HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2018 June 18.

Published in final edited form as:

Nat Genet. 2018 January ; 50(1): 20–25. doi:10.1038/s41588-017-0010-y.

Hidden genetic variation shapes the structure of functional elements in *Drosophila*

Mahul Chakraborty[1,*], **Nicholas W. VanKuren**[2], **Roy Zhao**[3,4], **Xinwen Zhang**[1,3], **Shannon Kalsow**[1], and **J.J. Emerson**[1,4,*]

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

[2]Department of Ecology and Evolution, The University of Chicago

[3]Graduate Program in Mathematical, Computational and Systems Biology, University of California, Irvine

[4]Center for Complex Biological Systems, University of California, Irvine

Abstract

Mutations that add, subtract, rearrange, or otherwise refashion genome structure often affect phenotypes, though the fragmented nature of most contemporary assemblies obscure them. To discover such mutations, we assembled the first new reference quality genome of *Drosophila melanogaster* since its initial sequencing. By comparing this genome to the existing *D. melanogaster* assembly, we create a structural variant map of unprecedented resolution, revealing extensive genetic variation that has remained hidden until now. Many of these variants constitute strong candidates underlying phenotypic variation, including tandem duplications and a transposable element insertion that dramatically amplifies the expression of detoxification genes associated with nicotine resistance. The abundance of important genetic variation that still evades discovery highlights how crucial high-quality references are to deciphering phenotypes.

Mutations underlying phenotypic variation remain elusive in trait mapping studies[1] despite the exponential accumulation of genomic data, suggesting that many causal variants are invisible to current genotyping approaches[2–5]. In fact, mutations like duplications, deletions, and transpositions[6,7] are systematically underrepresented by standard methods[7], even as a

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

URL

All codes used for variant calling and scaffolding have been deposited to GitHub (https://github.com/mahulchak). Codes used in temperature gradient experiment have been deposited to GitHub (https://github.com/jjemerson/TemperatureGradient). RNAi was designed using the E-RNAi server http://www.dkfz.de/signaling/ernai3/.

Author Contributions

MC and JJE conceived the project, designed the experiments, and wrote the paper. MC collected the sequencing data, assembled the A4 genome, designed the pipelines for calling SVs, and genotyped variants from genome alignment. NWV conceived and carried out the RNAi experiments. RZ performed the selective sweep analysis. RZ and JJE conceived and analyzed CNV genotypes based on paired end Illumina reads and RZ analyzed the frequencies of *Cyp6a17, Cyp28d1* and *Ugt86Dh*. XZ and MC measured the paralog specific expression pattern. SK generated the DNA for Bionano optical data.

Competing financial interests

The authors declare no competing financial interest.

consensus emerges that such structural variants (SVs) are important factors in the genetics of complex traits[2]. Addressing this problem requires compiling an accurate and complete catalog of genome features relevant to phenotypic variation, a goal most readily achieved by comparing nearly complete, high-quality genomes[7]. While development of high-throughput short-read sequencing led to a steep drop in cost and a commensurate increase in the pace of sequencing[8], it also led to a focus on single-nucleotide changes and small/insertion deletions (indels)[3,9]. Paradoxically, this also produced a deterioration of the contiguity and completeness in new genome assemblies, due primarily to read-length limitations[10].

Here we present a reference-quality assembly of a second *D. melanogaster* strain called A4 and introduce a comprehensive map of SVs that reveals a vast amount of hidden variation exceeding that due to SNPs and small indels, and includes strong candidates for explaining complex traits. The A4 strain is a part of the Drosophila Synthetic Population Resource (DSPR)[11], a resource for mapping phenotypically relevant variants. We assembled the new A4 genome using high-coverage (147×) long reads using Single Molecule Real-Time sequencing of DNA extracted from females (Supplementary Fig. 1), following an approach shown to yield complete and contiguous assemblies[12]. The A4 assembly is more contiguous than release 6 of the ISO1 strain[13] — which is arguably the best metazoan WGS assembly — with 50% of the genome contained in contiguous sequences (contigs) 22.3 Mbp in length or longer (Supplementary Figs. 2–3). Compared to ISO1, the A4 assembly comprises far fewer sequences (161 scaffolds vs. 1,857 non-Y scaffolds[14]) while maintaining comparable completeness (Supplementary Table 1)[15]. Both genomes are co-linear across all major chromosome arms, making large-scale misassembly unlikely (Fig. 1a). An optical map of the A4 genome also supports this (Supplementary Figs. 4–5).

Putative SVs were identified by classifying regions of disagreement in a genome-wide pairwise alignment between A4 and ISO1 assemblies as indels, copy number variants (CNVs), or inversions (Table 1). Reads spanning SVs show that genotyping error is rare (<2.5%; Supplementary Table 2). However, because extremely long repeats common in heterochromatin require specialized approaches for assembly and validation[16], we focus on euchromatin (Supplementary Table 3). We discovered 1,890 large (>100 bp) indels (Supplementary Table 4; Supplementary Fig. 6) affecting more than 7 Mbp. In contrast, mutations <100 bp affected only 1.4 Mbp (indels: 722 kbp; SNPs: 687 kbp). Among large indels, 79% (1,486/1,890) are transposable element (TE) insertions (Supplementary Figs. 7–17). A previously published catalog of TE insertions in A4 based on 70× short-read coverage failed to find 38% of the TE insertions in A4 reported here[17] (Fig. 1b, Supplementary Fig. 18, Supplementary Table 5). These insertions invisible to short-read approaches often (34%) occur when a TE is inserted near another TE, resulting in complex non-uniquely mapping reads that are difficult to interpret. One such insertion is found in the A4 allele of the gene Multidrug-Resistance like Protein 1 (*MRP*), which is a candidate gene for resistance to chemotherapy drug carboplatin[18] (Supplementary Fig. 17).

Many TE insertions affect introns (395/718 in ISO1, 435/768 in A4), often dramatically lengthening them (Fig. 1c; Supplementary Fig. 19). Additionally, TEs inserted into exons can be spliced out, effectively becoming new introns. We see evidence of this in cDNA from ISO1[19] and RNAseq reads in A4 that span TE insertions >1 kbp into exons in the other

genome (Supplementary Table 6; Supplementary Figs. 20–22), representing the first genome-wide glimpse of TE-derived introns segregating in a population. TE insertions within introns are associated with decreased transcription[20], possibly caused by a phenomenon called intron delay, which slows transcription in long introns[21]. TE insertions can affect phenotype directly[22], perhaps by modulating or disrupting the expression of important genes. Since most TEs are rare in *D. melanogaster*[23], they are poorly tagged by common variants, complicating GWAS approaches for mapping traits, mirroring results from human GWAS[24].

Non-TE insertions represent 20% of ISO1 and 23% of A4 insertions, accounting for 170 kbp of sequence variation. Though these mutations are much smaller than TEs (median 213 bp versus 4.7 kbp), they often affect genes and 23% even escape detection by short reads (Fig. 1b). For example, among both hidden and visible deletions are 18 genes that are present in ISO1 and partially or completely absent in A4 (Supplementary Table 7), including *Cyp6a17* (Fig. 2a, Supplementary Fig. 23). Knockouts of *Cyp6a17* in a previous study increased cold preference[25]. Indeed, A4 prefers colder temperatures than a strain carrying an intact copy of *Cyp6a17* (Fig. 2b, Supplementary Fig. 24). Furthermore, this mutation is more common than expected of a deleterious allele (Fig. 2c), suggesting that it plays an important role in how flies respond to temperature in the wild. One deletion missed by short-read genotyping removes the second exon (and 41 amino acids of the encoded protein) of *Mur18B*, a chitin-binding protein gene conferring resistance to high-temperature stress[26], (Supplementary Fig. 25), likely rendering the A4 *Mur18B* allele defective.

We discovered 27 inversions affecting 60 kbp of sequence, ranging from 100 bp to 21 kbp (Supplementary Table 4), only 4 of which are detected by paired-end methods (Fig. 1b, Supplementary Table 5). These inversions often (21/27) affect regions harboring genes, including 21 kbp spanning five gustatory receptor genes: *Gr22a, Gr22b, Gr22c, Gr22d*, and *Gr22e* (Supplementary Table 4). While such clusters of related sequences may obscure read mapping information used to detect inversions, we could not find genomic features that might explain why the other inversions were missed. The A4 optical map revealed a putative inversion not resolved by the A4 assembly occupying 300 kbp of the proximal end of the X chromosome scaffold (Supplementary Figs. 4–5). Failure to resolve this inversion is not unexpected, because assembly methods tuned for euchromatin perform poorly in heterochromatic regions[16].

We discovered 390 CNVs (209 in A4 and 181 in ISO1) affecting ~600 kbp (Fig. 1d, Supplementary Figs. 26–36, Supplementary Table 4). While some CNVs were missed by paired-end methods due to spacer sequences between copies that are longer than the library fragments (Fig. 3a,d), most (~90%) were missed because they occur in complex tandem repeats (Supplementary Fig. 37). Unlike indels, most CNVs (64%) affect exons. Additionally, short-read CNV genotyping methods missed 13/34 protein coding genes that are duplicated in A4. In total, only ~40% of CNVs were discoverable with high specificity split-read and read-orientation methods[27,28] (Fig. 1b, Supplementary Fig. 38). Consistent with previous observations[29], coverage-based methods are extremely non-specific (Supplementary Fig. 38) and were therefore excluded from analysis. We next compared published gene expression data from larvae of A4 to that of a DSPR strain called A3[30],

revealing 17 duplicate genes with elevated expression (Supplementary Table 8), including genes previously identified as candidates for cold adaptation, olfactory response, and toxin resistance, among others (Fig. 3a, 3d, Supplementary Tables 8–9). Interestingly, eight of these CNVs were invisible to short-read methods (Supplementary Table 8).

A longstanding concern in trait mapping studies is failure to genotype candidate mutations[2]. Because A4 is a parental line of the DSPR trait mapping panel[31], we can confront this problem directly. Among the eight duplicate genes in A4 with elevated expression that escape detection, *Cyp28d1* and *Ugt86Dh* fall under QTLs for resistance to nicotine, a plant defense toxin[30,32]. One QTL (Q1) contains two cytochrome P450 enzyme genes, *Cyp28d1* and *Cyp28d2*, both of which are upregulated[30]. The other major effect candidate region contains the *Ugt86D* gene cluster, which possesses several differentially regulated genes, including *Ugt86Dh* (Fig. 3d–e). Candidate mutations like these are of obvious interest to researchers trying to dissect any trait, and yet were not visible in the initial study[30].

In the A4 assembly, Q1 contains a 3,755 bp tandem duplication separated by a 1.5 kbp spacer region, creating two copies of *Cyp28d1* (Fig. 3a; Supplementary Figs. 39–41). We compared paralog-specific expression levels of the *Cyp28d1* copies in A4 to that of the single copy in A3. In the absence of nicotine, the proximal and distal copies exhibit ~41-fold and ~6.3-fold higher expression in A4 versus A3, respectively (Fig. 3b). The intervening spacer sequence proved to be the 5' end of *Accord*, a long terminal repeat (LTR) retrotransposon (Fig. 3a). Insertion of *Accord* upstream of another gene called *Cyp6g1* has been linked to upregulation of its Cytochrome P450 enzyme[33], suggesting that it may be responsible for the upregulation rather than the tandem duplication of the *Cyp28d* gene. The second nicotine resistance QTL contains several *Ugt* genes, including *Ugt86Dh*, which were previously implicated in increased resistance to DDT[34]. Interestingly, we find that *Ugt86Dh* is duplicated in A4 (Fig. 3d; Supplementary Figs. 42–43) and escapes detection by paired-end short reads (Supplementary Table 5). Though several *Ugt* genes in Q4 show higher expression in resistant A4 larvae than in sensitive A3 larvae[30] (Fig. 3e), candidate variants explaining these differences have yet to be identified.

Because nicotine analogs are widely used pesticides, we predict resistance mutations to be common, mirroring observations about DDT. Indeed, we find four duplicate alleles spanning *Cyp28d1* and *Cyp28d2* segregating at intermediate to high frequencies in multiple populations (Fig. 3c) in a 25 kbp region where we expect duplicate heterozygosity to be less than 0.1. Similarly, the single *Ugt86Dh* duplicate allele segregates at high or intermediate frequency in nearly all populations we examined[6] (Fig. 3f). Finally, patterns of SNP variation surrounding both *Cyp28d1* and *Ugt86Dh* are consistent with recent bouts of natural selection (Supplementary Figs. 44–45), suggesting recent adaptation to nicotinoids.

While we focus on genetic variation in A4 relative to ISO1, there is no biologically meaningful sense in which any individual of a species is a more appropriate reference than another. Yet despite the prevalence of heritable phenotypic variation, functional work often describes results derived from diverse genotypes as applying to an entire species[35]. Approaches like RNAi or CRISPR require precise sequence information about their targets that can be easily misled by hidden SV. One study on the origin of new genes in *Drosophila*

argues that new genes rapidly become essential, even reporting a new gene called *p24-2* so young that it is present only in *D. melanogaster*[36]. Experiments targeting *p24-2* using RNAi constructs suggested that, although new, *p24-2* is essential. However, *p24-2* is absent in eight of ten strains we examined, including A4 and Oregon-R (Supplementary Figs. 46–47), questioning its essentiality to *D. melanogaster*. Because the original construct actually targets both *p24-2* and its essential paralog *eca*[37,38] (Supplementary note), we tested two other constructs targeting *p24-2*, neither of which showed any viability reduction (Supplementary Table 10), bolstering the suggestion that *p24-2* is not essential.

The ubiquity of hidden variation in genome structure is merely a first glimpse beneath the tip of an iceberg of genetic variation governing phenotypes. Together with careful phenotypic measurements, a new generation of high-quality genomes will reveal previously invisible heritable phenotypic variation. Our results show that popular genotyping approaches miss a significant fraction of SVs (Fig. 1b, Supplementary Figs. 18 and 38, Supplementary Table 5), including those affecting gene expression and organismal phenotype (Supplementary Tables 8–9), suggesting that previous estimates of the contribution to regulatory[39] and phenotypic variation by SVs are misleading[40]. The extensive hidden variation we observe segregates in *D. melanogaster*, a species likely harboring fewer complex structural features than humans or livestock and crop species like wheat and maize. Consequently, we suggest that the true medical and agricultural impact of structural variation is likely much greater than the already considerable estimates made without recourse to multiple reference-grade assemblies[29].

Online Methods

DNA sequencing and genome assembly

A4 DNA was extracted from females and used in SMRTbell library preparation following[12]. We sequenced this library on 30 SMRTcells using P6-C4 chemistry on a Pacific Biosciences RSII platform at the UC Irvine Genomics Facility, yielding 18.7 Gb of sequence. We then followed[12] to assemble the A4 genome. We assembled a draft genome using PBcR-MHAP[41] in *wgs* 8.3rc1 and PacBio reads only (NG50 = 13.9 Mb, 147 Mb total), then generated a hybrid assembly with *DBG2OLC*[42] using the longest 30× PacBio reads and 75× paired end Illumina reads from[12] (assuming 130 Mb genome size; NG50 = 4.23 Mb, 129 Mb total). We merged the two assemblies using *quickmerge* v0.1 with default settings except hco = 5, c = 1.5, and l = 2 Mb. The merge yielded an assembly (NG50 = 21.3 Mb, 130 Mb total) that was both smaller than expected[43] and smaller than the PacBio-only assembly. Therefore, we added contigs unique to the PacBio assembly to the hybrid assembly using *quickmerge* as above but with I = 5 Mb. Finally, we generated the final assembly by running *finisherSC*[44] with default settings, polishing the assembly twice with *quiver* (smrtanalysis v2.3), and finally finishing with *Pilon* v1.3[45] (using A4 reads from[12]). This yielded a final assembly of 144 Mb with N50 = 22.3 Mb (Supplementary Table 1).

Bionano data

A4 embryos less than 12h old were collected on apple juice/agar Petri dishes, dechorionated using 50% bleach, rinsed with water, then stored at −80 °C. DNA was extracted from frozen

embryos using the Animal Tissue DNA Isolation kit (Bionano Genomics, San Diego, CA). Bionano Irys optical data was generated and assembled with IrysSolve 2.1 at Bionano Genomics. We then merged the Bionano assembly with the final assembly contigs (above) using IrysSolve, retaining Bionano assembly features when the two assemblies disagreed.

Comparative scaffolding

The A4 assembly was scaffolded with the software *mscaffolder* (see URL) using the release 6 *D. melanogaster* genome (r6.09) assembly[13] as the reference. Prior to scaffolding, transposable elements and repeats in both assemblies were masked using default settings for *Repeatmasker (v4.0.6)*. The repeatmasked A4 assembly was aligned to the repeatmasked major chromosome arms (X,2L,2R,3L,3R,4) of *D. melanogaster* ISO1 assembly using *MUMmer*[46]. Alignments were further filtered using the delta-filter utility with the -m option and the contigs were assigned to the specific chromosome arms based on the mutually best alignment. Contigs showing less than 40% of the total alignment for any chromosome arms could not be assigned a chromosomal location and therefore were not scaffolded. The mapped contigs were ordered based on the starting coordinate of their alignment that did not overlap with the preceding reference chromosome-contig alignment. Finally, the mapped contigs were joined with 100 Ns, a convention representing assembly gaps. The unscaffolded sequences were named with a 'U' prefix.

BUSCO analysis

We used *busco* (v1.22)[15] to evaluate completeness and accuracy of the A4 and ISO1 release 6 assemblies. ISO1 contains 5 BUSCOs (BUSCOaEOG75R3J9, BUSCOaEOG7SJRJ9, BUSCOaEOG7SJRK2, BUSCOaEOG7WMR0H, BUSCOaEOG71S8ZH) that are missing from the A4 assembly. To validate the absence of these 5 BUSCOs in the A4 assembly, the full-length sequence of the ISO1 genes (*Ftz-f1, CG7627, Raw, Maf1, Cv*-c) were downloaded from FlyBase[14] and searched against the A4 assembly with *MUMmer*. Surprisingly, *MUMmer* found all five 'missing BUSCOs' in the A4 assembly in single copies. Consequently, the BUSCO counts for A4 were adjusted accordingly.

Structural variant detection

CNVs via whole genome alignment—We aligned ISO1 and A4 using *MUMmer*[46] (mummer -mumreference -l 20 -b), then clustered maximal exact matches (MEMs) between the two *mgaps* (*mgaps* -C -s 200 -f .12 -l 100). The *l* parameter in *mgaps* was set to 100 to detect duplicates that are 100bp or longer. We used a pipeline called *svmu* (Structural Variants from MUMmer; see URL) to automate CNV detection from overlapping *mgaps* clusters. When reference sequence regions in two separate alignment clusters overlapped, the overlapping segment of the reference sequence regions was inferred as duplicated in the query sequence. This approach can also identify 1) a duplicated sequence that is present in the both genomes but has diverged due to the presence of repeats or indels and 2) CNVs containing TE sequences. We filtered the latter using *Repeatmasker* (v4.0.6). We identified false positives duplication calls by aligning the putatively duplicated reference sequences back to ISO1 and A4 genomes using *nucmer* (*nucmer* –maxmatch –g 200) and then counting the copy number using *checkCNV*, which is also included in the *svmu* pipeline. *svmu* was

run with the default parameters; *checkCNV* was run with c = 500 (max copy number 500), qco = 10000 (10 kbp of insertion/deletion allowed within a copy), rco = 0.2 (unaligned length of up to 20% of the sequence length between reference and query copies is allowed). CNVs occurring within 2 kbp of each other were designated as "complex events" and combined (*bedtools* merge –d 2000)[47] for the purpose of counting total CNVs present in the genome (Supplementary Table 11). However, total sequence affected by CNVs was counted before merging. Functional annotation of CNVs was based on gene annotation of ISO1 release 6.

Indels via whole genome alignment—Insertions (>100 bp) in A4 appear as alignment gaps between two adjacent syntenic blocks when ISO1 is aligned to A4 (and vice versa). We aligned A4 to ISO1 using *nucmer* (default parameters), then identified adjacent syntenic blocks with gaps >100 bp between them in A4 but <10% the gap length in ISO1. Indel detection was carried out by the *svmu* utility *findInDel*. A deletion was inferred for a specific gene (e.g. *Cyp6a17*), when an ortholog of the gene was present in closely related species *D. simulans*.

Inversions via whole genome alignment—We identified inversions in the A4 genome by aligning it to the ISO1 genome using *nucmer* (-mumreference), then processing the outputted delta file using *findInDe*l. A4 regions that ran in the reverse direction with respect to ISO1 were recorded as inversions. TEs were removed from this list using *Repeatmasker* annotations for ISO1.

Genotyping CNVs, indels, and inversions using Illumina reads—Three common, complementary strategies are typically employed to discover CNVs using paired-end Illumina reads: read depth, read pair mapping orientation, and split-read mapping[7]. We identified duplications (100bp to 25kbp long) in A4 using 70× paired-end reads[11] with *CNVnator*[48] for read depth, *pecnv*[28] for read pair orientation, and *Pindel*[49] for split-read mapping approaches. We mapped reads to ISO1 release 6 using *bwa mem* for *CNVnator* and *pindel* and *bwa aln* for *pecnv*[50]. We required at least 3 supporting read pairs for *pecnv* calls[28] and used a bin size 100 for *CNVnator* due to the data's high coverage. Furthermore, we used *CNVnator* and *Pindel to identify* large (>100bp) indels and *Pindel* to identify inversions. We manually compared these short-read-based calls to our alignment-based CNV calls for all of chromosome arm 2L.

TE insertion coordinates for A4 were obtained from flyrils.org[17]. We manually compared our TE insertion calls and those from[17] for all of chromosome arm 2L.

SNP and small indel detection

SNPs and small (<100bp) indels in the A4 assembly were identified using the *show-snps* utility from *MUMmer*[46]. We aligned A4 scaffolds to ISO1 scaffolds using *nucmer (-mumreference*), then filtered repeats using *delta-filter* in conjunction with the –r and –q options. SNPs and small indels were called from the filtered data using *show-snps* with –Clr options.

Validation of duplicates and indels

Dotplots between A4 and ISO1 for all SV loci on chromosome arm 2L were manually inspected to confirm the accuracy of the *MUMmer*-based genotyping. All manually inspected loci corresponded to the automated genotype calls. To quantify the effect of assembly errors in A4 on SV calls, we required that unassembled corrected long reads from A4 agree with the A4 assembly in the region spanning the entire mutation. To do this, we mapped the *PBcR-MHAP* corrected long reads to the A4 assembly using *blasr* v1.3.1.142244 (-bestn 1 –sam) and identified all reads spanning the mutation region with anchors in the flanking sequence of at least 250 bp on each side. For our stringent validation criteria, we require at least two fully spanning reads to overlap each SV (Supplementary Fig. 48A). These fully spanning reads must possess at least 99.5% alignment coverage ($P_{Aligned}$) and less than a ratio 0.005 of gaps to read length (R_{Gaps}; Supplementary Fig. 48A). For our standard validation criteria, we permit validation under the following relaxed criteria: 1) overlap spanning reads (at least two on each side) that otherwise fit the stringent criteria above; 2) fully spanning reads must possess at least 97.5% alignment coverage ($P_{Aligned}$) and less than a ratio 0.025 of gaps to read length (R_{Gaps}; Supplementary Fig. 48B).

Half of our sequencing data is present in reads of 17,885 bp or longer, which is enough to achieve more than 60-fold coverage across the entirety of the euchromatin, and more than 10-fold coverage of the genome in reads 30 kbp or longer. Such long reads contain unique sequence flanking each side of the mutation as well as the mutation breakpoints and the mutation itself, making this a powerful approach to validating SV calls.

PCR validation

We assayed presence and absence of *Cyp28d1* and *p24-2* copies using PCR (Supplementary Table 12; Supplementary Figs. 41 and 47). We extracted DNA from 25 flies from each strain using Magattract HMW DNA kit (Qiagen) and used Phusion (New England Biolabs) for PCRs and an amplification time of 15 seconds for the *Cyp28d1* PCRs and 30 seconds for *p24-2* PCRs.

Temperature preference assay

We created a linear temperature gradient on a solid aluminum bar (total dimensions: 24” × 4” × 4”) by placing 4” of one end of the bar inside a reservoir containing ice water (0°C) and 4” of the other end inside a reservoir containing warm water (35°C) (Supplementary Fig. 24). This left ~40 cm of aluminum bar exposed between the baths. Temperatures along the bar were measured by 11 temperature sensors (Tmp36 analog temperature sensors from Adafruit) evenly spaced at 4 cm intervals sealed into holes drilled into the bar and secured with thermal epoxy (OMEGABOND 101 Two-Part Epoxy). The probes were connected to three 4-channel 16-bit analog-to-digital converters (ADS1115 from Adafruit), which were in turn calibrated and monitored by a Raspberry Pi 3 single-board computer. Automated temperatures were recorded every second using a custom Python script (see URLs) during the experiment to verify the stability of the gradient. The temperature measurements at the end of the experiment were used in assigning temperatures to individual flies. The temperature gradient on the aluminum bar ranged from 9°C to 30°C (Fig. 2b). We compared the preference of A4, which lacks the *Cyp6a17* gene, to w^{1118} (BDSC stock 5905), which

has an intact copy of *Cyp6a17*[25]. We collected groups of 100 1–3 days old flies of mixed sexes and kept them at 25°C for 24 hours. Before the assay, flies were immobilized with light anesthesia and placed between a thin aluminum sheet cut into the shape of the aluminum bar surface and an acrylic lid possessing a partition to create two "lanes" for the flies to behave without interacting with each other. Quinine sulfate was applied to the roof and walls of each channel in the lid so that flies avoided these surfaces and were constantly contacting the aluminum surface. Flies were allowed to recover on the aluminum sheet in a 25°C incubator for 40 minutes after anesthesia. The aluminum sheet was then placed on top of the aluminum bar and left for 40 minutes in the dark. A photo was taken to record the positions of the flies on the block after 40 minutes. We recorded fly positions and interpolated their temperatures using linear regression based on temperature probe readings.

Statistical analyses

We replicated the temperature preference assay experiment six times. Three replicates were conducted with A4 in lane 1 and w^{1118} in lane 2, and three replicates were conducted with the lane assignments reversed. We performed a nonparametric Wilcoxon rank-sum test, which does not assume a particular distribution for the data, on each of these six replicates to test for a difference in temperature preference between the two strains. These six individual tests produced *p*-values of 2.12e-10, 6.76e-10, 1.89e-06, 9.21e-14, 1.96e-06, and 1.25e-24. To obtain a combined *p*-value, we performed a meta-analysis using Fisher's method, producing a very low meta-*p*-value ($p << 10^{-16}$).

RNAi strain construction and screening

Strain 60100 (Vienna Drosophila Resource Center) contains two attP sites, at 2L: 22,019,296 (near tiptop; VIE260B) and 2L:9,437,482 (VIE260B-2). Activation of RNAi constructs inserted into VIE260B results in ectopic activation of tiptop and phenotypes independent of the RNAi target[51]. PCR screening showed that KK109179 contained insertions at both sites and likely caused the lethal phenotype observed by[36] (Supplementary Fig. 49). We removed the insertion at VIE260B following the crossing scheme outlined by[51] and kept two of the resulting lines with insertions only at VIE260B-2 (Supplementary Fig. 49).

We generated a new *p24-2* RNAi line following[52]. We designed the RNAi construct CG33105_RNAi using the E-RNAi server (see URLs). CG33105_RNAi was the only possible construct >50 bp with 100% of the possible 19-mers uniquely matching *p24-2*. CG33105_RNAi was cloned into pKC26, then injected into 60100 at 250 ng/μL. We isolated transformants using Bloomington Drosophila Stock Center (BDSC) balancer stock 9325, ensuring that the RNAi construct was inserted only at VIE260B-2 using PCR54. NV-CG33105-2 and NV-CG33105-6 are derived from different transformants, but carry the same CG33105_RNAi construct. We drove RNAi using lines constitutively expressing GAL4 under control of *Act5C* or *αTub84B* promoters (BDSC lines 4414 and 5138). Five males and five virgin driver females were allowed to cross for 9 days at 25 °C and 12h:12h light:dark cycle, then removed from vials. F1s were counted 19 days after crossing. The proportion of wild-type (RNAi-active) F1s was compared to the proportion of wild-type F1s from control crosses between 60100 males and driver strains. We confirmed presence of the

p24-2 duplicate in each of these lines using PCR (Supplementary Table 12) and Sanger sequencing.

Expression analysis

Genome-wide gene expression difference between A3 and A4 larvae were analyzed following[30]. Sequences of the A3 genes were obtained from an A3 genome assembly constructed with publicly available A3 Illumina paired end reads. To compare the expression levels of the *Cyp28d1, CG7742*, and *Ugt86Dh* gene copies, we aligned publicly available 100bp RNAseq reads[30] to A4 mRNA sequences using *bowtie2*[53] (with --score-min L,0,0 to ensure that only perfectly-aligned unique, i.e. copy-specific, reads were kept for FPKM calculations). We adjusted transcript length by subtracting the length of regions to which no SNP-covering read aligned, because only reads overlapping SNPs could be included in FPKM calculations. For example, *Cyp28d1* gene copies are distinguishable by 15 SNPs. When regions that cannot be spanned by perfectly-aligned unique reads are removed from the effective transcript length, 310bp are subtracted from the total 1509bp transcript length, leaving an effective transcript length of 1199bp. Similarly, for *Ugt86Dh* and *CG7742*, transcript lengths of 1065 bp and 755bp were used to calculate FPKM, respectively. No such adjustments were made for the single copy genes not segregating for duplications. The total number of reads aligned to the genomes was calculated based on the alignment of the single-ended RNAseq reads aligned to the A4 and A3 genomes using *TopHat*[54].

Testing for selective sweeps

We used the composite likelihood ratio (CLR) statistic of SweepFinder2 v1.0 to test for recent selective sweeps[55,56]. CLR values were calculated using the frequency of SNPs present in each sample over a grid with 250 bp increments. Sites were polarized using *D. simulans, D. yakuba*, and *D. erecta*. Invariant sites that differed from the inferred ancestral state (substitutions) were included in the analysis, thus improving power and robustness to bottlenecks[55,57]. The significance of the results was evaluated by comparing the CLR values to 100 coalescent neutral simulations generated using *ms*[58]. Estimates of the effective population size, neutral mutation rate, and recombination rate were taken from previous publications[59]. The 95% confidence intervals were computed using the largest CLR values from each neutral simulation.

Estimating duplicate allele frequencies

The frequency of duplicate alleles was estimated from next-generation Illumina data (see supplementary note) by analyzing the density of divergently mapped read pairs. Reads were mapped against the release 6 ISO1 reference genome using *bwa mem*[50]. Divergent read pairs were selected by taking the complement of paired reads in the BAM file that mapped with proper orientation, defined as pairs of reads that mapped to the same chromosome on opposite strands and were flagged by the aligner as being properly aligned with respect to the each other. Duplications were called for samples that showed a clear peak and high signal-to-noise ratio in the coverage density for divergent read pairs at breakpoints surrounding genes that were found to be duplicated in A4. The divergent read pair signals for several duplicate alleles for *Cyp28d1* from various populations are shown in Supplementary Fig. 50. Samples with low genomic coverage (less than 10 Mb over the

chromosome containing the duplication) or inferred to be identical by descent to other samples over a region containing the duplication, using estimates of homozygous coverage and IBD from[60], were excluded from analysis. Populations were excluded from this analysis if they contained fewer than 10 samples.

Data availability

All single molecule sequence data has been deposited to NCBI SRA and can be found under the accession number SRX2729308. The A4 scaffolded assembly has been deposited in NCBI WGS under the accession no. GCA_002300595.1. All the variant calls are provided in the supplementary files.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank L.T. Ngo, J. Yan, and A. Yue for help with fly maintenance and SV analysis and J. Mohammed for providing the multiple sequence alignment of the Drosophila species group. We thank A. Carrillo, E. Azizi, D. German, and M. McHenry for assistance in assembling the temperature gradient instrument, N. Nirale for uploading the sequencing data and assembly, B. Gaut, G.C.G. Lee, A. Long, K. Thornton, and three anonymous reviewers for thoughtful comments on the manuscript, and M. Long for discussion and for permission to use the RNAi data. The work was supported by US National Institutes of Health (NIH) grant R01GM123303-1 and University of California, Irvine setup funds (J.J.E), National Science Foundation (NSF) Graduate Research Fellowships (DGE-1321846 to R.Z. and DGE-1144082 to N.W.V), and NIH Genetics and Regulation Training Grant T32-GM007197 (N.W.V). 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.

References

1. Rockman MV. THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S GOLD DOES NOT GLITTER. Evolution. 2012; 66
2. Eichler EE, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics. 2010; 11:446–50.
3. Wray NR, et al. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013; 14:507–15. [PubMed: 23774735]
4. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–53. [PubMed: 19812666]
5. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–69. [PubMed: 18398418]
6. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science. 2008; 320:1629–31. [PubMed: 18535209]
7. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–76. [PubMed: 21358748]
8. The human genome at ten. Nature. 2010; 464:649–650. [PubMed: 20360688]
9. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009; 10:241–51. [PubMed: 19293820]
10. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nature methods. 2011; 8:61–5. [PubMed: 21102452]

11. King EG, et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 2012; 22:1558–66. [PubMed: 22496517]
12. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016
13. Hoskins RA, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome research. 2015; 25:445–58. [PubMed: 25589440]
14. dos Santos G, et al. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015; 43:D690–7. [PubMed: 25398896]
15. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31:3210–2. [PubMed: 26059717]
16. Khost DE, Eickbush DG, Larracuente AM. Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. bioRxiv. 2016
17. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. Mol Biol Evol. 2013; 30:2311–27. [PubMed: 23883524]
18. King EG, Kislukhin G, Walters KN, Long AD. Using Drosophila melanogaster To Identify Chemotherapy Toxicity Genes. Genetics. 2014; 198:31-+. [PubMed: 25236447]
19. Stapleton M, et al. The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res. 2002; 12:1294–300. [PubMed: 12176937]
20. Cridland JM, Thornton KR, Long AD. Gene expression variation in Drosophila melanogaster due to rare transposable element insertion alleles of large effect. Genetics. 2015; 199:85–93. [PubMed: 25335504]
21. Swinburne IA, Silver PA. Intron delays and transcriptional timing during development. Dev Cell. 2008; 14:324–30. [PubMed: 18331713]
22. Long AD, Lyman RF, Morgan AH, Langley CH, Mackay TFC. Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete-scute complex are associated with variation in bristle number in Drosophila melanogaster. Genetics. 2000; 154:1255–1269. [PubMed: 10757767]
23. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. Population genomics of transposable elements in Drosophila melanogaster. Molecular biology and evolution. 2011; 28:1633–44. [PubMed: 21172826]
24. Lohmueller KE, et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. Am J Hum Genet. 2013; 93:1072–86. [PubMed: 24290377]
25. Kang J, Kim J, Choi KW. Novel Cytochrome P450, cyp6a17, Is Required for Temperature Preference Behavior in Drosophila. Plos One. 2011; 6
26. MacMillan HA, et al. Cold acclimation wholly reorganizes the Drosophila melanogaster transcriptome and metabolome. Scientific Reports. 2016; 6
27. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–71. [PubMed: 19561018]
28. Rogers RL, et al. Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. Mol Biol Evol. 2014; 31:1750–66. [PubMed: 24710518]
29. Huddleston J, Eichler EE. An Incomplete Understanding of Human Genetic Variation. Genetics. 2016; 202:1251–4. [PubMed: 27053122]
30. Marriage TN, King EG, Long AD, Macdonald SJ. Fine-mapping nicotine resistance loci in Drosophila using a multiparent advanced generation inter-cross population. Genetics. 2014; 198:45–57. [PubMed: 25236448]
31. King EG, et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome research. 2012; 22:1558–66. [PubMed: 22496517]
32. Glendinning JI. How do herbivorous insects cope with noxious secondary plant compounds in their diet? Entomologia Experimentalis Et Applicata. 2002; 104:15–25.

33. Chung H, et al. Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1. Genetics. 2007; 175:1071–7. [PubMed: 17179088]
34. Pedra JHF, McIntyre LM, Scharf ME, Pittendrigh BR. Genom-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:7034–7039. [PubMed: 15118106]
35. mod EC, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–97. [PubMed: 21177974]
36. Chen S, Zhang YE, Long M. New genes in Drosophila quickly become essential. Science. 2010; 330:1682–5. [PubMed: 21164016]
37. Saleem S, et al. Drosophila melanogaster p24 trafficking proteins have vital roles in development and reproduction. Mechanisms of Development. 2012; 129:177–191. [PubMed: 22554671]
38. Bartoszewski S, Luschnig S, Desjeux I, Grosshans J, Nusslein-Volhard C. Drosophila p24 homologues eclair and baiser are necessary for the activity of the maternally expressed Tkv receptor during early embryogenesis. Mechanisms of Development. 2004; 121:1259–1273. [PubMed: 15327786]
39. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–853. [PubMed: 17289997]
40. Gamazon ER, Nicolae DL, Cox NJ. A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. Plos Genetics. 2011; 7
41. Berlin K, et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015; 33:623–30. [PubMed: 26006009]
42. Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci Rep. 2016; 6:31900. [PubMed: 27573208]
43. Hoskins RA, et al. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. 2002; 3:RESEARCH0085. [PubMed: 12537574]
44. Lam KK, LaButti K, Khalak A, Tse D. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. Bioinformatics. 2015; 31:3207–9. [PubMed: 26040454]
45. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9:e112963. [PubMed: 25409509]
46. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]
47. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014; 47:11 12 1–34.
48. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research. 2011; 21:974–984. [PubMed: 21324876]
49. Ye K, Schulz MH, Long Q, Apweiler R, Ning ZM. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. [PubMed: 19561018]
50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
51. Green EW, Fedele G, Giorgini F, Kyriacou CP. A Drosophila RNAi collection is subject to dominant phenotypic effects. Nature Methods. 2014; 11:222-+. [PubMed: 24577271]
52. Dietzl G, et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature. 2007; 448:151–6. [PubMed: 17625558]
53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357–U54. [PubMed: 22388286]
54. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012; 7:562–578. [PubMed: 22383036]

55. Nielsen R, et al. Genomic scans for selective sweeps using SNP data. Genome Res. 2005; 15:1566–75. [PubMed: 16251466]

56. DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics. 2016; 32:1895–7. [PubMed: 27153702]

57. Huber CD, DeGiorgio M, Hellmann I, Nielsen R. Detecting recent selective sweeps while controlling for mutation rate and background selection. Molecular ecology. 2016; 25:142–56. [PubMed: 26290347]

58. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–8. [PubMed: 11847089]

59. Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. Drosophila melanogaster recombination rate calculator. Gene. 2010; 463:18–20. [PubMed: 20452408]

60. Lack JB, et al. The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics. 2015; 199:1229–41. [PubMed: 25631317]

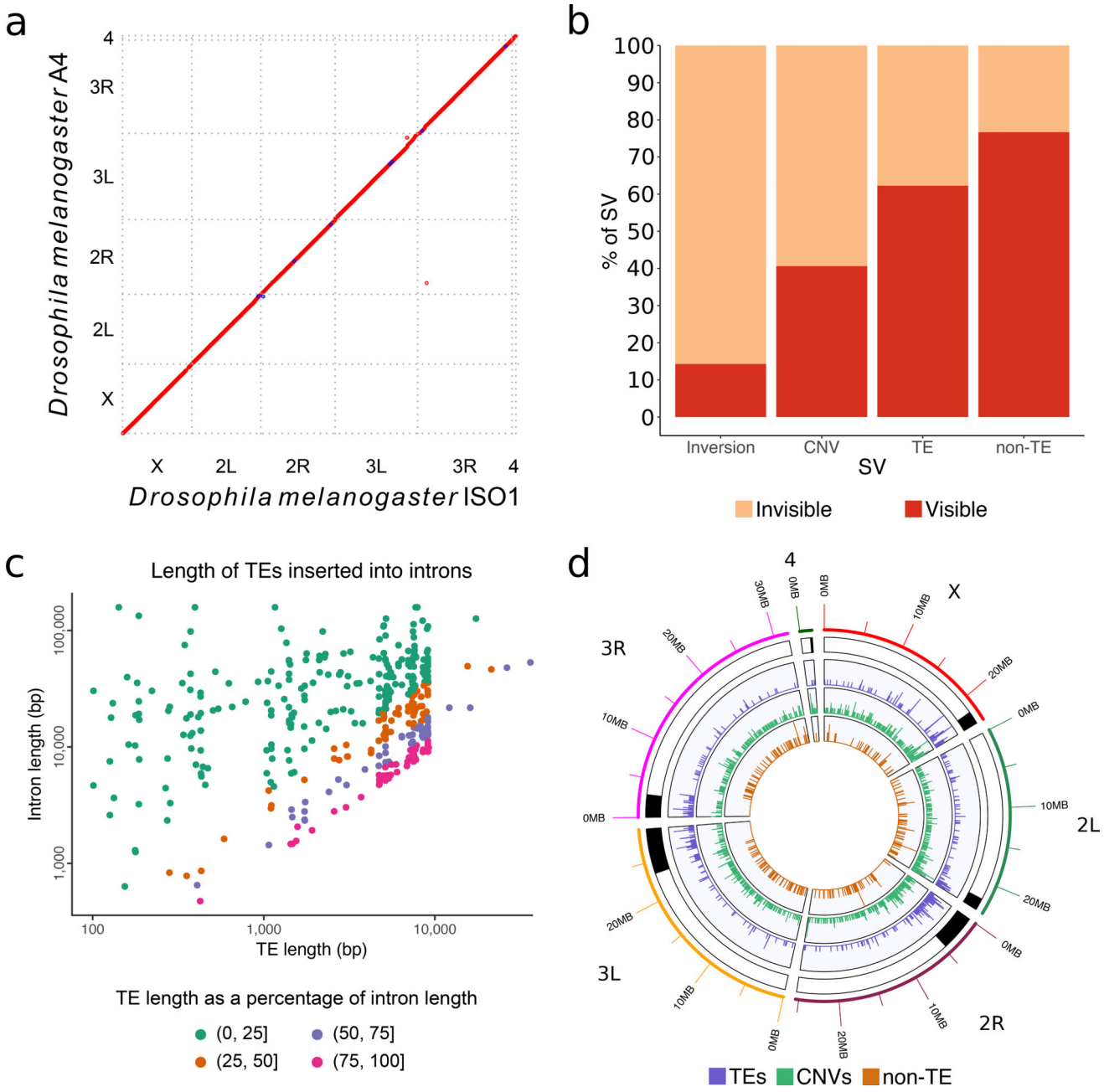

Figure 1. A4 assembly quality and structural variation (SV)

a) Dot plot between the *D. melanogaster* reference (ISO1) and A4 assemblies. The A4 assembly is as contiguous as the ISO1 assembly (scaffold N50 = 25.4Mb vs 25.2Mb; Supplementary Table 1). Repeats and transposable elements were masked to highlight the correspondence of the two genomes. **b)** The proportions of large (>100 bp) SVs in the A4 chromosome 2L assembly relative to ISO1 2L that were identified (visible) or missed (invisible) by short read methods (Online Methods). **c)** Relationship between the length of TEs in ISO1 (median 5.1 kbp) and the lengths of the introns they are inserted into. Nearly equal intron and TE lengths indicate that many introns are comprised of mainly TEs. **d)** Distribution of SVs (>100 bp) across A4 chromosome arms. Track 1 shows pericentric

heterochromatin (black). Tracks 2–4 show TEs, duplicate CNVs (relative to ISO1), and non-TE indels greater than 100 bp, respectively. CNVs and TEs are present in higher densities in heterochromatin, whereas non-TE indels are less numerous.

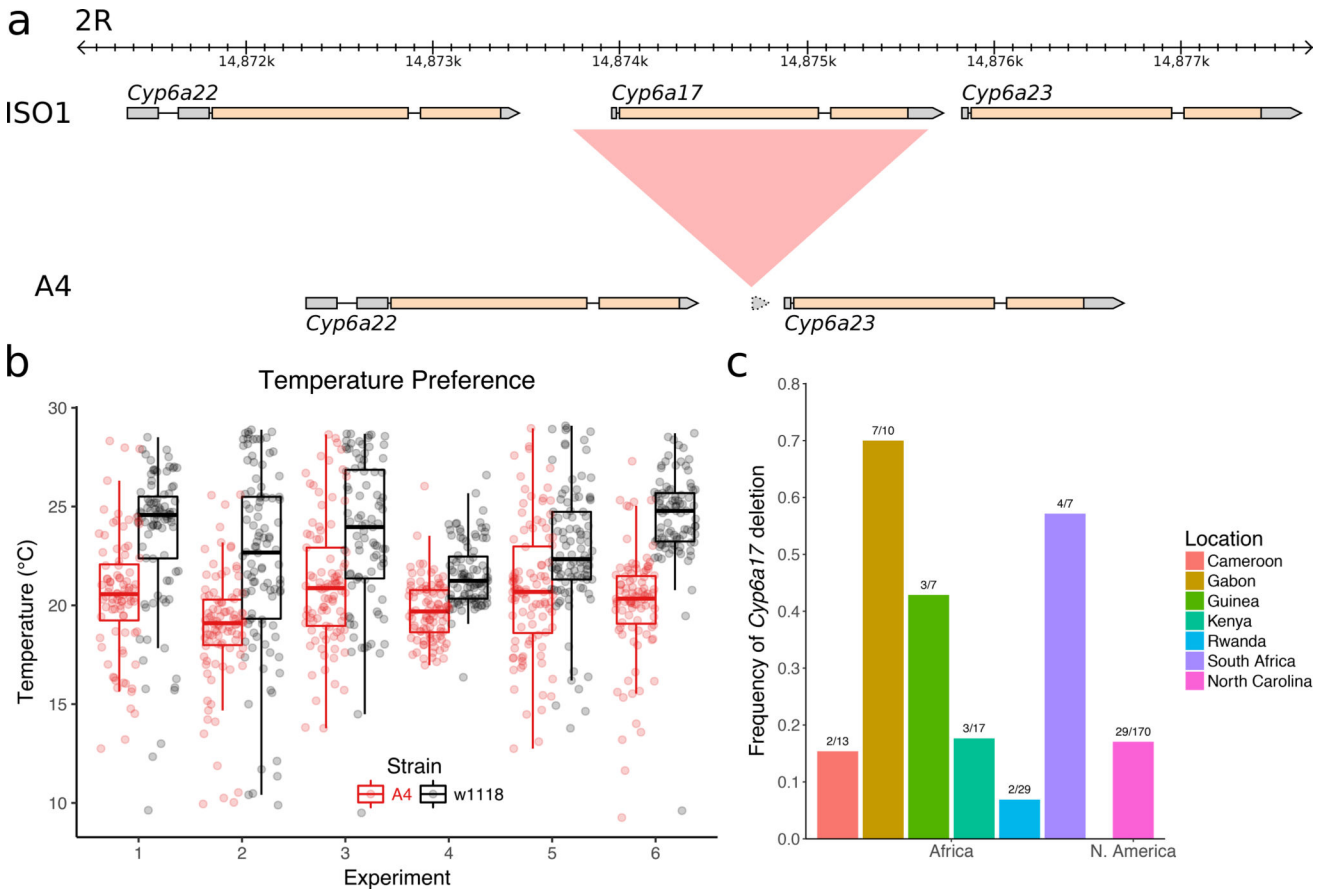

Figure 2. Copy number variation of *Cyp6a17* and its functional consequences

a) *Cyp6a17* is deleted in A4 relative to ISO1. Alignment between annotated ISO1 and A4 assemblies on 2R shows a large ISO1 region (red) missing in A4. Gene models are shown (gray - noncoding, yellow - coding). **b)** Temperature preference of strains A4 (Δ*Cyp6a17*) and w^{1118} (*Cyp6a17*$^{+}$; ref. 23). Preference was measured by recording the position of flies along a linear 8°–30°C temperature gradient after an adjustment period (Online Methods). Each dot represents the position of a fly along the gradient. Each experiment number is an independent pairwise trial. A4 flies occupy colder regions of the gradient than w1118 flies (Fisher's method on Wilcoxson rank-sum tests, meta-*p*-value << 10^{-16}). Upper and lower hinges of the boxplots represent 25% and 75 % quantiles, respectively; upper whisker = largest observation less than or equal to upper hinge + 1.5 * IQR; lower whisker = smallest observation greater than or equal to lower hinge − 1.5 * IQR; middle = median, 50% quantile. **c)** Frequency of the *Cyp6a17* deletion in African (DPGP2) and North American (DGRP) populations.

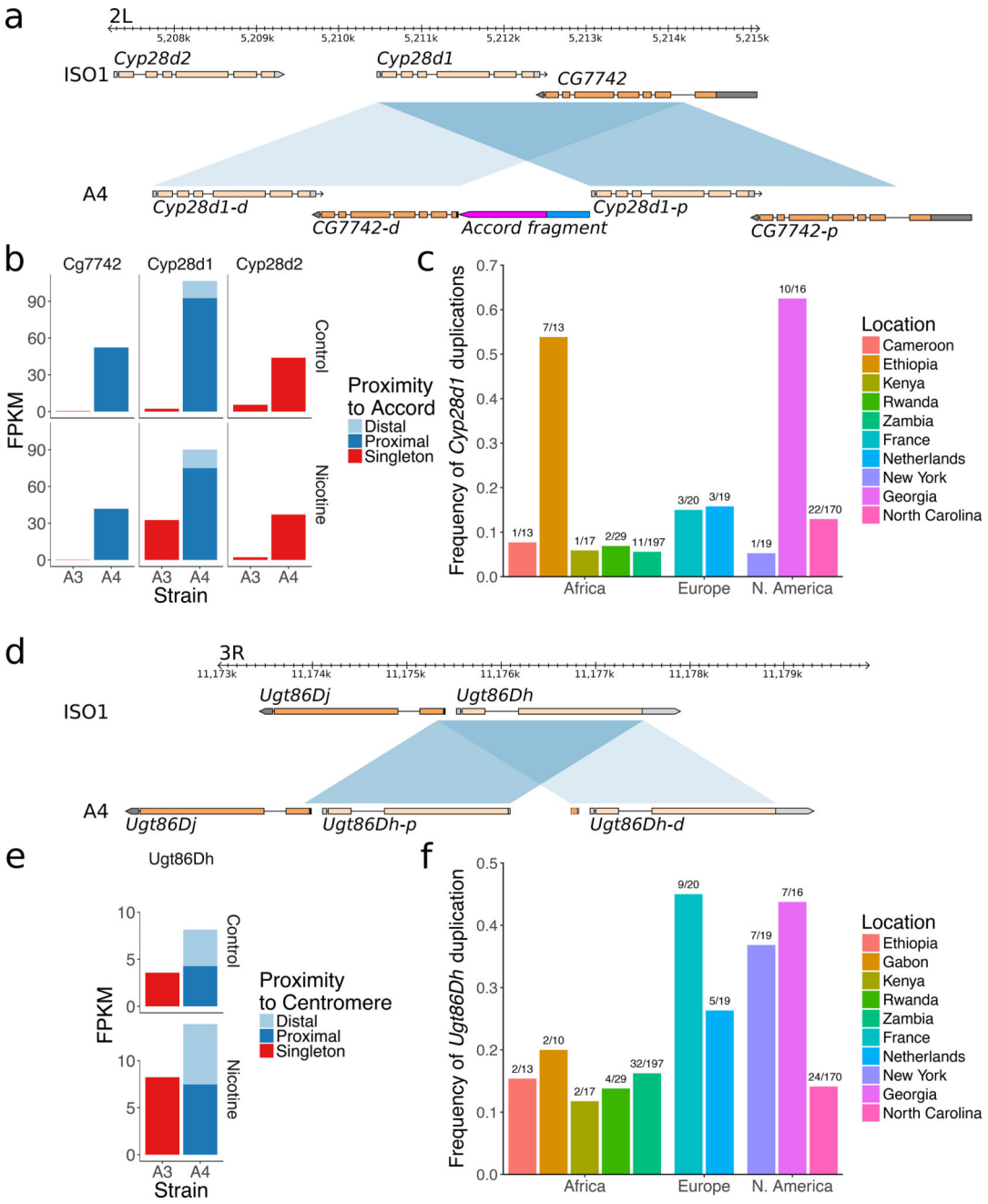

Figure 3. Copy number variation in the *Ugt86Dh* and *Cyp28d1* and its effect on gene expression variation

Shaded parallelograms (distal copy: light blue, proximal copy: dark blue) indicate the single and duplicated regions in ISO1 and A4, respectively. **a)** Duplication of *Cyp28d1* and *CG7742* in A4. ISO1 and strain A3 possess one copy of *Cyp28d1*, whereas A4 has two copies. A 1.5 kbp *Accord* fragment (pink) containing an LTR (blue) is located between the proximal *Cyp28d1* and the distal *CG7742*. Gene models are shown with gray (non-coding) and orange (coding) rectangles. **b)** Paralog specific expression of candidate QTL genes at Q1 in A4 and A3 in the presence of nicotine in the food. *CG7742* and *Cyp28d1* copies

located nearer the *Accord* element are transcribed at higher levels than those more distal. **c)** Combined frequency of four *Cyp28d* duplicate alleles in African (DPGP2, DPGP3) and North American populations. **d)** Tandem duplication of *Ugt86Dh* in A4 created *Ugt86Dh-d*. **e)** In contrast to *Cyp28d1* duplicates, both copies of *Ugt86Dh* are expressed at similar levels and their expression nearly doubles in the presence of nicotine. **f)** Frequency of the *Ugt86Dh* duplicate.

Table 1

Number of different types of SVs uncovered by A4-ISO1 genome alignment.

Mutation type (>100 bp)	# of mutation in A4 euchromatin
Insertion (non-TE)	768
Deletion (non-TE)	718
Insertion (TE)	223
Deletion (TE)	181
CNV (more copy)	209
CNV (less copy)	181
Inversion	27