genes (39.6 burdened genes/Mbp of genic DNA vs 27.3 burdened genes/Mbp of genic DNA, p = 0.017; Supplementary Data 2). Finally, candidate genes also exhibit greater SV mutation density compared to the expected SV density in the Monte Carlo samples (205.8 SVs/Mbp vs 138.4 SV/Mbp, p = 0.047; Supplementary Data 2), an enrichment of 49%. These observations suggest that SVs may be disproportionately associated with QTL candidate genes even after accounting for gene size. However, due to the limited sample size of QTL candidate genes, the power to detect even strong effects (e.g. the 45% enrichment reported here) is limited. Consequently, further QTL studies identifying candidate genes would substantially improve our understanding of this effect. These observations suggest that the contribution of rare SVs of large effect to complex traits could be pervasive.

Fig. 2 Ubiquity of SVs in D. melanogaster genes and QTL candidate genes. a 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 Data 4. b 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. c Monte Carlo distribution of number of QTL candidate genes possessing at least one SV in a sample of 31 genes. In total, 2.6% of the samples exhibit at least 15 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, 15 QTL candidate genes possessed one or more SVs.
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. 3c). Consistent with this, DSPR Recombinant Inbred Lines (RILs) bearing the A2 genotype show the highest average resistance to nicotine toxicity among the DSPR Recombinant Inbred Lines (RILs) bearing the A2 genotype all founder genotypes measured (Fig. 3c). Consistent with this, female heads of RILs bearing the A2 genotype is highest among founder allele at the QTL harboring Cyp28d1 and Cyp28d2. Genotype ordering (L to R) follows genotype ordering (top to bottom) in Fig. 2a. A2 and B4 alleles are most resistant to nicotine, followed by A4. A2 possesses a 15 Kbp 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. Error bars indicate standard errors of genotypic means.

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 founder48. Among these is Insulin Receptor (InR), which plays a key role in several life history traits related to lifespan and is likely a key molecular mediator of the tradeoff between reproductive success and longevity49–51. Amino acid polymorphism in InR evolves under positive selection and some non-synonymous variants affect fecundity and stress response51,52. Expression variation of InR also affects body size,
lifespan, and fecundity33,54, suggesting that natural cis-regulatory variation might also be under selection. We discovered a 215 bp fragment of a DOC6 element within a second intron enhancer55 (Fig. 4b, c) of InR on the AB8 haplotype, and this allele exhibits reduced gene expression relative to reference genotypes (Fig. 4b). This mutation potentially 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 insert within known cis-regulatory elements55 (Fig. 4b, c). Except for A2 and A6, all strains, including ISO1, harbor an FB-NOF element (FB|1698) inside the first intron of InR (Fig. 4a). Like many genes, the first intron of InR possess several transcription factor binding sites (TFBS), including those for factors Nejire and Caudal56 (Fig. 4c). The FB-NOF element is inserted within this dense cluster of TFBS and active enhancer marks (Fig. 4c). Furthermore, the FB element is segregating at high frequency in the strains discussed here (13/15), a North American population57 (125/170), and a French population58 (4/9), but is rare in populations derived from D. melanogaster's ancestral range in Africa58,59 (Cameroon: 0/10, Rwanda: 1/27, Zambia: 10/139) (Fig. 4d). 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 InR55. In total, InR harbors a remarkable amount of potentially functional structural diversity. Including these variants described above, there are nine TE insertions and two deletions throughout the gene, many of which impinge on candidate regulatory regions or transcribed portions of the gene (Fig. 4a, c).

Public resources like modENCODE annotate molecular phenotypes (e.g., RNA-seq, ChIPseq, DNase-seq) against reference genomes which are often genetically different than the strains assayed26–28,56. Canton-S (our DSPR founder A1) and Oregon-R are strains commonly used in phenotypic assays26–28, 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 is just one example among hundreds in modENCODE that likely suffer from misleading annotations.
structurally distinct Gss alleles in our dataset, ISO1 is the sole representative of its allele (Fig. 5a). The two most common Gss alleles include one that contains only a single Gss 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. 5a). The remaining six alleles have SV genotypes represented by only a single individual in the sample. Collectively, this sample represents a haplotype network of structural variation involving five 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 14 kb 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. 5b).

**Discussion**

Despite claims that a significant proportion of complex trait variation in humans, model organisms, and agriculturally important animals and plants is likely due to rare SVs of large effect\(^1\), systematic inquiry of this hypothesis has been impeded by genotyping approaches attuned to SNP detection\(^2\). 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 (Figs. 2a, 3a, 4a, and 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 for variation in complex traits 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 effect62,63. Furthermore, models for complex traits under either stabilizing63,64 or purifying selection12,43 with constant mutational input predict the existence of genes segregating several individually rare causative alleles that account for a sizable fraction of 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.

Methods

DNA extraction. Genomic DNA was extracted from females following the protocols described previously22 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). All testing and research involving flies were performed in compliance with relevant ethical regulations. 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 Table S6.

Genome assembly. The genomes were assembled following the approach described in Chakraborty et al.32. For all calculations of sequence coverage, a genome size of 130Mbp is assumed (G = 130 × 10^6 bp). For individual strain, we generated a hybrid assembly with DRCO2 v2.4.2 and longest W30X PacBio reads, and a PacBio assembly with canu v1.3966 (Supplementary Data 3). The paired end Illumina reads were obtained from King et al.24. The hybrid assemblies were merged with the PacBio only assemblies with quickmerge v0.2222,67 ((l = 2 Mb, 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 quickmerge67. For the second round of quickmerge (l = 5 mb, 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 reads68. The assemblies were then polished twice with quiver (SMRTanalyisis v3.3.0p5) and once with Pilon v1.1609 using the same Illumina reads as used for the hybrid assemblies.

Comparative scaffolding. We scaffolded the contigs for each assembly based on the scaffolds from the reference assembly29,25, by following a previously described approach15. 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 MUMmer71. 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 alignments. The unscaffolded contigs were joined by 100 Ns, a convention representing assembly gaps. Of the unscaffolded sequences were named with a “U” prefix.

BUSCO analysis. We ran BUSCO (v3.0.2)72 on the Pilon polished pre-scaffolding assemblies to evaluate the completeness of all the assemblies relative to the ISO1 release 6 (v6.13). Assembly was used both the arthropoda and diptera datasets for the BUSCO evaluation. For the arthropoda database, three ortho- loops (EOG090X0822NZ, EOG090X0800M0, and EOG090X0949L0) were not 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 (EOG090X0822NZ is CG3223, EOG090X0800M0 is Pa1 and EOG090X0949L0 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 assembly release 6 (release 6.13)70 using nucmer (nucmer -maxmatch -noendcent71. We identified and classified the variants using SVUM 0.2beta (Structural variants from MUMmer) (n = 10)42. SVUM classifies the structural differences between two assemblies as insertion, deletion, or inversion based on whether the TE, shorter, more copy, or inverted sequence, respectively, with respect to the reference genome. The variant calls for individual genomes were combined using bedtools merge and converted into a vcf file using a custom script (https://github.com/mahulchak/dspr-asr). 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 D. simulans strains for the BUSCO evaluation. For the arthropoda database, three ortho- loops (EOG090X0822NZ, EOG090X0800M0, and EOG090X0949L0) were not 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 (EOG090X0822NZ is CG3223, EOG090X0800M0 is Pa1 and EOG090X0949L0 is CG40178). Consequently, we removed these three genes from consideration as uninformative.

Comparing SV genotypes from de novo assemblies to short read only calls. TE genotypes for the founders14 were downloaded from flyrils.org and the insertion coordinates were lifted over to the current release (release 6) of the reference genome20 using UCSC liftover tool72. For detection of the duplicates, we have previously found that discordant read pair based method (Pecnv)55 was comparable to split read mapping and more reliable than methods based on coverage alignment62. We used Pecnv. The alignment was run using the settings described previously15. Because svmu reports tandem duplicate CNVs as insertions (with appropriate CN 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. SVMU genotypes converted to the format of the 15 Drosophila species were present, expression data only for the longest transcript that is expressed in R. floccosus using both methods, 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 heads74 and IIS/TOR expression data48 from whole bodies were downloaded from www. flyrils.org. Expression QTL analysis (Supplementary Fig. 11) for Cy28d2 and Gs1 using the head expression data were performed using the R package DSRPqtl following the instructions provided in the manual (DSRPqalc, model = gene –1design = “Abcross”). When expression data for multiple isolate was 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 DSRPpeaks included in the DSRPqtl package. No eQTL were found for InR so the genotype values for the InR expression data were obtained by assigning the founder genotype to the RIL using the ISOPeach dataset, using the posterior probabilities of the forward-backward decoding of the HMM for the panel B RILs available on www.flyrils.org. Drsl expression levels in A4 and A3 were obtained from a publicly available RNAseq dataset47.

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 file24 using vcf tools and bedtools50,80. 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-asam). Complex mutations (CE = 1 and CE = 2) were excluded from the analysis. Let N be the sample size and xi 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 χ² 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 0 ≤ 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 differs from the gene set without the GstE1-10 array, calculated from the Monte Carlo samples. Although we present the enrichment results from analysis performed on the gene set without the GstE1-10 array, 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 Data 2) as well as the dataset excluding the GstE1-10 genes (Supplementary Data 2). 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 in the rest of the genome, we analyzed the candidate gene dataset without the GstE1-10 gene array (Table S4). The genes comprising GstE1-10 were excluded to avoid confounding effects of the complex structure of the locus and the identity of GstE1, GstE3, GstE5, and GstE6. As described below, the conclusions of the enrichment analysis do not depend on this locus, so we report the results of excluding it. A Fisher’s Exact Test was applied to the counts in categories of candidate gene vs. rest of the genome and SV-free vs. SV-burdened genes. To account for the lengths of the candidate genes being longer than the rest of the genome, we performed a Monte Carlo sampling of the whole genome according to the histogram of gene sizes in the candidate gene list (Supplementary Data 2). We then tallied up the number of observed SVs across all bins. 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 100,000 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). Expected density of SV-burdened genes (number of burdened genes per Mb of gene spans) and expected SV density (total number of SVs per Mb of gene spans) were also calculated from the Monte Carlo samples. Although we present the enrichment results from analysis performed on the gene set without the GstE1-10 array, inclusion of the array either as a single 13kb locus or as individual genes does not alter the conclusion of the enrichment analysis (single Gst locus: enrichment p-value = 0.021, length p-value = 6.5 × 10⁻⁵; individual Gst genes: length p-value = 0.034 and enrichment p-value = 2.9 × 10⁻⁴).

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 such diploid pairing, the number of 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 D. melanogaster release 6.16 gff file.68 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 ≥2 for a particular diploid combination is considered SV-burdened for that diploid.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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

Code availability
All scripts and codes have been deposited to GitHub and freely accessible from https://github.com/mahulchak/dspr-asam.

Received: 27 December 2018; Accepted: 25 September 2019; Published online: 25 October 2019

References
1. Mauricio, R. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat. Rev. Genet. 2, 370–381 (2001).
2. Mackay, T. F., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. 10, 565–577 (2009).
3. Goddard, M. E. & Hayes, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10, 381–391 (2009).
4. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187, 367–383 (2011).
5. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145 (2008).
6. Bansal, V. et al. Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res. 20, 537–545 (2010).
7. Goddard, M. E., Matukumalli, L. K., May, S. D. & Jackson, S. A. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 27, 522–530 (2009).
8. Day-Williams, A. G. & Zeggini, E. The effect of next-generation sequencing technology on complex trait research. Eur. J. Clin. Investig. 41, 561–567 (2011).
9. Davey, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499–510 (2011).
10. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
11. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450 (2010).
12. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. 10, 241–251 (2009).
13. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLoS Genet. 5, doi:ARTN e1000477 https://doi.org/10.1371/journal.pgen.1000477 (2009).
14. Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. Genetics 202, 1251–1254 (2016).
15. Chakraborty, M. et al. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nat. Genet. 50, 20–25 (2018).
16. Emerson, J. J., Cardoso-Moreira, M., Bobreitw, J. O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320, 1629–1631 (2008).
17. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712 (2010).
18. Cridland, J. M., Macdonald, S. J., Long, A. D. & Thornton, K. R. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. Mol. Biol. Evol. 30, 2311–2327 (2013).
19. Rogers, R. L. et al. Tandem duplications and the limits of natural selection in Drosophila yakuba and Drosophila simulans. PLoS One. 10, e0132184 (2015).
20. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 27, 677–685 (2017).
21. Hilden, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376 (2011).
22. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw654 (2016).
23. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
24. King, E. G. et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 22, 1558–1566 (2012).
25. Long, A. D., Macdonald, S. J. & King, E. G. Dissecting complex traits using the Drosophila Synthetic Population Resource. Trends Genet. 30, 488–495 (2014).
26. mod, E. C. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787–1797 (2010).
27. Graveley, B. R. et al. The developmental transcriptome of Drosophila melanogaster. Nature 471, 473–479 (2011).
28. Schwartz, Y. B. et al. Nature and function of insulator protein binding sites in the Drosophila genome. Genome Res. 22, 2188–2192 (2012).
Author contributions
M.C., J.J.E., S.J.M., A.D.L. conceived of the work and wrote the paper. M.C. assembled the genomes and wrote the variant caller.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-12884-1.

Correspondence and requests for materials should be addressed to M.C. or A.D.L.

Peer review information Nature Communications thanks Wen Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019