File S5

Expanded Methods Section

cDNA library construction and sequencing

For each parental strain (*mal0-sc2* and *bip3-isoA*), we extracted total RNA from pools of 20-30 randomly picked flies using the standard TRIzol (Invitrogen) protocol. RNA quality and concentration were measured on a Bioanalyzer (Agilent) and NanoDrop (Thermo Scientific). cDNA libraries were synthesized from 5 µg of total RNA using the MINT cDNA synthesis kit (Evrogen) and normalized using the TRIMMER kit (Evrogen) following the manufacturer's protocols. AMPure beads (Agencourt) were used for purification steps. We then used a Bioruptor (Diagenode) to shear the cDNA, and agarose gel electrophoresis to select a narrow range of ~300 bp fragments. Paired-end libraries were prepared from the sheared, normalized cDNA using the standard Illumina protocol, and sequenced with 85-base paired-end reads on an Illumina Genome Analyzer II at the UC Davis Genome Center (Table S1).

Data filtering and transcriptome assembly

Illumina and Evrogen adapter sequences used during library construction were trimmed from the Illumina reads using custom scripts. We then filtered out any reads shorter than 36 bases to minimize repetitive read mapping. Read quality appeared poor because many cDNA fragments contained Evrogen adapters, creating repetitive sequences that influenced basecalling probabilities. We therefore relaxed the standard Illumina 3' trimming of contiguous bases with Phred scores (COCK *et al.* 2010; EWING and GREEN 1998; EWING *et al.* 1998) of 15 or less, and fed the remaining reads into the assembly pipeline without quality filtering.

The partial transcriptomes of *mal0-sc2* and *bip3-isoA* were assembled *de novo* using ABySS (BIROL *et al.* 2009; MILLER *et al.* 2010; SIMPSON *et al.* 2009) followed by Trans-ABySS (ROBERTSON *et al.* 2010) to merge multiple unscaffolded assemblies with different K-mer sizes. The assembly procedure was as follows:

1. We ran paired-end ABySS assemblies with kmer sizes increasing from 29 to 63 in increments of two (n=10, q=0, s=160, c=2, ABYSS_OPTIONS= "--no-chastity --no-trim-masked --illumina-quality").

2. Next we ran Trans-ABySS phase 1, which leverages the BLAT (KENT 2002) algorithm, to exclude repetitive and inferior contigs generated by the individual assemblies.

3. We selected contigs 100 bp or longer from the combined assembly, and ran them through CAP3 (HUANG and MADAN 1999) in order to merge overlapping contigs.

4. We selected contigs 200 bp or longer, and removed contigs that were identical except for one mismatch, keeping the copy with the fewest ambiguous bases.

5. Finally, we grouped all the contigs by homology using NCBI BLASTn (ALTSCHUL *et al.* 1990; ALTSCHUL *et al.* 1997) (word_size= 14, gapopen=12, gapextend=4, penalty=-6, reward=4, minimal 110 bp overlap, minimal bitscore of 130, mismatch to identity ratio ≤ 0.05, allowing for large gaps), uniformly oriented all contigs within each group using a directed graph, and assigned unique identifiers to these groups (Table S2 and Files S1A-B and S2A-B).

This procedure produces a unique set of genes, some of which have multiple isoforms. Subsequent read mapping assigns most reads unambiguously to contigs, but can map the same read redundantly to a group of transcripts corresponding to a single alternatively spliced locus.

Identification of fixed SNP differences

For our genetic analysis, we required fixed differences (FD) between parental strains, i. e. nucleotide positions where *bip3-isoA* is fixed for one allele and *mal0-sc2* for a different allele. To identify FDs for the first round of genotyping, *mal0-sc2* and *bip3-isoA* read libraries were each mapped separately to both *bip3-isoA* and *mal0-sc2 de novo* transcriptome assemblies. Mapping was performed using SOAP2 (LI *et al.* 2009) (m=1, x=1200, v=3, l=40, g=300, s=50, r=2). In order to eliminate ambiguously mapping poly(A+) tails, we first aligned all reads to the *de novo* transcriptomes, then removed any reads that mapped to multiple contig groups and had GC content below 10%. Same-species mapping (*mal0-sc2* reads to *mal0-sc2* transcriptome and *bip3-isoA* reads to *bip3-isoA* transcriptome) serves to correct assembly errors, gauge coverage at each position, and identify nucleotide positions where multiple alleles segregate within each parental strain despite inbreeding. Cross-species mapping allows us to identify polymorphic positions and select FDs that meet coverage cut-offs (File S3).

FDs are positions in cross-species alignments where the consensus basecall is monomorphic and different from the reference base, subject to the following additional requirements: cross-species alignment depth of at least 20; a minimal mean Phred score of 20; mean depth of at least 5 at the 20 flanking bases; and contig coverage extending at least 70 bases upstream and downstream from the FD. We allowed for errors based on the depth of sequencing at the position of interest and on the expected number of individual chromosomes in the pooled samples (KOFLER *et al.* 2011). If depth (d) was lower than the expected number of chromosomes (c), the number of permissible errors for the position to be considered an FD was determined as ln(d), rounded down to the nearest whole number. If depth at the position was higher than the expected number of chromosomes, the number of permissible errors was ln(c)+(d/c), rounded down to the nearest whole number. For positions with more than two alleles, all minor alleles falling below this cut-off were considered sequencing errors if their frequencies were equal. If a majority of non-consensus basecalls supported one of the minor alleles, this position was considered polymorphic rather than FD and excluded from further analysis.

To identify FDs for the second round of genotyping, we first mapped *bip3-isoA* reads to the *D. bipectinata* reference genome (GenBank AFFE00000000.1) using SOAP2 with the same parameters as above, and examined positions

where the pileup consensus did not match the genome reference base. At these positions, we replaced the reference base with the new consensus base if it had a minimum depth of 5 and primary allele frequency above 0.5. This replacement was performed only for exonic positions, as determined by alignment to the annotated genome of *D. ananassae*. Finally, we mapped *mal0-sc2* reads to this modified reference genome and identified FDs as described above (File S4).

We used only FDs that had no other FDs within 70 bp on either side in order to avoid amplification biases during Sequenom genotyping. Since genotyping could also be hampered by intraspecific polymorphisms, we generated a reference sequence where IUPAC ambiguity codes were substituted at positions that were polymorphic in either one or both parental strains. Genotyping primers were designed based on this modified reference (Files S1C and S2C).

Marker design and genotyping

To identify the genomic locations of FDs, we BLASTed the sequence flanking each FD against the *D. ananassae* FlyBase 1.3 July 2011 reference genome and transcriptome. We assigned each FD to the chromosome arm of its BLAST match (SCHAEFFER *et al.* 2008), and used the *D. ananassae* transcriptome to make sure that target amplicons did not span splice junctions. Subsequently, marker flanking sequences were also BLASTed against the *D. bipectinata* genome. For the first round of genotyping, we selected 32 FDs that were evenly distributed among the major chromosome arms (Muller elements A-E) and were located at least 2.4Mb from each other in the *D. ananassae* genome (Table S3). Linkage mapping confirmed all homology-based chromosome assignments except for marker E-2, which was predicted to be located on Muller D (2R) but was linked to multiple markers on Muller E (2L). Because the initial analysis suggested the presence of one or more strong QTLs on Muller E or proximal Muller D, we performed a second round of genotyping with 32 additional FDs concentrated on these chromosome arms (Table S3). For these markers, we chose transcriptome contigs that were either located on different scaffolds in the modified *D. bipectinata* genome, or were separated by at least 0.75 Mb if located on the same scaffold.

Genotyping was performed by the UC Davis Veterinary Genetics Laboratory. Primers were designed using TYPER (Sequenom) based on the sequences of at least 70 bp upstream and 70 bp downstream from each candidate FD, after accounting for intraspecific polymorphisms with IUPAC codes as described above (Table S4). All primers were ordered from IDT. Individual flies were genotyped using MASSARRAY (Sequenom) single base extension in a 32-plex format using standard protocols. Genotype calls were made using default peak intensity thresholds (Table S5).

Construction of linkage maps

Genetic maps were constructed based on two separate F_2 backcrosses and a long-term introgression cross (see Results). Markers were assembled into linkage groups using R-QTL (BROMAN and SEN 2009; BROMAN *et al.* 2003); markers that deviated strongly from the expected Mendelian ratios were excluded. When marker density is low, recombination

frequency will underestimate the actual genetic distance due to the inability to detect double crossovers. We therefore used the standard Haldane (HALDANE 1919) and Kosambi (KOSAMBI 1944) mapping functions to estimate genetic distances between markers; the two formulas produced nearly identical results. We observed very high amounts of recombination on Muller D (chromosome arm 2R). This is probably caused by the interchromosomal effect (STEVISON *et al.* 2011), since all chromosome arms except 2R carry large inversions. To place minimal bounds on the large genetic distances on 2R, we used a simplified Haldane map function that does not include crossover interference or correction for multiple exchanges between adjacent markers.

Within each linkage group, we used the R-QTL "ripple" function to determine the marker order that minimized the number of crossovers. Due to the presence of chromosomal inversions, several potential marker orders had similar likelihoods. In such cases, we tested all possible orders on the chromosome arm and chose the one with the lowest total map length. On Muller A (X chromosome), where very little crossing-over was detected, we selected marker order at random from a set of possible orderings with the highest likelihood. Some pairs of markers showed no recombination and had identical map positions, suggesting they were located inside the same chromosomal inversion. We assigned such markers the same relative order in which their homologs appear in the *D. ananassae* genome. Finally, markers located on the same scaffold in the *D. bipectinata* genome assembly were placed adjacent to each other in the order in which they appear in that scaffold.

Linkage maps were constructed independently for each of the two backcrosses. To create a linkage map for the introgression panel, we merged the two backcross maps as follows. First, we determined the consensus marker order compatible with data from both backcrosses. Second, when data for a pair of adjacent markers was available in both datasets, we averaged the genetic distances between these markers from the two backcrosses. Finally, we used this low-resolution map to anchor additional, more densely spaced markers that were only genotyped in the introgression panel.

Genotyping on the dot chromosome (Muller F)

We used the FD alignments mapped to the *D. bipectinata* genome to identify FDs that were mapped by BLASTn to the fourth chromosome (Muller F) in both *D. melanogaster* and *D. ananassae*. Cleaved amplified polymorphism sequences (CAPSs) were selected by feeding the 500 bp upstream and downstream of the FD into SNP Cutter (ZHANG *et al.* 2005). Two potential CAPS sites were chosen, and gradient PCR was run to determine the optimal conditons. The sites were amplified and digested in single-plex format using restriction enzyme identified by SNP Cutter (PstI for markers F-1 and F-2, with PstI cutting the *D. bipectinata* but not the *D. malerkotliana* allele). Digested amplicons were examined on a 1.8% Agarose gel. CAPS genotyping was performed on eight of the lightest and eight of the darkest individuals in each of the two backcrosses. DNA extracted from pooled *bip3-isoA* flies, pooled *mal0-sc2* flies, and pooled *bip3-isoA* / *mal0-sc2* F_1

flies was used for controls. For every marker, post-restriction amplicon sizes matched the predicted values for each genotype.

QTL mapping

We applied the Haley-Knott, multiple imputation, and expectation-maximization models (DEMPSTER *et al.* 1977; HALEY and KNOTT 1992; SEN and CHURCHILL 2001) to our data using the R-QTL package (BROMAN and SEN 2009; BROMAN *et al.* 2003). All three methods gave nearly identical peak locations, LOD scores, and significance levels, indicating that the data are robust to over-parameterization. We first performed single-QTL scans to identify likely regions of genotype-phenotype association. For each detected QTL, we performed composite interval mapping and determined that genotypes at neighboring markers did not significantly affect the peak LOD score or width. To calculate the statistical significance of QTL peaks, we used a genome scan-adjusted P-value corresponding to an observed LOD score. The null distribution was derived through standard permutation test. The P-value represents the chance, under the null hypothesis of no QTL, of obtaining an LOD score that large or larger somewhere in the genome.

In the introgression cross, but not in either backcross, we detect a possible small-effect association between marker D-6 (most distal 2R) and sex combs. While the LOD score on the distal 2R is significant in both the single-QTL genome scan and the two-QTL genome scan that takes both the distal 2L and distal 2R peaks into account, the genotype ratios at marker D-6 diverge significantly from expectation ($P < 0.001$). We therefore discarded this marker. Marker D-4 (distal 2R) maintained the *D. bipectinata* allele through the introgression and is present in the introgression cross but is not associated with the phenotype. In both backcrosses, but not in the introgression cross, a two-QTL genome scan accounting for the 2L QTL reveals another significant QTL (P-value < 0.001) on proximal Muller C (3L), and to a lesser extent Muller B (3R). The relationship between the 2L and proximal 3L/3R QTL is additive. It appears that the *D. bipectinata* genes responsible for this weak QTL were lost from the introgression strain.

Power analysis was performed using R/qtlDesign (SEN *et al.* 2007) and confirmed using calculations set forth in (LYNCH and WALSH 1998). This analysis leverages the theories described in (DARVASI *et al.* 1993; HALEY and KNOTT 1992; REBAI *et al.* 1995; SIMPSON 1989; SIMPSON 1992). The `detectable()` function in R/qtlDesign was used to determine the minimum effect size of the QTL compared to the Muller E QTL. Environmental variance was estimated from the variance within each parental strain: $\sigma^2 = 1.967$ for *D. malerkotliana* homozygotes and $\sigma^2 = 2.857$ for *D. bipectinata* homozygotes. Genetic variance was calculated by taking the average of the means of the heterozygote and the homozygote, squaring it, and dividing by 4 (as mandated by the backcross model). We assumed a complete linkage between the QTL and a genotyping marker (recombination fraction of 0), and treated all QTLs as additive. The selection fraction was 1 for Muller A-E; for Muller F, it was set to 16/188 for the *D. malerkotliana* backcross and 16/163 for the *D. bipectinata* backcross. For QTLs located on Muller A-E, the minimal detectable effect size is nearly identical for a power of 0.95 and a power of 1. For

Muller F, selective genotyping of extreme individuals (DARVASI and SOLLER 1992) meant that requiring a power of 1 rather than 0.95 would dramatically reduce the probability of detecting QTLs of small effect. In the interest of consistency, we used the power of 0.95 for all chromosome arms.

Mapping candidate genes to the *D. bipectinata* genome

To determine the locations of *Scr* and *dsx* on our linkage maps, we BLASTed the full-length sequences of the *D. ananassae* genomic regions encompassing each gene, as well as the mature transcript of each gene, against the modified *D. bipectinata* genome assembly. Both the genomic and the transcript sequences mapped unambiguously to genome scaffolds that contained several of our genotyping markers. *Scr* mapped to scaffold scf7180000396708 (http://www.ncbi.nlm.nih.gov/nuccore/358402995), which also contained markers E-In(2L)D-u7 through E-In(2L)D-u13. This scaffold is built of 31 contigs that are stitched together with 30 short stretches (mean = 58.66 bp, mode = 20, min = 20, max = 523) of unknown bases (Ns). *Scr* almost wholly resides in a single contig, ctg7180000390941, which also includes marker E-In(2L)D-u10. *dsx* mapped to scaffold scf7180000395971 (http://www.ncbi.nlm.nih.gov/nuccore/358403364), which also contained marker E-In(2L)D-u16. This scaffold is composed of three contigs with a mean separation of only 20 bp. *dsx* and marker E-In(2L)D-u16 both lie on the same contig, ctg7180000389680. On the linkage map, both scf7180000396708 and scf7180000395971 are located in the distal-most, non-recombining segment of Muller E (2L) corresponding to the inversion in In(2L)D. A similar BLAST analysis shows that genotyping markers linked to the major Muller E QTL are located in a different, more proximal non-recombining region corresponding to the inversion In(2L)M (Table S6).

REFERENCES

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

BIROL, I., S. D. JACKMAN, C. B. NIELSEN, J. Q. QIAN, R. VARHOL *et al.*, 2009 De novo transcriptome assembly with ABySS. Bioinformatics **25:** 2872-2877.

BROMAN, K. W., and Ś. SEN, 2009 *A Guide to QTL Mapping with R/qtl.* Springer, New York.

BROMAN, K. W., H. WU, Ś. SEN and G. A. CHURCHILL, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889-890.

COCK, P. J. A., C. J. FIELDS, N. GOTO, M. L. HEUER and P. M. RICE, 2010 The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research **38:** 1767-1771.

DARVASI, A., and M. SOLLER, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. TAG Theoretical and Applied Genetics **85:** 353-359.

DARVASI, A., A. WEINREB, V. MINKE, J. I. WELLER and M. SOLLER, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943-951.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) **39:** 1-38.

EWING, B., and P. GREEN, 1998 Base-Calling of Automated Sequencer Traces UsingPhred.II. Error Probabilities. Genome Research **8:** 186-194.

EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-Calling of Automated Sequencer Traces UsingPhred.I. Accuracy Assessment. Genome Research **8:** 175-185.

HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. Journal of Genetics **8:** 299-309.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315-324.

HUANG, X., and A. MADAN, 1999 CAP3: A DNA Sequence Assembly Program. Genome Research **9:** 868-877.

KENT, W. J., 2002 BLAT—The BLAST-Like Alignment Tool. Genome Research **12:** 656-664.

KOFLER, R., P. OROZCO-TERWENGEL, N. DE MAIO, R. V. PANDEY, V. NOLTE *et al.*, 2011 PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. PLoS ONE **6:** e15925.

KOSAMBI, D. D., 1944 The estimation of map distance from recombination values. Ann. Eugen. **12:** 172–175.

LI, R., C. YU, Y. LI, T.-W. LAM, S.-M. YIU *et al.*, 2009 SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25:** 1966-1967.

LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantatative Traits*. Sinauer Associates, Inc., Sunderland, Massachusetts.

MILLER, J. R., S. KOREN and G. SUTTON, 2010 Assembly algorithms for next-generation sequencing data. Genomics **95:** 315-327.

REBAI, A., B. GOFFINET and B. MANGIN, 1995 Comparing Power of Different Methods for QTL Detection. Biometrics **51:** 87-99.

ROBERTSON, G., J. SCHEIN, R. CHIU, R. CORBETT, M. FIELD *et al.*, 2010 De novo assembly and analysis of RNA-seq data. Nat Meth **7:** 909-912.

SCHAEFFER, S. W., A. BHUTKAR, B. F. MCALLISTER, M. MATSUDA, L. M. MATZKIN *et al.*, 2008 Polytene Chromosomal Maps of 11 *Drosophila* Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics **179:** 1601-1655.

SEN, S., and G. A. CHURCHILL, 2001 A Statistical Framework for Quantitative Trait Mapping. Genetics **159:** 371-387.

SEN, S., J. M. SATAGOPAN, K. W. BROMAN and G. A. CHURCHILL, 2007 R/qtlDesign: inbred line cross experimental design. Mamm Genome **18:** 87-93.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. JONES *et al.*, 2009 ABySS: a parallel assembler for short read sequence data. Genome Res **19:** 1117-1123.

SIMPSON, S. P., 1989 Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. TAG Theoretical and Applied Genetics **77:** 815-819.

SIMPSON, S. P., 1992 Correction: Detection of linkage between quantitative trait loci and restriction fragment length polymorphism using inbred lines. TAG Theoretical and Applied Genetics **85:** 110-111.

STEVISON, L. S., K. B. HOEHN and M. A. NOOR, 2011 Effects of inversions on within- and between-species recombination and divergence. Genome Biol Evol **3:** 830-841.

ZHANG, R., Z. ZHU, H. ZHU, T. NGUYEN, F. YAO *et al.*, 2005 SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design. Nucleic Acids Res **33:** W489-492.