Hidden genetic variation shapes the structure of functional elements in Drosophila

Mahul Chakraborty1*, Nicholas W. VanKuren2, Roy Zhao3,4, Xinwen Zhang1,3, Shannon Kalsow1 and J. J. Emerson2,4*

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

Mutations underlying phenotypic variation remain elusive in trait-mapping studies despite the exponential accumulation of genomic data, suggesting that many causal variants are invisible to current genotyping approaches8,9. In fact, mutations like duplications, deletions, and transpositions8,10 are systematically under-represented by standard methods, even as a consensus emerges that such structural variants (SVs) are important factors in the genetics of complex traits. Addressing this problem requires compiling an accurate and complete catalog of the genomic features that are relevant to phenotypic variation, a goal most readily achieved by comparing nearly complete high-quality genomes. Although the development of high-throughput short-read sequencing led to a steep drop in cost and a commensurate increase in the pace of sequencing8, it also led to a focus on single-nucleotide changes and small indels3,9. Paradoxically, this has also resulted in deterioration of the contiguity and completeness of new genome assemblies, due primarily to read-length limitations10.

Here we present a reference-quality assembly of a second D. melanogaster strain called A4 and introduce a comprehensive map of SVs, which identifies a large amount of hidden variation exceeding that due to SNPs and small indels, and which includes strong candidates to explain 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 through single-molecule real-time sequencing of DNA extracted from females (Supplementary Fig. 1), following an approach that has been shown to yield complete and contiguous assemblies12. The A4 assembly is more contiguous than release 6 of the ISO1 strain13—which is arguably the best metazoan whole-genome sequence assembly—with 50% of the genome contained in contiguous sequences (contigs) 22.3 Mb in length or longer (Supplementary Figs. 2 and 3). As compared to the ISO1 assembly, the A4 assembly comprises far fewer sequences (161 scaffolds versus 1,857 non-Y-chromosome scaffolds13) while maintaining comparable completeness (Supplementary Table 1). The two genomes are collinear across all major chromosome arms, making large-scale misassembly unlikely (Fig. 1a). An optical map of the A4 genome also supported the accuracy of the assembly (Supplementary Figs. 4 and 5).

We identified putative SVs by classifying regions of disagreement in a genome-wide pairwise alignment of the A4 and ISO1 assemblies as indels, copy number variants (CNVs), or inversions (Table 1). Reads spanning SVs showed that genotyping error was rare (<2.5%; Supplementary Table 2). However, because extremely long repeats are common in heterochromatin and require specialized approaches for assembly and validation14, we focused on euchromatin (Supplementary Table 3). We discovered 1,890 large (>100-bp) indels (Supplementary Fig. 6 and Supplementary Table 4), which affected more than 7 Mb. In contrast, mutations <100 bp in length affected only 1.4 Mb (indels, 722 kb; SNPs, 687 kb). Among large indels, 79% (1,486/1,890) were transposable element (TE) insertions (Supplementary Figs. 7–17). A previously published catalog of TE insertions in A4 based on 70× short-read coverage15 failed to find 38% of the TE insertions in A4 reported here (Fig. 1b, Supplementary Fig. 18, and Supplementary Table 5). These insertions, which are invisible to short-read approaches, often occur (in 34% of instances) when a TE is inserted near another TE, resulting in complex, non-uniquely mapping reads that are difficult to interpret. One such insertion was found in the A4 allele of the MRP gene (encoding multidrug-resistance-like protein 1), which is a candidate gene for resistance to the chemotherapeutic drug carboplatin16 (Supplementary Fig. 17).

We found that many TE insertions affected introns (395/718 in ISO1, 435/768 in A4), often greatly lengthening them (Fig. 1c and Supplementary Fig. 19). Additionally, TEs inserted into exons can be spliced out, effectively becoming new introns. We saw evidence of this in cDNA from ISO117 and in RNA-seq reads in A4 that showed exon junctions flanking TE insertions (Supplementary Figs. 20–22 and Supplementary Table 6), which represents a genome-wide view of TE-derived introns segregating in a population. TE insertions within introns are associated with decreased transcription18, possibly caused by a phenomenon called intron delay, which slows transcription in long introns. TE insertions can affect phenotype directly19, perhaps by modulating or disrupting the expression of important genes. Because most TEs are rare in D. melanogaster20,
they are poorly tagged by common variants, complicating genome-wide association study (GWAS) approaches for mapping traits; this mirrors similar complications in human GWAS.

Non-TE insertions represented 20% of ISO1 and 23% of A4 insertions, and they accounted for 170kb of sequence variation (Fig. 1d and Table 1). Although these mutations were much smaller than TEs (median 213 bp versus 4.7 kb), they often affected genes, and 23% even escaped detection by short reads (Fig. 1b). For example, among both hidden and visible deletions, there were 18 genes that were present in ISO1 and partially or completely absent in A4 (Supplementary Table 7), including Cyp6a17 (Fig. 2a and Supplementary Fig. 23). Knockout of Cyp6a17 in a previous study increased cold preference25. Indeed, A4 flies preferred colder temperatures than flies from a strain carrying an intact copy of Cyp6a17 (Fig. 2b and Supplementary Fig. 24). Furthermore, this mutation was more common than expected for a deleterious allele (Fig. 2c), suggesting that it has a role in regulating how flies respond to temperature in the wild. One deletion missed by short-read genotyping removed the second exon of Mur18B (and 41 amino acids of the encoded chitin-binding protein that confers resistance to high-temperature stress26) (Supplementary Fig. 25), likely rendering the A4 Mur18B allele defective.

We discovered 27 inversions, ranging from 100 bp to 21 kb in length (Supplementary Table 4), that affected 60 kb of sequence, and were present in higher densities in heterochromatin as compared to euchromatin, whereas non-TE indels are less numerous in heterochromatin.

Table 1 | Number of different types of structural variants uncovered by aligning the A4 and ISO1 genomes

| Mutation type (>100 bp) | Number of mutations in A4 euchromatin |
|-------------------------|-------------------------------------|
| Insertion (TE)          | 768                                 |
| Deletion (TE)           | 718                                 |
| Insertion (non-TE)      | 223                                 |
| Deletion (non-TE)       | 181                                 |
| CNV (more copies in A4) | 209                                 |
| CNV (fewer copies in A4)| 181                                 |
| Inversion               | 27                                  |

TE, transposable element; CNV, copy number variation.
Copy number variation of Cyp6a17 and its functional consequences. a. Cyp6a17 is deleted in the A4 genome relative to the ISO1 genome. Alignment between annotated ISO1 and A4 assemblies on chromosome arm 2R shows a large ISO1 region (red) missing in A4. Gene models are shown (gray indicates noncoding sequences, and yellow indicates coding sequences). b. Temperature preference of strains A4 (∆Cyp6a17) and w^118 (intact Cyp6a17^17^). Preference was measured by recording the position of 100 flies along a linear 8 °C–30 °C temperature gradient after an adjustment period (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 w^118 flies (Fisher’s method on Wilcoxon rank-sum tests, meta P value ≪ 10^−5^). Upper and lower hinges of the box plots represent 25% and 75% quantiles, respectively; the upper whisker indicates the largest observation less than or equal to the upper hinge + 1.5 times the interquartile range (IQR); the lower whisker indicates the smallest observation greater than or equal to the lower hinge – 1.5 times the IQR; and the middle horizontal bar indicates the median, 50% quantile. c. Frequency of the Cyp6a17 deletion in African (DPGP2) and North American (DGRP) populations.

We discovered 390 CNVs (209 in A4 and 181 in ISO1) that affected ~600 kb (Fig. 1d, Supplementary Figs. 26–36, and Supplementary Table 4). Although some CNVs were missed by paired-end methods owing to spacer sequences between copies that were longer than the library fragments (Fig. 3a,d), most (~90%) of the CNVs were missed because they occurred in complex tandem repeats (Supplementary Fig. 37). Unlike indels, most CNVs (64%) affected exons. Additionally, short-read CNV genotyping methods missed 13 of 34 protein-coding genes that were duplicated in A4. In total, only ~40% of CNVs were discoverable with high-specificity split-read and read-orientation methods^27,28 (Fig. 1b and Supplementary Fig. 38). Consistent with previous observations^29, coverage-based methods were extremely nonspecific (Supplementary Fig. 38) and were therefore excluded from analysis. We next compared published gene expression data from larvae of A4 to expression data for a DSPR strain called A3^30 and identified 17 A4 duplicate genes that are single copy in ISO1 with increased expression (Supplementary Table 8), including genes previously identified as candidates for cold adaptation, olfactory response, and toxin resistance, among others (Fig. 3a,d and Supplementary Tables 8 and 9). Notably, 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^1. Because A4 is a parental line of the DSPR trait-mapping panel^31, we could confront this problem directly. Among the eight duplicate genes with increased expression in A4 that escaped detection, Cyp28d1 and Ugt86Dh fell under quantitative trait loci (QTLs) for resistance to nicotine, a plant defense toxin^30,31. One QTL (Q1) contains two genes, Cyp28d1 and Cyp28d2, that encode cytochrome P450 enzymes, both of which were upregulated^30. The other candidate region that showed a major effect contains the Ugt86D gene cluster, which includes 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 they were not visible in the initial study^30.

In the A4 assembly, Q1 contains a 3,755-bp tandem duplication in which the duplicated regions are separated by a 1.5-kb spacer, resulting in two copies of Cyp28d1 (Fig. 3a and Supplementary Figs. 39–41). We compared paralog-specific expression levels of the Cyp28d1 copies in A4 to expression of the single copy in A3. In the
Fig. 3 | Copy number variation in Ugt86Dh and Cyp28d1 and its effect on gene expression variation. Shaded parallelograms (light blue, distal copy; dark blue, proximal copy) correspond to the single and duplicated regions in ISO1 and A4. a, Schematic showing duplication of Cyp28d1 and CG7742 in A4. ISO1 and strain A3 possess one copy of Cyp28d1, whereas A4 has two copies. A 1.5-kb Accord fragment (pink) containing an LTR (blue) is located between the proximal Cyp28d1 and the distal CG7742. Gene models are shown with gray (noncoding) and orange (coding) rectangles. b, Paralog-specific expression of candidate QTL genes at Q1 in A4 and A3 in the presence and absence (control) of nicotine in the food. CG7742 and Cyp28d1 copies located nearer the Accord element are transcribed at higher levels than those that are more distal. FPKM, fragments per kilobase of transcript per million mapped reads. c, Combined frequency of four Cyp28d1 duplicate alleles in African (DPGP2 and DPGP3) and North American populations. d, Schematic showing that tandem duplication of Ugt86Dh in A4 created Ugt86Dh-d. e, Paralog-specific expression of candidate QTL gene Ugt86Dh in A4 and A3 in the presence and absence (control) of nicotine in the food. In contrast to Cyp28d1 duplicates, the two copies of Ugt86Dh are expressed at similar levels, and their expression nearly doubles in the presence of nicotine. f, Frequency of the Ugt86Dh duplicate in African (DPGP2 and DPGP3) and North American populations.
Letters

absence of nicotine, the proximal and distal copies in A4 exhibited ~41-fold and ~6.3-fold higher expression, respectively, than the single copy in A3 (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 the encoded cytochrome P450 enzyme32, suggesting that the retrotransposon may be responsible for the upregulated expression rather than the tandem duplication of the Cyp28d gene. The second nicotine-resistance QTL contains several Ugt genes, including Ugt86Dh, which have previously been implicated in increased resistance to the pesticide DDT26. Of note, we found that Ugt86Dh was duplicated in A4 (Fig. 3d and Supplementary Figs. 42 and 43); this mutation escaped detection by paired-end short reads (Supplementary Table 5). Although several Ugt genes in the Q4 QTL showed higher expression in nicotine-resistant A4 larvae than in sensitive A3 larvae24 (Fig. 3e), candidate variants that explain these differences have yet to be identified.

Because nicotine analogs are widely used pesticides, we predict that resistance-conferring mutations are common, mirroring observations for DDT. Indeed, we found that four duplicate alleles spanning Cyp28d1 and Cyp28d2 segregated at intermediate to high frequencies in multiple populations (Fig. 3c) in a 25-kb region where we expected duplicate heterozygosity to be less than 0.1. Similarly, the single duplicate allele of Ugt86Dh segregated at high or intermediate frequency in nearly all of the populations we examined (Fig. 3f). Finally, patterns of SNP variation surrounding both Cyp28d1 and Ugt86Dh are consistent with recent bouts of natural selection (Supplementary Figs. 44 and 45), suggesting recent adaptation to nicotine.

Although 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 individuals with diverse genotypes as applying to an entire species35. Approaches like RNA interference (RNAi) or gene editing with CRISPR require precise sequence information about their targets and can be easily misled by hidden structural variation. One study on the origin of new genes in D. melanogaster argues that new genes rapidly become essential, and the authors even report a new gene called p24-2 that is so young that it is present in only D. melanogaster33. Experiments targeting p24-2 using RNAi constructs suggested that, although new, p24-2 is essential. However, p24-2 was absent in eight of the ten strains we examined, including A4 and Oregon-R (Supplementary Figs. 46 and 47), which calls into question its essential nature in D. melanogaster. Because the original construct actually targeted both p24-2 and its essential paralog eca36–38 (Supplementary Note), we tested two other constructs targeting p24-2, neither of which resulted in any reduction in viability (Supplementary Table 10), thus bolstering the suggestion that p24-2 is not essential.

The ubiquity of hidden variation in genome structure is merely an indication of the extent of the underlying genetic variation governing phenotypes. Together with careful phenotypic measurements, a new generation of high-quality genomes will identify 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, and Supplementary Table 5), including ones that affect gene expression and organisinal phenotype (Supplementary Tables 8 and 9), suggesting that previous estimates of the contribution of SVs to regulatory39 and phenotypic variation are misleading38. The extensive hidden variation we observe segregates in D. melanogaster, a species that likely harbors fewer complex structural features than humans or livestock, as well as crop species like wheat and maize. Consequently, we suggest that the true medical and agricultural impact of structural variation is likely to be much greater than the already considerable estimates made without recourse to multiple reference-grade assemblies38.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-017-0010-y.

Received: 15 March 2017; Accepted: 10 November 2017; Published online: 18 December 2017

References

1. Rockman, M. V. The QTN program and the alleles that matter for evolution: all that’s gold does not glitter. Evolution 66, 1–17 (2012).
2. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450 (2010).
3. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14, 507–515 (2013).
4. Caicedo, A. L. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
5. McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9, 356–369 (2008).
6. Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. & Long, M. Natural selection shapes genome-wide patterns of copy number polymorphism in Drosophila melanogaster. Genetics 139, 1629–1631 (2008).
7. Alkan, C., Cole, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376 (2011).
8. Anonymous. The human genome at ten. Nature 464, 649–650 (2010).
9. 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).
10. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61–65 (2011).
11. King, E. G. et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genes Res. 22, 1538–1566 (2012).
12. 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. 44, e147 (2016).
13. Hoskins, R. A. et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genes Res. 25, 445–458 (2015).
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. 43, D690–D697 (2015).
15. Simão, 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).
16. Khoost, D. E., Eickbush, D. G. & Larraquente, A. M. Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. Preprint at bioRxiv https://doi.org/10.1101/054155 (2016).
17. 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).
18. King, E. G., Kisulkhin, G., Walters, K. N. & Long, A. D. Using Drosophila melanogaster to identify chemotherapy toxicity genes. Genetics 198, 313–323 (2013).
19. Stapleton, M. et al. The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res. 12, 1294–1300 (2002).
20. Cridland, J. M., Thornton, K. R. & Long, A. D. Gene expression variation in Drosophila melanogaster due to rare transposable element insertion alleles of large effect. Genetics 199, 85–93 (2015).
21. Swinburne, I. A. & Silver, P. A. Intron delays and transcriptional timing during development. Dev. Cell 14, 324–330 (2008).
22. Long, A. D., Lyman, R. F., Morgan, A. H., Langley, C. H. & Mackay, T. F. C. Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at theachaete–scute complex are associated with variation in bristle number in Drosophila melanogaster. Genetics 154, 1255–1269 (2000).
23. Petrov, D. A., Fiston-Lavié, A.-S., Lipatov, M., Lenkov, K. & González, J. Population genomics of transposable elements in Drosophila melanogaster. Mol. Biol. Evol. 28, 1633–1644 (2011).
24. Lohmueller, K. E. 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. 93, 1072–1086 (2013).
25. Kang, J., Kim, J. & Choi, K. W. Novel cytochrome P450, cyp6a17, is required for temperature preference behavior in Drosophila. PLoS One 6, e19889 (2011).
26. MacMillan, H. A. et al. Cold acclimation wholly reorganizes the Drosophila melanogaster transcriptome and metabolome. Sci. Rep. 6, 28999 (2016).

© 2017 Nature America Inc., part of Springer Nature. All rights reserved.
27. Ye, K., Schulz, M. H., 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 25, 2865–2871 (2009).
28. Rogers, R. L. et al. Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. Mol. Biol. Evol. 31, 1750–1766 (2014).
29. Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. Genetics 202, 1251–1254 (2016).
30. Marriage, T. N., King, E. G., Long, A. D. & Macdonald, S. J. Fine-mapping nicotine resistance loci in Drosophila using a multiparent advanced generation inter-cross population. Genetics 198, 45–57 (2014).
31. Glendinning, J. I. How do herbivorous insects cope with noxious secondary plant compounds in their diet? Entomol. Exp. Appl. 104, 15–25 (2002).
32. 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 175, 1071–1077 (2007).
33. Pedra, J. H. F., McIntyre, L. M., Scharf, M. E. & Pittendrigh, B. R. Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant Drosophila. Proc. Natl. Acad. Sci. USA 101, 7034–7039 (2004).
34. modENCODE Consortium. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787–1797 (2010).
35. Chen, S., Zhang, Y. E. & Long, M. New genes in Drosophila quickly become essential. Science 330, 1682–1685 (2010).
36. Saleem, S. et al. Drosophila melanogaster p24 trafficking proteins have vital roles in development and reproduction. Mech. Dev. 129, 177–191 (2012).
37. Bartoszewski, S., Luschnig, S., Desjues, I., Grosshans, J. & Nüsslein-Volhard, C. Drosophila p24 homologues eclair and baiser are necessary for the activity of the maternally expressed Tkv receptor during early embryogenesis. Mech. Dev. 121, 1259–1273 (2004).
38. Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, 848–853 (2007).
39. Gamazon, E. R., Nicolae, D. L. & Cox, N. J. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genet. 7, e1001292 (2011).

Acknowledgements

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, and K. Thornton 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 (J.J.E.), 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-GM07197 (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 S10RR025496-01, S10OD010794-01, and S10OD021718-01.

Author contributions

M.C. and J.J.E. conceived the project, designed the experiments, and wrote the paper. M.C. collected the sequencing data, assembled the A4 genome, designed the pipelines for calling SVs, and genotyped variants from genome alignment. N.W.V. conceived and performed the RNAi experiments. R.Z. performed the selective sweep analysis. R.Z. and J.J.E. conceived and analyzed CNV genotypes based on paired-end Illumina reads, and R.Z. analyzed the frequencies of Cyp6a17, Cyp28d1, and Ugt86Dh. X.Z. and M.C. measured the paralog-specific expression patterns. S.K. generated the DNA for the Bionano optical data.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-017-0010-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C. or J.J.E.

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/.
Methods

DNA sequencing and genome assembly. A4 DNA was extracted from females and used in SMRTbell library preparation as described previously. We sequenced this library on 30 SMRTcells using P6-C4 chemistry on a Pacific Biosciences RSII platform at the University of California High-Throughput Genomics Facility, yielding 18.7 Gb of sequence. We then followed the method described previously to assemble the A4 genome. We assembled a draft genome using PBcR-MHAP in vgs 8.3.1c and PacBio reads (NG50 = 13.9 Mb, 147 Mb in total; NG50 is the contig length such that 50% of an assembled contig assembly is contained within contigs of this length or longer) and then generated a hybrid assembly with DBG2OLC using the longest 30x PacBio reads and 75x paired-end Illumina reads from ref. 6, (assuming a genome size of 130 Mb; NG50 = 4.23 Mb, 129 Mb in 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 in total) that was both smaller than expected and smaller than the PacBio-only assembly. Therefore, we identified contigs that were likely to the hybrid assembly using quickmerge as described above but with I = 5 Mb. Finally, we generated the final assembly by running finisherSC with default settings, polishing the assembly twice with quiver (SMRT Analysis v2.3), and with Pilon v1.3 (using A4 reads from ref. 6). This yielded a final assembly of 144 Mb with N50 = 22.3 Mb (Supplementary Table 1).

Bionano data. A4 embryos less than 12 h old were collected on Petri dishes containing apple juice and agar, dechorionated using 50% bleach, rinsed with water, and stored at −80 °C. DNA was extracted from frozen embryos using the Animal Tissue DNA Extraction kit (Bionano). Bionano Irys data were generated and assembled with IrysSolve 2.1 at Bionano Genomics. We then merged the Bionano assembly with the final assembly contigs (described in "DNA sequencing and genome assembly") using IrysSolve, retaining Bionano assembly features when the two assemblies disagreed.

Comparative scaffolding. The scaffold for the A4 assembly was prepared with the software mscassembler (see URLs) using the release 6. D. melanogaster genome (r6.09) assembly as the reference. Prior to scaffolding, TEs and repeats in both assemblies were masked using default settings for RepeatMasker (v4.0.6). The repeat-masked A4 assembly was aligned to the repeat-masked major chromosome arms (X, 2L, 2R, 3L, 3R, and 4) of the D. melanogaster ISO1 assembly using MUMmer. Alignments were further filtered using the delta-filter utility with the -m option, and the contigs were assigned to specific chromosome arms based on the basis of the mutually best alignment. Contigs showing less than 40% of the total alignment for any chromosome arm could not be assigned a chromosomal location and therefore were not scaffolded. The mapped contigs were ordered on the basis of 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 contigs were named with a ‘U’ prefix.

Benchmarking universal single-copy orthologs (BUSCO) analysis. We used BUSCO v3.0.3-d22 for the completeness and accuracy of the A4 and ISO1 release 6 assemblies. ISO1 contains five BUSCOs (BFC0aE0G78J3R9, BFC0aE0G75J9R19, BFC0aE0G75J9RK2, BFC0aE0G79W9ROH, and BFC0aE0G718Z82H) that are missing from the A4 assembly. To validate the absence of these five BUSCOs in the A4 assembly, we aligned the full-length sequences of the BUSCOs to the PacBio reads from 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 blast v3.1.1.42244 (−best 1 −sodm) and identified all of the reads that spanned the mutation-containing region with anchors in the flanking sequence of at least 250 bp on each side. For our stringent validation criteria, we required at least two fully spanning reads to overlap each SV (Supplementary Fig. 48a). Since 250 bp was 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 that were 30 kb or longer. Such long reads contained unique sequences 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 for the presence and absence of Cyp28d1 and p24-2 copies using PCR (Supplementary Figs. 41 and 47, and Supplementary Table 12). We extracted DNA from 25 flies from each strain using the Magattract HMW DNA kit (Qiagen), and we used Phusion (New England Biolabs) for PCRs that had an amplification time of 15 s for the Cyp28d1 reactions and 30 s for the p24-2 reactions.

Temperature-preference assay. We created a linear temperature gradient on a solid aluminum bar (total dimensions: 24 inches × 4 inches × 4 inches) by placing 4 inches of one end of the bar inside a reservoir containing ice water (0 °C) and 4 inches of the other end inside a reservoir containing warm water (35 °C)
Letters

Statistical analyses. We replicated the temperature preference assay experiment six times. Three replicates were conducted with A4 flies in lane 1 and w1118 flies 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 normal distribution, to compare the overall temperature preference between the two strains. The test was performed using the R statistical software package.

Estimating frequencies of duplicate alleles. The frequency of duplicate alleles was estimated from next-generation Illumina data (Supplementary Note) by analyzing the density of divergently mapped read pairs. Reads were mapped against the release 6 ISO1 reference genome using bwa-mem59. 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 each other. Duplicates 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 the A4 sequence. 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 (<100 Mb over the chromosome containing the duplication) or inferred to be identical by descent to other samples over a region containing the duplication, used estimates of homozygous coverage identity by descent from ref. 36, were excluded from analysis. Populations were excluded from this analysis if they contained fewer than ten samples.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. All single-molecule sequence data have been deposited to the NCBI Sequence Read Archive (SRA) and can be found under accession SRX2735574. The A4 scaffolded assembly has been deposited in the NCBI Assembly database under accession GCA_002300595.1. All the variant calls are provided in the supplementary files.

References
40. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630 (2015).
41. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: efficient assembly of large genomes using long error-prone reads of the third-generation sequencing technologies. Sci. Rep. 6, 31900 (2016).
42. Hoskins, R. A. et al. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. 3, RESEARCH0085.1–RESEARCH0085.16 (2002).
43. Lam, K. K., LaButti, K., Khachaturov, A. & Tse, D. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. Bioinformatics 31, 3207–3209 (2015).
44. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9, e112963 (2014).
45. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
46. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.2.1–11.2.12 (2014).
47. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator—an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984 (2011).
48. Li, H. & Durbin, R. Fast and accurate short-read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
49. Green, E. W., Fedele, G., Giorgini, F. & Kyriacou, C. P. A Drosophila RNAi collection is subject to dominant phenotypic effects. Nat. Methods 11, 222–223 (2014).
50. Dietzl, G. et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature 448, 151–156 (2007).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. 
   *Nat. Methods* 9, 357–359 (2012).
52. Trapnell, C. et al. Differential gene and transcript expression analysis of 
   RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 
   562–578 (2012).
53. Nielsen, R. et al. Genomic scans for selective sweeps using SNP data. *Genome 
   Res.* 15, 1566–1575 (2005).
54. DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. 
   SweepFinder2: increased sensitivity, robustness, and flexibility. *Bioinformatics* 
   32, 1895–1897 (2016).
55. Huber, C. D., DeGiorgio, M., Hellmann, I. & Nielsen, R. Detecting recent 
   selective sweeps while controlling for mutation rate and background 
   selection. *Mol. Ecol.* 25, 142–156 (2016).
56. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of 
   genetic variation. *Bioinformatics* 18, 337–338 (2002).
57. Fiston-Lavier, A. S., Singh, N. D., Lipatov, M. & Petrov, D. A. *Drosophila 
   melanogaster* recombination rate calculator. *Gene* 463, 18–20 (2010).
58. Lack, J. B. et al. The *Drosophila* genome nexus: a population genomic 
   resource of 623 *Drosophila melanogaster* genomes, including 197 from a 
   single ancestral range population. *Genetics* 199, 1229–1241 (2015).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size
   Describe how sample size was determined.
   - We followed a published temperature preference assay protocol (doi:10.1038/ng1513) and used the largest sample size (100) from the protocol.

2. Data exclusions
   Describe any data exclusions.
   - No data was excluded

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   - We performed the temperature preference assay six times independently using flies from different generations.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   - We collected 100 randomly chosen flies from each strain for each experiment. These flies were collected from a pool of >500 flies originating from 10 vials of adult flies from each strain. We switched the sides of the fly strains on the aluminum block to ensure a balanced design (A4 on the top panel 3/6 times and on the bottom panel 3/6 times and vice versa for w1118).

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - The flies at the end of the temperature preference assays were counted independently by at least two persons. The genotype was withheld to the counters.
   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   n/a | Confirmed
   - The exact sample size \( n \) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - A statement indicating how many times each experiment was replicated
   - The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - The test results (e.g. \( P \) values) given as exact values whenever possible and with confidence intervals noted
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
7. Software

Describe the software used to analyze the data in this study.

All custom software used in this study have been deposited in GitHub. The software used in this study include:
- PBcR-MHAP v8.3rc1
- DBG2OLC v1.0
- IrysSolve 2.1
- mscaffolder
- bedtools v2.25.0
- Repeatmasker v4.0.6
- BUSCO v1.22
- MUMmer v3.23
- SVMU 0.1beta
- smrtanalysis v2.3
- quickmerge v0.1
- Pilon v1.3
- CNVnatorv0.3
- Pindel v0.2.4
- Pecnv 0.1.8
- SweepFinder2

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All sequence data have been deposited to NCBI and are publicly available. The fly strains used in this study are available from Bloomington Stock Center and the transgenic fly strains we generated for the p24-2 RNAi experiments are available upon request.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A
c. Report whether the cell lines were tested for mycoplasma contamination.

N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

We used fruit fly D. melanogaster strains for our experiments.

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A