Supplementary Materials:

Evidence for the fixation of gene duplications by positive selection in *Drosophila*

Margarida Cardoso-Moreira[1,2], J. Roman Arguello[1,2], Srikanth Gottipati[1,3], L. G. Harshman[4], Jennifer K. Grenier[1] and Andrew G. Clark[1]

1. Additional methods on the identification of CNVs segregating in the GDL

1.1. Pindel, Delly and the In-house pipeline

Pindel detects CNVs using the information contained in the reads that span the breakpoints of CNVs (split-read detection). We ran Pindel on genome alignments generated by Novoalign (v2.07.11, www.novocraft.com) against the release 5 of the *D. melanogaster* genome. We ran Novoalign with default parameters with the exception of the option x6, which lowers the penalty for gap extension. We ran Pindel (version 024q and pindel2vcf 033) on all genomes simultaneously (each chromosome arm ran separately). We used the following Pindel parameters: -n 25 and -x 7. We accepted all calls supported by at least 3 reads that were larger than 25 bp and that when complex (*i.e.* accompanied by additional nucleotides inserted or deleted at the breakpoints) had a stretch of nucleotides inserted/deleted equal or smaller to half the size of the CNV.

Delly detects CNVs using the information from pairs of reads that map discordantly to the reference genome (paired-end detection). We ran Delly (version 0.0.7) on each line independently, using the same genomic alignments used for Pindel. We ran Delly using the following parameters: -p –q 20. We used the set of calls produced by Delly that have paired-end support.

Our in-house pipeline also uses a split-read approach to detect CNVs (Cardoso-Moreira *et al.* 2012). Unlike the two other pipelines, our in-house pipeline starts from the set of reads that were not mapped to the reference genome (release 5) using Mosaik (Lee *et al.* 2014). We ran Mosaik (version 1.0) using a jump library and the following parameters: -hs 15 –mm 15 –mhp 100 –act 35 and –bw 35. For each line, we took the set of unmapped reads and re-aligned them to the same reference genome using BLAT (blat-3.4, oneOff=1 (Kent 2002)), bwa-sw (bwa-0.5.7 (Li and Durbin 2010)) and SSAHA2 (ssaha2, -output sam_soft (Ning *et al.* 2001). We then used custom perl scripts (modified from Cardoso-Moreira *et al.* 2012) to identify CNV breakpoints. Around 98%

of CNVs identified using this approach derived from the re-alignment of the unmapped reads with BLAT.

Our CNV dataset is comprised of all CNV calls made by at least two of the three CNV pipelines described above. Because Delly only provides approximate CNV breakpoints, we used the precise breakpoint coordinates generated by Pindel or by our in-house pipeline. When both Pindel and the in-house pipeline predicted the same variant we used the information provided by Pindel because in addition to the CNV coordinates it also detects the presence of microhomology and inserted/deleted nucleotides at the breakpoints. The overlap between the three sets of calls was done using intersectBed from the BEDTools suite (Quinlan and Hall 2010) and is described in Supp. Fig. 2. We considered that a given CNV was supported by at least two pipelines when the extent of the overlap between the two pipelines was at least 70% (intersectBed –f 0.7 –r) if the calls were supported by the two split-read pipelines, and 50% (intersectBed –f 0.5 –r) if the calls were supported by one split-read pipeline and Delly.

The set of CNVs was then filtered to only include variants between 25 bp and 25 kb (the number of variants larger than 25 kb is low: 103 deletions and 36 duplications) and to exclude variants with breakpoints associated with transposable elements or other classes of repeats (Supp. Fig. 2). The latter filter was imposed by removing CNV breakpoints overlapping transposable elements and repeats identified by Flybase (release 5.52, dos Santos *et al.* 2015) or overlapping repeats identified by running RepeatMasker (Smit *et al.* 2015) on 100 bp upstream and downstream the CNV breakpoints. We also excluded CNVs where at least 50% of the sequence matched a transposable element.

1.2 Determining the coverage support of CNV calls

In addition to split-read and paired-end methods, the depth of read coverage of a given region can also be used to infer the presence of CNVs. The sensitivity and accuracy of this method depends however on the average depth of coverage across the genome, which in our case is not sufficient to make this method accurate (12.5x), especially within heterozygous

blocks. Although we cannot use depth of coverage to identify CNVs, we can use it to distinguish retrogenes from intron deletions and to distinguish gene fusions from gene conversion events that mimic the signal of deletion. This information can also be useful to strengthen the confidence in a given CNV call though we did not use the depth of read coverage to further filter the CNV dataset. For duplications and deletions we determined whether there is a higher or lower read coverage in that region than expected. Using BEDTools (Quinlan and Hall 2010) we determined the number of reads mapping to three positions within each CNV: the middle position, halfway between the start of the CNV and the middle and halfway between the end of the CNV and the middle. If for the lines without the CNV the median coverage of at least 2 of the 3 positions was between 5 and 50 reads we accepted we had enough information to make a coverage call. We classified a deletion as having coverage support when the median coverage was smaller or equal to 2 reads in at least 2 of the 3 positions. If a given deletion was located within an inversion, and therefore potentially heterozygous, we lowered the threshold so that a deletion was classified as having coverage support if the median coverage was smaller or equal to 8 reads and lower than the median coverage for the lines without the deletion. We classified a duplication as having coverage support if the median coverage was at least 1.5X higher than the median coverage for the lines without the duplication. Again, we lowered the threshold for duplications located within inversions by requiring the median coverage of the lines carrying the duplication to be at least 1.3X higher than the median coverage of the lines without the duplication. Although informative (the false positive rate of CNVs with coverage support is lower at 8%), most CNVs without coverage support tested by PCR were confirmed (8/10 duplications and 13/18 deletions) suggesting that using coverage in our dataset is of limited value.

1.3 Polarization of the CNV calls

Our CNV detection pipelines identify CNVs by comparison with the reference genome, which means that a small fraction of our calls are expected to correspond to novel variants carried by the reference genome. These CNV calls correspond to the ancestral state and are not

new polymorphisms (*i.e.* the derived state). Our pipelines are biased against detecting variants segregating in the reference genome. They can only identify small insertions (within the length of a read), so all medium to large deletions specific to the reference genome will not be identified in our dataset. Similarly, tandem duplications that have not yet accumulated nucleotide differences are often collapsed in the reference assembly. Regardless of these caveats, it is important to quantify the fraction of CNV calls that correspond to the ancestral state. We started polarizing our CNVs using syntenic alignments between *D. melanogaster* and *D. simulans*. We downloaded the pairwise alignments in axt format from http://hgdownload-test.sdsc.edu/goldenPath/droSim1/vsDm3/axtNet/, and converted them into multiple alignment format (MAF) using the "axtToMaf" utility retrieved from http://genomewiki.ucsc.edu/index.php/The_source_tree. We extracted the sub-alignments for each CNV region (with an additional 100 bp 5' and 3' to the start and end coordinates) using the "maf_parse" utility, and within each CNV's sub-alignment generated summaries of the available length of sequence for the two species including the number of gaps. For alignments that spanned the CNV region, we labeled any instances where *D. melanogaster* gaps amounted to ≥ 70% the size of the CNV as *D. melanogaster* deletions. Similarly, we labeled any alignments with gaps in *D. simulans* ≥ 50% of the CNV as *D. melanogaster* insertions (alignments that did not span the CNV region were counted as uninformative). If no gaps were identified in the alignment it means that the polymorphic deletion call in our CNV dataset corresponds to the derived state (and the opposite reasoning applies to insertions and duplications). Using this strategy only ~1% of the calls were classified as corresponding to the ancestral state, but 14% of the calls could not be polarized. Since our work focuses on new genes, for both new genes created by duplications and deletions we performed an additional check. We blasted our set of new genes against the reference genomes of *D. melanogaster* (r.5), *D. simulans* (r.1), *D. sechellia* (r.1) and *D. yakuba* (r.1) using standalone ncbi-blast-2.2.25+ (Camacho *et al.* 2009). We then determined if there were alignments in addition to the expected self-hit/orthologous sequence with a sequence identity higher than 80% and matching at least 80% of the blasted sequence. Using this approach

we determined that all new genes corresponded to the derived state, with the exception of 3 gene duplications and 6 gene deletions that could not be polarized.

2. Additional methods on the identification of new genes

2.1. Or22a:22b fusion

When CNVs have complex breakpoints their detection is much harder (*i.e.* in addition to the CNV there are additional nucleotides inserted or deleted at the breakpoint). This is particularly the case with short read technology. In our set of CNV calls, the fusion between the sensory genes *Or22a* and *Or22b* appears as segregating in 20 of the 84 lines, which includes only 1 Zimbabwe line. However, using read coverage (*i.e.* requiring no reads covering this region) we were able to detect this fusion in 13 additional lines, including 11 Zimbabwe lines. It is not clear why we have such a high false negative rate for this deletion (and why it is mostly associated with the Zimbabwe lines). However, it could be because it is associated with a 25 bp insertion that makes its detection more difficult. Overall, this fusion is nearly fixed in the Zimbabwe population (92% frequency) and is present at appreciable, though lower, frequencies in the other populations (61% in Tasmania, 32% in Ithaca, 16% in the Netherlands and 7% in Beijing).

2.2. Identification of retrogenes

We identified retrogenes using the set of CNV calls before filtering for transposable elements and other repeats. We identified all deletions where at least 90% of its sequence matched at least 90% of an intronic sequence (intersectBed -f 0.9 –r). Using this approach we identified 68 genes with at least one intron carrying the signal of being deleted. Of the 68, 42 genes had one single intron deletion that matched the sequence of a transposable element. These events do not correspond to either retroposition or intron loss events but instead to the insertion/deletion of mobile elements. For the remaining 26 genes we distinguished between intron loss and retroposition by determining the read coverage within the “deleted” intron. If there

is truly intron loss there should be no read coverage, but if there is a retroposition event the coverage should be close to the genome average. 9 of the 26 genes are carrying intron deletions, whereas the remaining 17 have polymorphic retrogenes (Supp. Table 4). For this set of 17 genes we went back to the set of calls made by the three CNV detection pipelines and allowed for additional intron deletions within these genes supported by only one of the pipelines (Supp. Table 4). In order to map the insertion sites of the retrogenes we ran another pipeline designed to identify structural variants called Hydra (Quinlan *et al.* 2010). We ran Hydra (version 0.5.3) on the set of Mosaik alignments using default parameters. Hydra identified the insertion of the *CG33969*-retrogene within the intron of the gene *CBP*, the insertion of the *eIF-4E*-retrogene within rRNA sequences and confirmed the co-retroposition of the genes *Cf2* and *Pen*. Using the coordinates suggested by Hydra, we designed primers to confirm the insertion of the *CG33969*-retrogene within the intron of *CBP* and sequenced the locus with Sanger sequencing (Supp. Fig. 5). We investigated the expression profile of the parental genes of retrogenes using the modEncode (Brown *et al.* 2014) and FlyAtlas (Chintapalli *et al.* 2007) datasets accessed from within Flybase ("FlyAtlas Anatomy Microarray" and "modENCODE Anatomy RNA-Seq" tracks, r.5.52). We also investigated the expression profile of the parental genes throughout sperm development using the SpPress database (Vibranovski *et al.* 2009).

3. Additional methods on the gene expression analyses

We applied standard normalization routines from the R package 'limma' (Smyth and Speed 2003, Ritchie *et al.* 2015). The expression value ratios were first subjected to within array normalization with print-tip loess and then to between array normalization using quantile normalization. After these steps, a dye bias correction was performed. We applied two filters to the expression data. The first filter removed all genes identified as being either lowly expressed or not expressed at all. We identified the 25th quantile of intensity values across all lines to be ~ 8 and removed all genes where at least one line had an intensity value lower than 8. We also explored using higher cutoffs (9 and 10) but our results remained qualitatively the same. The

second filter was aimed at preventing cross-hybridization effects. We blasted all probes in the arrays against the *D. melanogaster* CDS fasta sequence (release 5.33) using standalone blastn with default values (ncbi-blast-2.2.25+ (Camacho *et al.* 2009)). We removed all probes with a second alignment (*i.e.* non-self) in a different gene when the alignment was larger than 80% of the probe size with a sequence identity of more than 85%. These corresponded to 224/14735 probes in the array. After the intensity and cross hybridization filters were applied we were left with 10,997/14,735 probes.

4. Additional notes on detecting signals of positive selection

In addition to the LD analyses described in the Results and Methods sections, we also compared the LD distribution between the whole set of duplications and those of high-frequency duplications using only SNPs that are at least 700 bp away from each other. In all populations, LD between SNPs located at this distance is very low, so by limiting our analysis to this set of SNPs we can avoid the potential confounding effects of seeing higher LD associated with some high-frequency duplications simply because there are more SNPs segregating in this region. All of our results (*i.e.* the identity of the high-frequency duplications associated with higher LD) remained the same when using this second approach.

5. A portrait of copy number variation in five world populations of *Drosophila*

In agreement with previous work, we find that purifying selection is pervasive across the CNV dataset (*e.g.* Emerson *et al.* 2008, Zichner *et al.* 2013). Simulations show that CNVs are strongly depleted among coding regions. For example, 4% of deletions overlap coding exons when 24% would be expected in the absence of purifying selection (Fisher's exact test, $p < 2.2 \times 10^{-16}$, Supp. Table 3). This depletion of CNVs in coding regions is significantly stronger for deletions/insertions than for duplications ($p < 2.2 \times 10^{-16}$, Supp. Table 3). This is to be expected because the deletion of a coding region will almost always be deleterious, whereas a partial duplication can often be neutral. Although we only identified variants that are at least 25 bp in

length, we still detect a clear excess of CNVs *within* coding exons that are multiples of 3 bp, and that therefore are less likely to lead to frame-shifts or premature stop codons (Supp. Fig. 21). Deletions and insertions are also significantly depleted in UTR exons and the depletion is significantly stronger for 5'UTR exons than 3'UTR exons ($p < 2.2 \times 10^{-16}$, Supp. Table 3). In contrast, we found an excess of duplications overlapping 3'UTR exons (p=0.001). This observation is in agreement with a previous study (Emerson *et al.* 2008) and merits additional future work. Most notably, we observed that, although partial gene duplications are significantly depleted in our dataset ($p = 3 \times 10^{-12}$), there is a clear excess of complete gene duplications (Supp. Table 3). Given the size of the duplications in our dataset and their chromosomal locations, we would expect that ~ 5% would encompass complete genes; instead 14% of the duplications create new complete gene duplications ($p < 2.2 \times 10^{-16}$, Supp. Table 3).

There are notable differences between the five populations in the numbers of CNVs detected. The Zimbabwe population is segregating for a significantly higher number of CNVs than the remaining 4 populations (Supp. Fig. 22). This result is consistent with the demographic history of these populations. *D. melanogaster* is originally from sub-Saharan Africa and so Zimbabwe, by being closest to the center of origin of the species, is expected to show the highest genetic diversity. Higher levels of diversity for the Zimbabwe population were also observed for the set of SNPs and indels segregating in these populations (Grenier *et al.* 2015). In addition to the difference in CNV abundance between the populations, we also identified a difference in the number of CNVs present on the X chromosome vs. the autosomes. All non-African populations have a significantly lower density of CNVs on the X when compared to the autosomes (Supp. Fig. 23), with the Zimbabwe population showing lower CNV density on both the X and chromosome 2R. This observation is in line with previous studies in *Drosophila,* which have identified lower densities of CNVs on the X (*e.g.* Emerson *et al.* 2008, Zichner *et al.* 2013). In previous studies, this difference was attributed to a higher efficiency of selection on the X, a consequence of the immediate expression of X-linked recessive mutations in males that only carry one copy of this chromosome. There are, however, demographic reasons to expect the same result. The non-

African populations underwent a bottleneck when they moved out of Africa, and the recovery of diversity levels after a bottleneck occurs more slowly for the X than the autosomes (Pool and Nielsen 2007). These two explanations are not mutually exclusive but lead to different predictions in terms of the variants that would be most affected. Demography affects equally all variants, whereas selection acts mostly on variants that affect fitness, the majority of which are mutations affecting coding regions. In agreement with a role for selection (that does not preclude additional demographic effects), we observe that the fraction of deletions and duplications that affect coding sequences is significantly lower on the X than on autosomes for most populations.

6. Notes on the set of polymorphic retrogenes

In order for retrogenes to be heritable, the retroposition has to occur in the germline, which means that the probability that a given gene generates a duplicate retrogene depends on its expression level in the germline (Kaessmann *et al.* 2009). Consistent with this expectation, we observed that the parental genes of retrogenes are all highly expressed in the germline based on modEncode expression data (Brown *et al.* 2014) (Supp. Table 4). Surprisingly however, this analysis shows that parental genes are more highly expressed in ovary than testis, with some parental genes being highly transcribed only in ovary (Supp. Table 4). If taken at face value, these data could suggest that the female germline contributes disproportionately to the creation of retrogenes compared to the male germline. Alternatively, the apparent difference in expression between the two germline tissues could derive from differences in the cellular composition of these tissues. When the adult whole testis transcriptome is profiled, most of the expression signal derives from the most common cell type, which are meiotic cells, whereas most retrogenes are likely to be created earlier during sperm development. Using the SpPress database (Vibranovski *et al.* 2009) we investigated the expression levels of parental genes in mitotic, meiotic and post-meiotic testicular cells. We found that parental genes are significantly more highly expressed during sperm development than those genes that do not create retrogenes, and that parental

genes are most highly expressed in mitotic cells (Supp. Fig. 24). Although these data are not directly comparable with the modEncode data for ovary, the most parsimonious explanation for the difference in expression of parental genes of retrogenes in ovary vs. testis is the under-representation of mitotic cells in adult whole testis.

7. Supplementary References

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, *et al.* 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**:393-399.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.

Cardoso-Moreira M, Long M. 2012. The origin and evolution of new genes. *Methods Mol Biol.* **856**:161-186.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet* **39**:715-720.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM; FlyBase Consortium. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* **43**:D690-697.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320**:1629-1631.

Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG. 2015. Global diversity lines - a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3* **5**:593-603.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**:19-31.

Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* **12**:656-664.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* **9**:e90581.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. **26**:589-595.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**:1725-1729.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution* **6**:3001-3006.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**:623-635.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842.

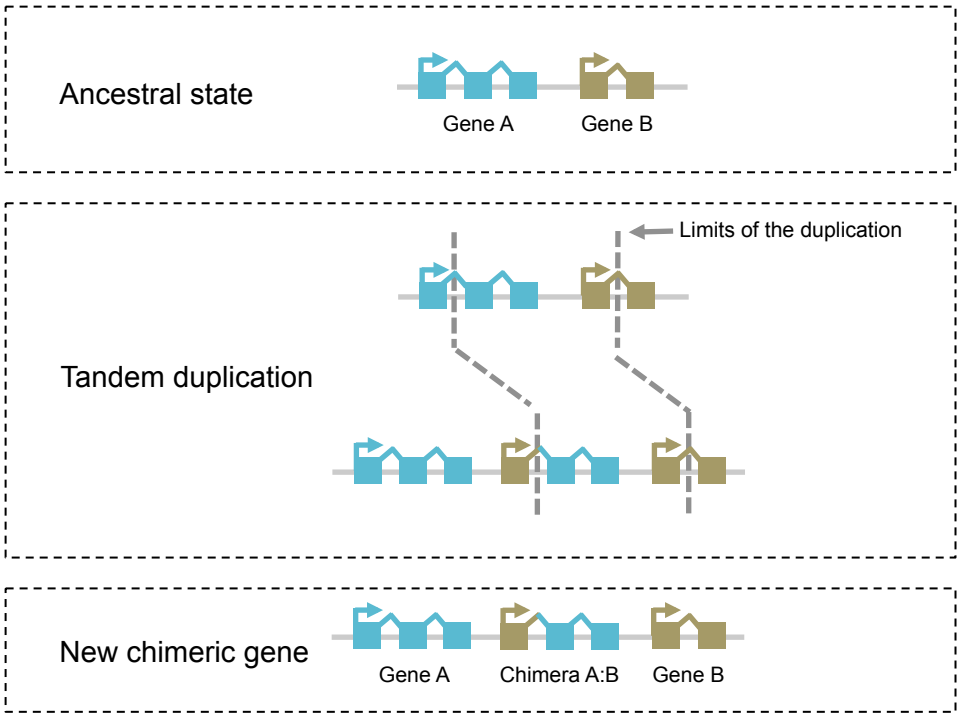
Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W,Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**:e47.

Smit, AFA, Hubley, R, Green, P. 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org

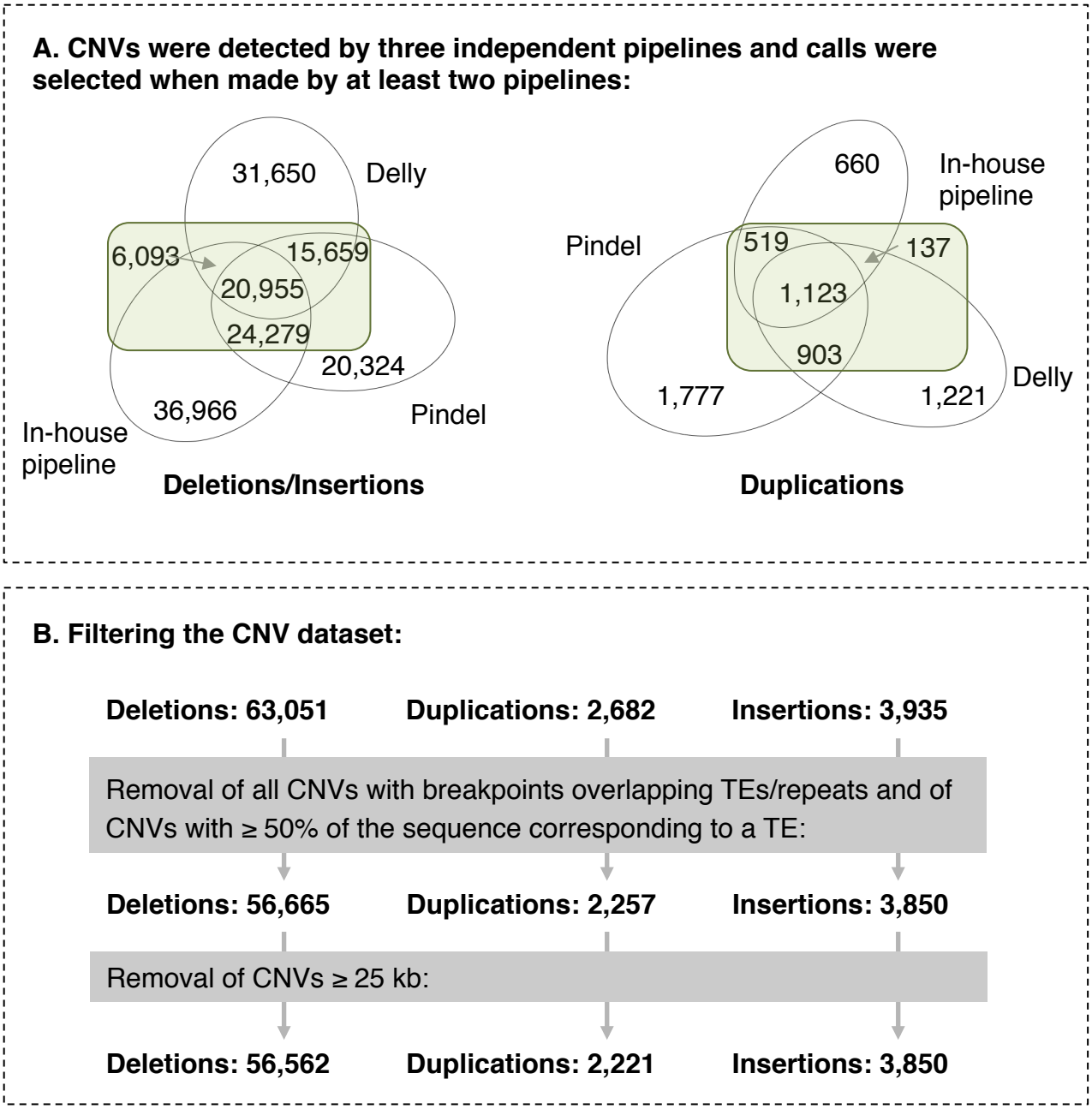
Smyth GK, Speed T. 2003. Normalization of cDNA microarray data. *Methods* **31**:265-273.

Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet* **5**:e1000731.

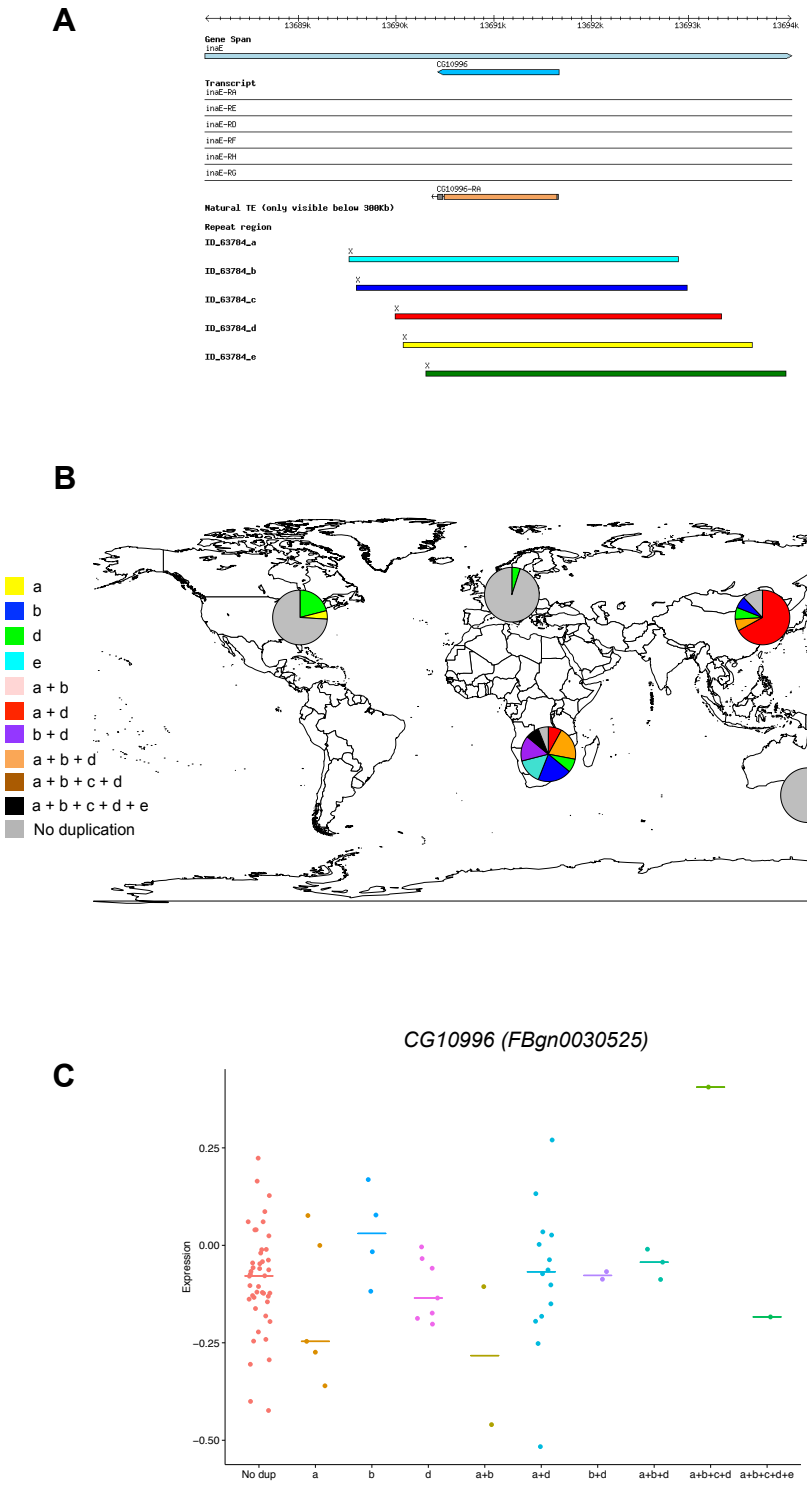
Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EE, Korbel JO. 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* **23**:568-579.



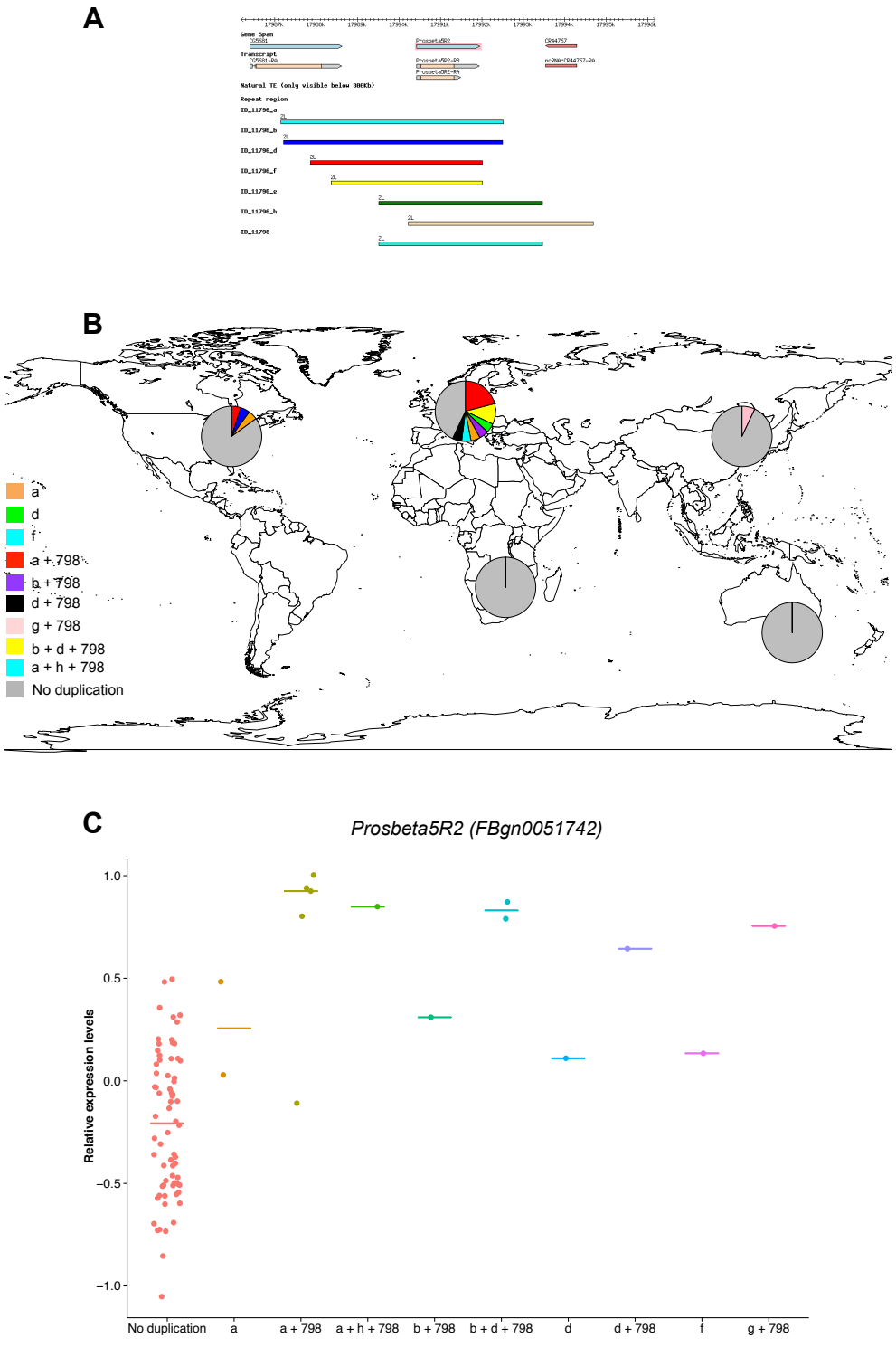
Supp. Figure 1. Schematic depicting how duplications can create new chimeric genes



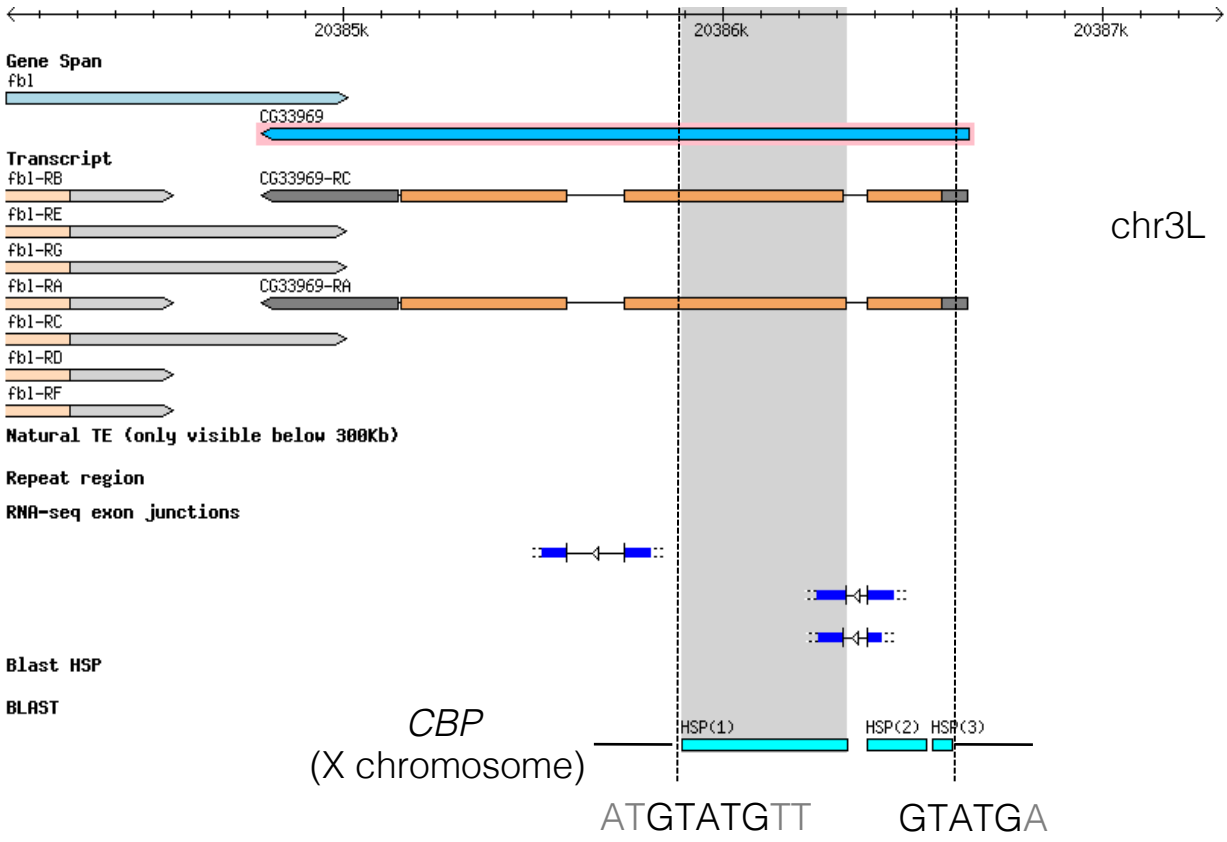
Supp. Figure 2. Description of the pipeline used to identify the set of CNVs segregating in the GDL. (A) Venn diagrams of the number of deletions/insertions and duplications identified by the three CNV detection pipelines. We selected all calls supported by at least two pipelines (calls in the green box). (B) Description of the filters used on the set of calls produced following (A) and the numbers of CNVs remaining after they were applied.



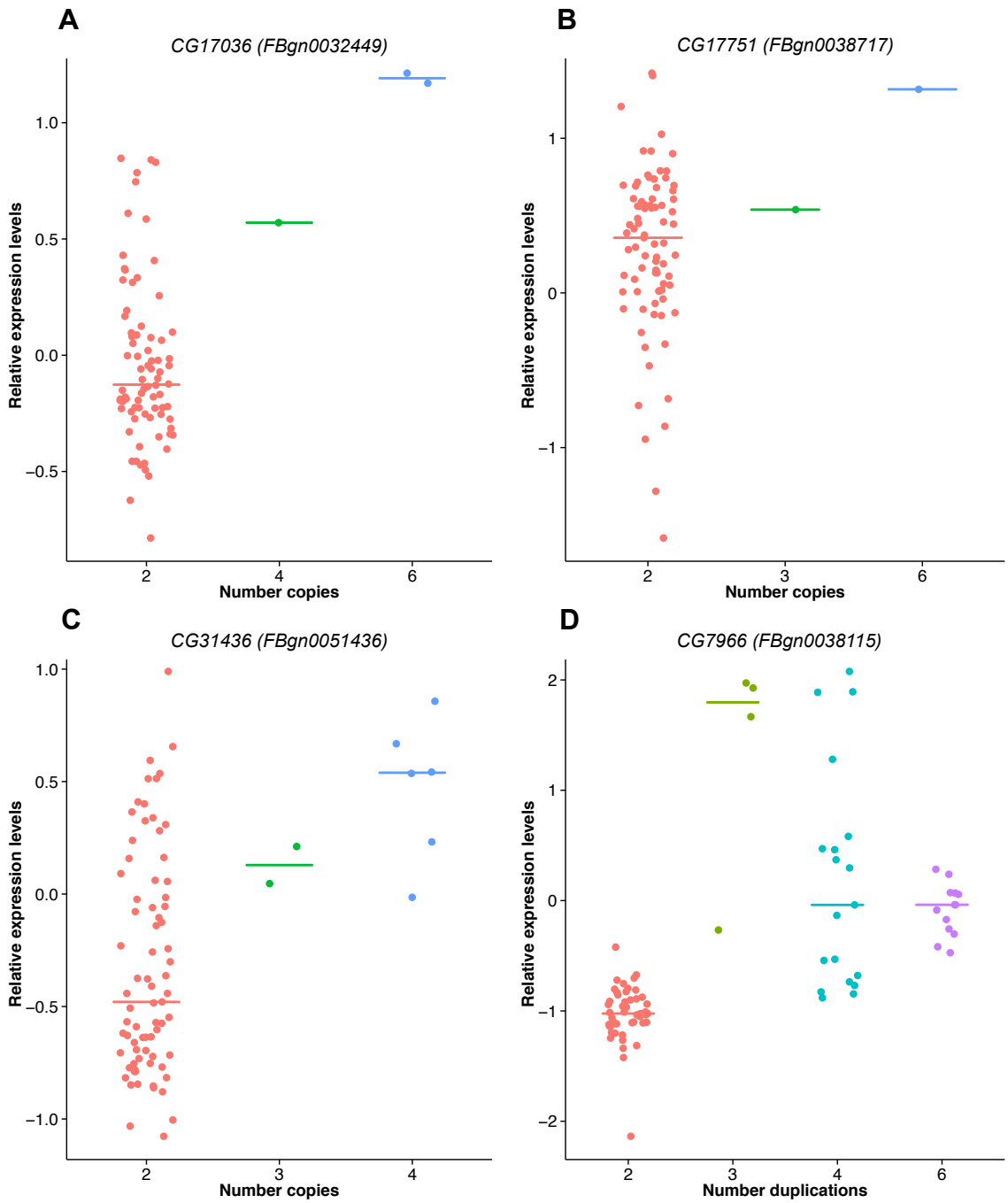
Supp. Figure 3. Characterization of the *CG10996* duplications. (A) Genome Browser (Flybase) view of the 5 duplications of *CG10996*. (B) Distribution of the gene duplications in the five populations. (C) The duplication of *CG10996* is not associated with significant changes in the expression levels of this gene.



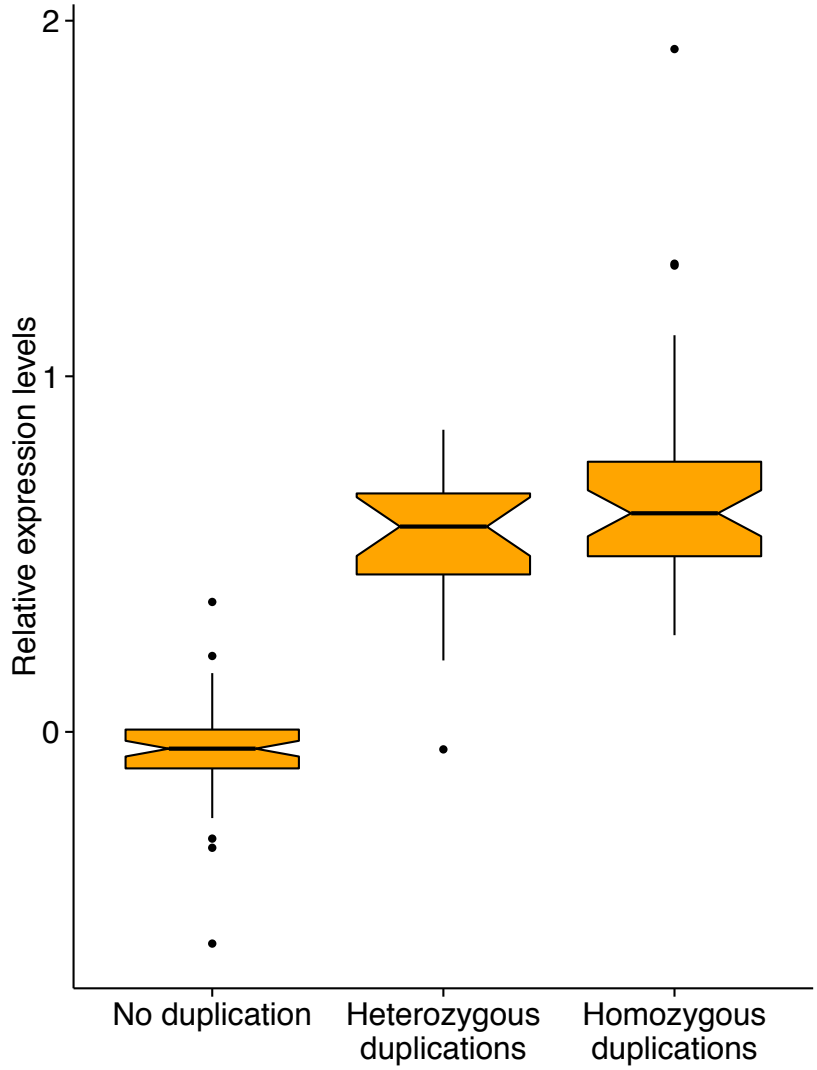
Supp. Figure 4. Characterization of the *Prosbeta5R2* duplications. (A) Genome Browser (Flybase) view of the 7 independent gene duplications. (B) Distribution of the gene duplications in the five populations. (C) Expression levels of *Prosbeta5R2* in individuals carrying different duplication genotypes.



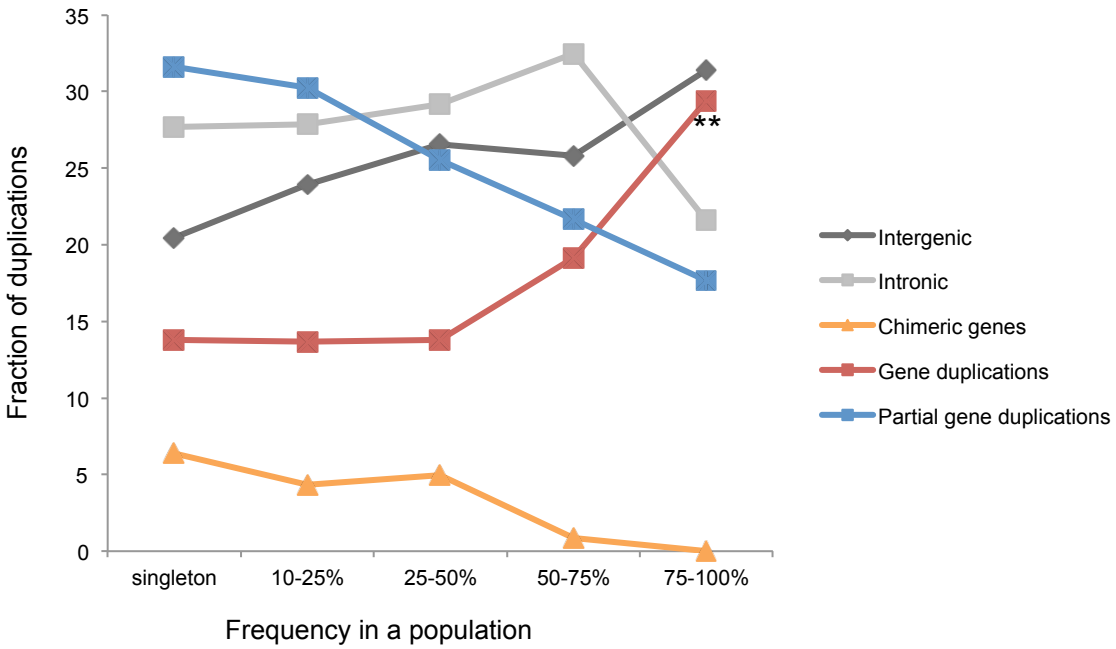
Supp. Figure 5. Retroposition of *CG33969*. The retroduplication of *CG33969* (on chromosome 3L) is partial, encompassing only the first exon and part of the second exon (region between the vertical dotted lines). This retrocopy was inserted into an intron of the gene *CBP* located on the X chromosome. In addition to the absence of the intron (the hallmark of retroposition), Sanger sequencing revealed a small deletion in the first exon of the retrocopy. Immediately flanking the retrocopy are direct repeats.



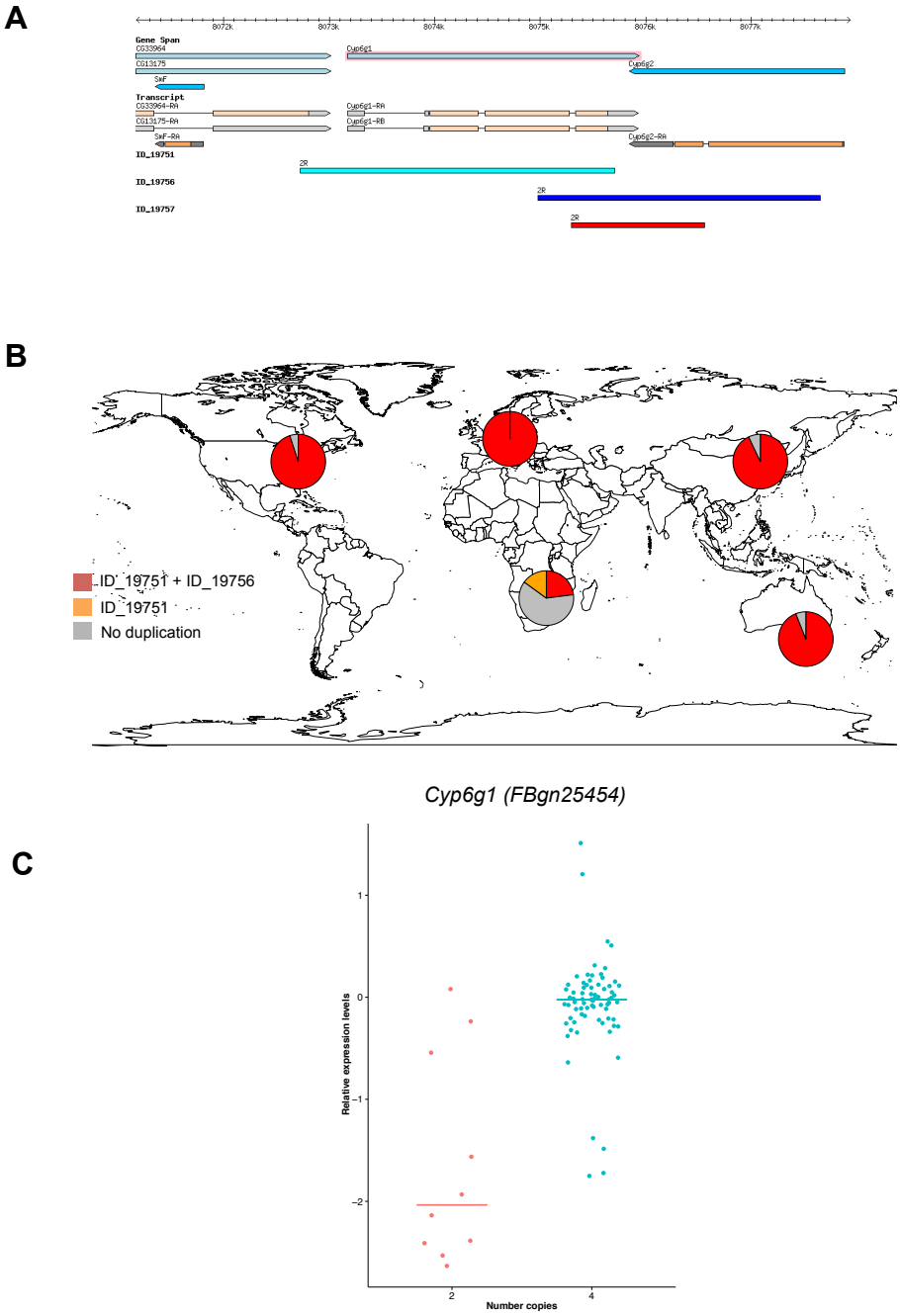
Supp. Figure 6. Relationship between gene copy number and expression levels for 4 genes segregating multiple independent duplications in the same line(s). (A-C) For these 3 genes there is a positive correlation between gene copy number and expression levels. (D) For *CG7966* the lines with the highest expression levels are those carrying 3 copies of the gene, those carrying 4 and 6 copies have, on average, intermediate expression levels between lines carrying no duplication and those carrying a heterozygous duplication.



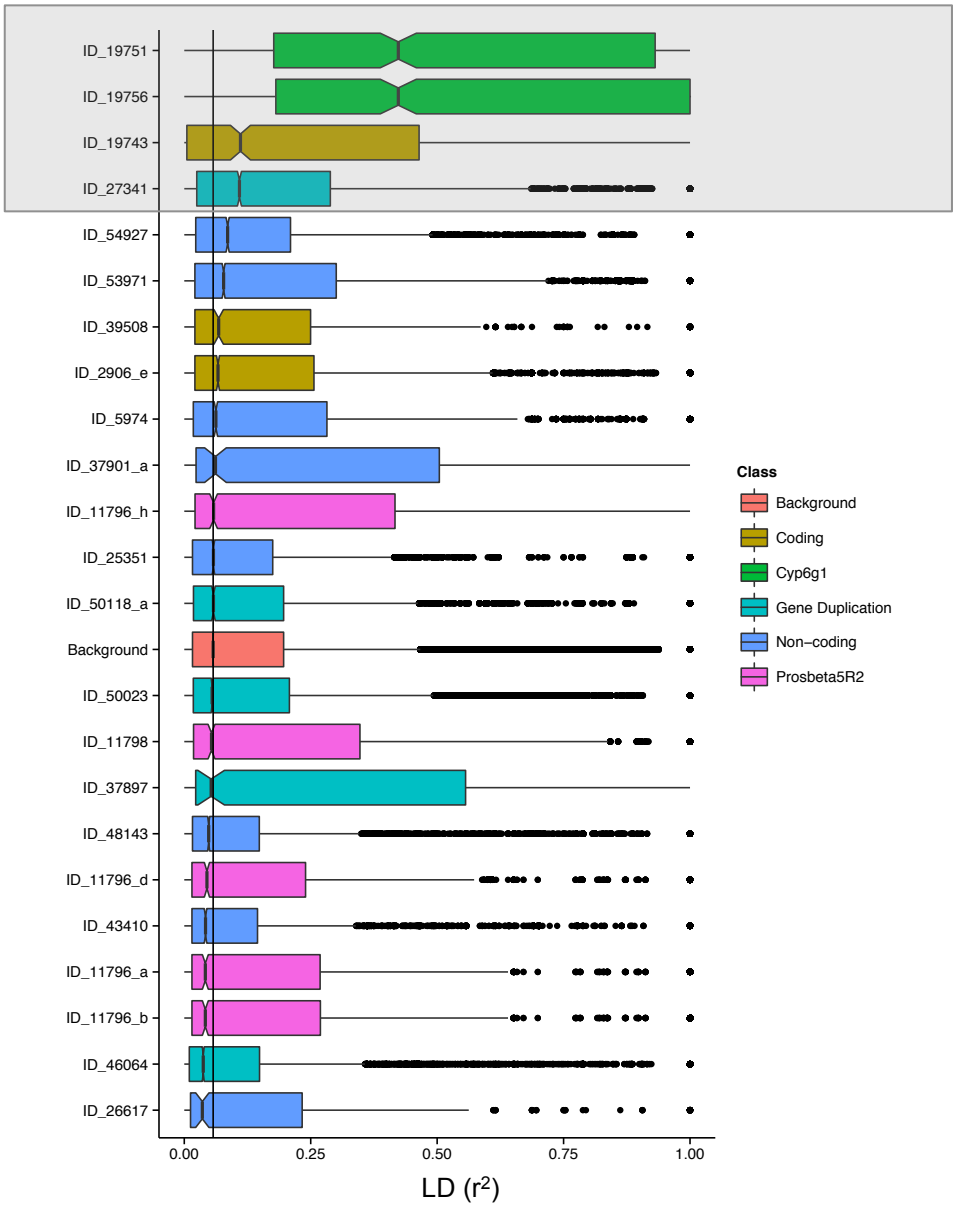
Supp. Figure 7. Relationship between gene copy number and expression levels for duplications classified as leading to significant changes in gene expression. Expression levels for genes duplicated in only one line in lines not carrying the duplication and in lines carrying heterozygous and homozygous duplications. These data differs from those on Fig. 2 because it only includes gene duplications classified as leading to significant changes in gene expression.



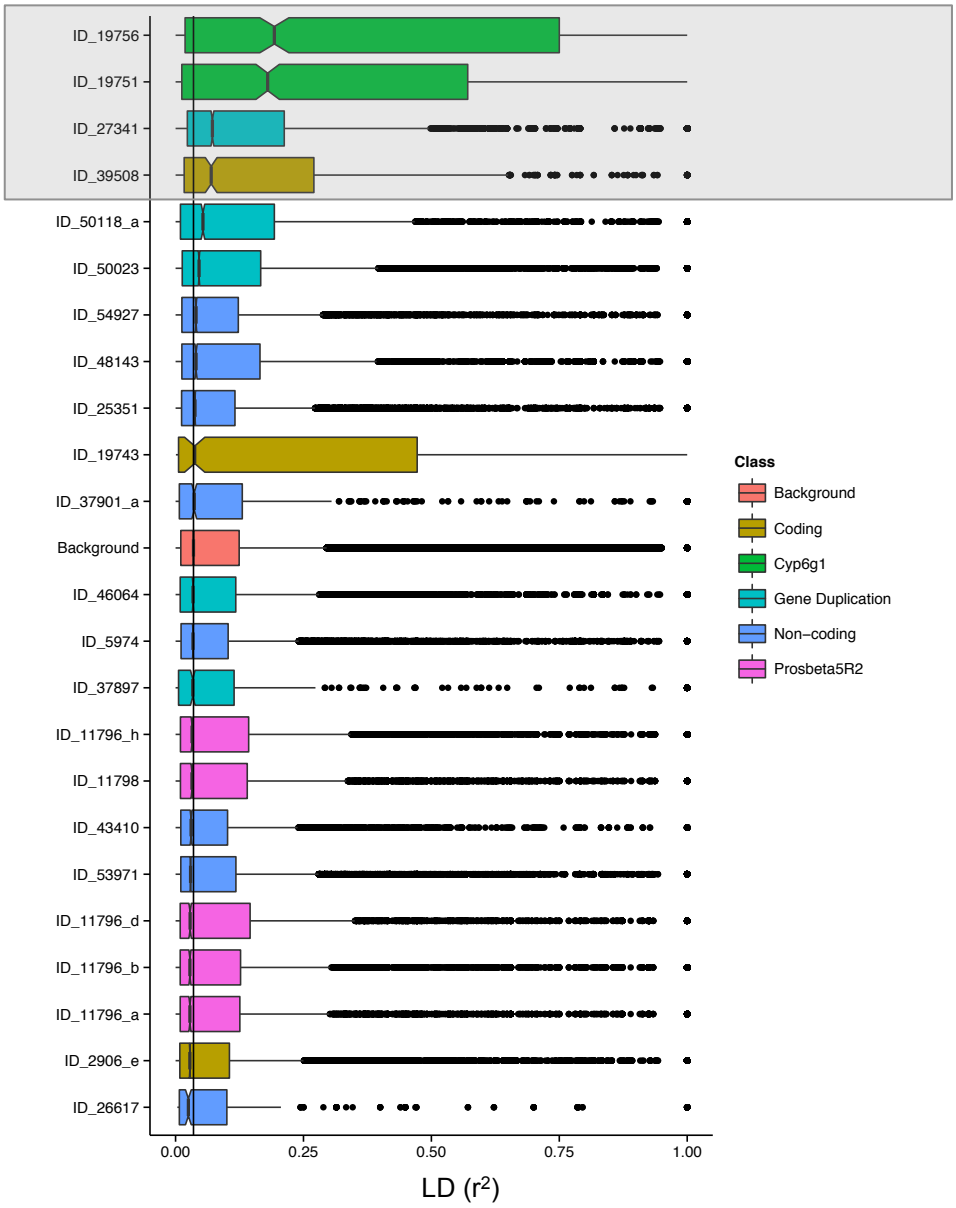
Supp. Figure 8. Gene duplications are enriched among high-frequency duplications (whole genome). Fraction of duplications overlapping different genomic contexts in bins of increasing frequency. The data shown represent the combined data from the five populations and for all duplications (differs from Fig. 3 which only shows the data for autosomal duplications).



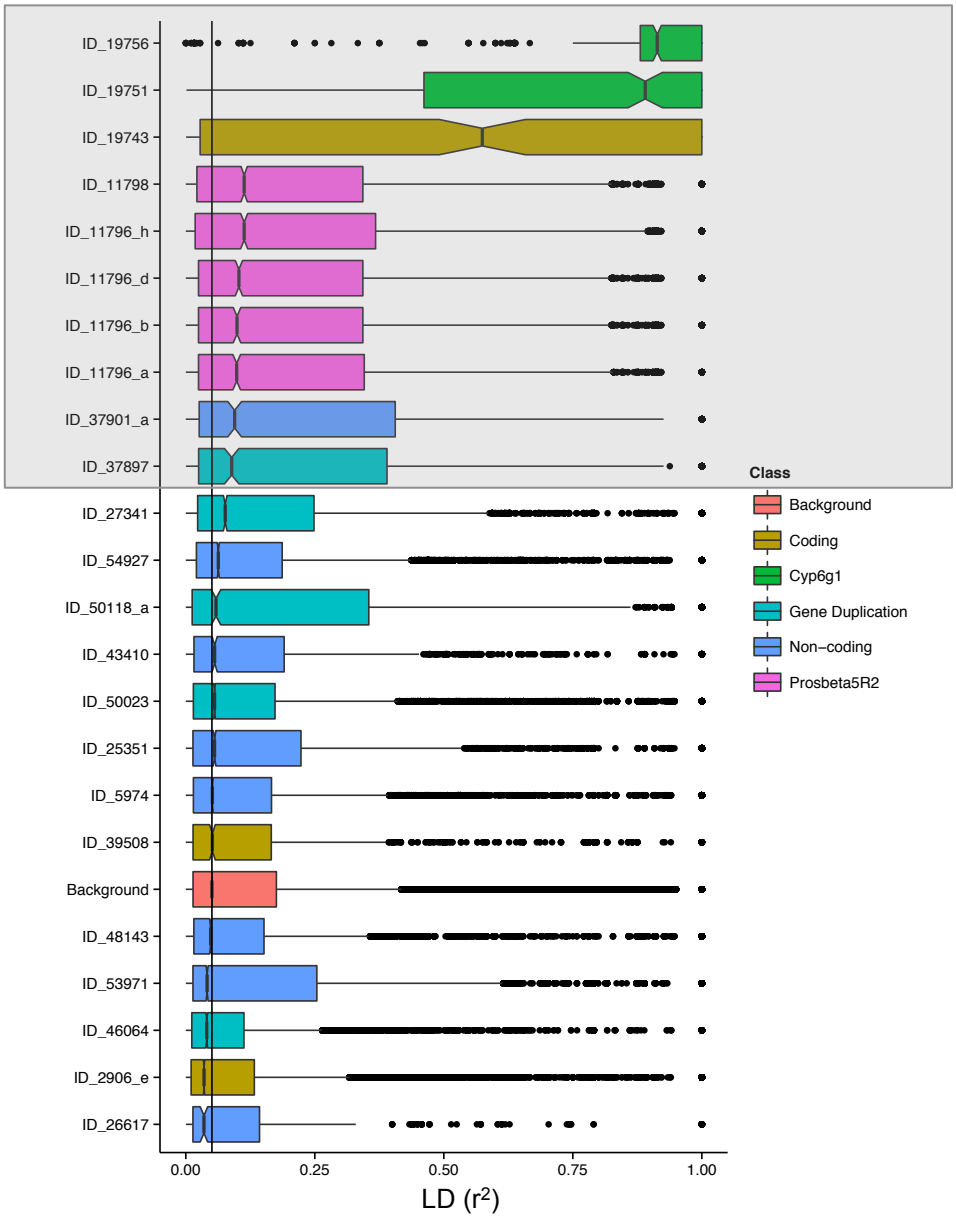
Supp. Figure 9. Characterization of the *Cyp6g1* duplication. (A) Genome Browser (Flybase) view of the 3 duplications overlapping *Cyp6g1*. Only ID_19751 duplicates the whole coding region of the gene. (B) Frequency of the *Cyp6g1* duplication (ID_19751) in the five populations. (C) Expression levels in lines with and without the *Cyp6g1* duplication (ID_19751).



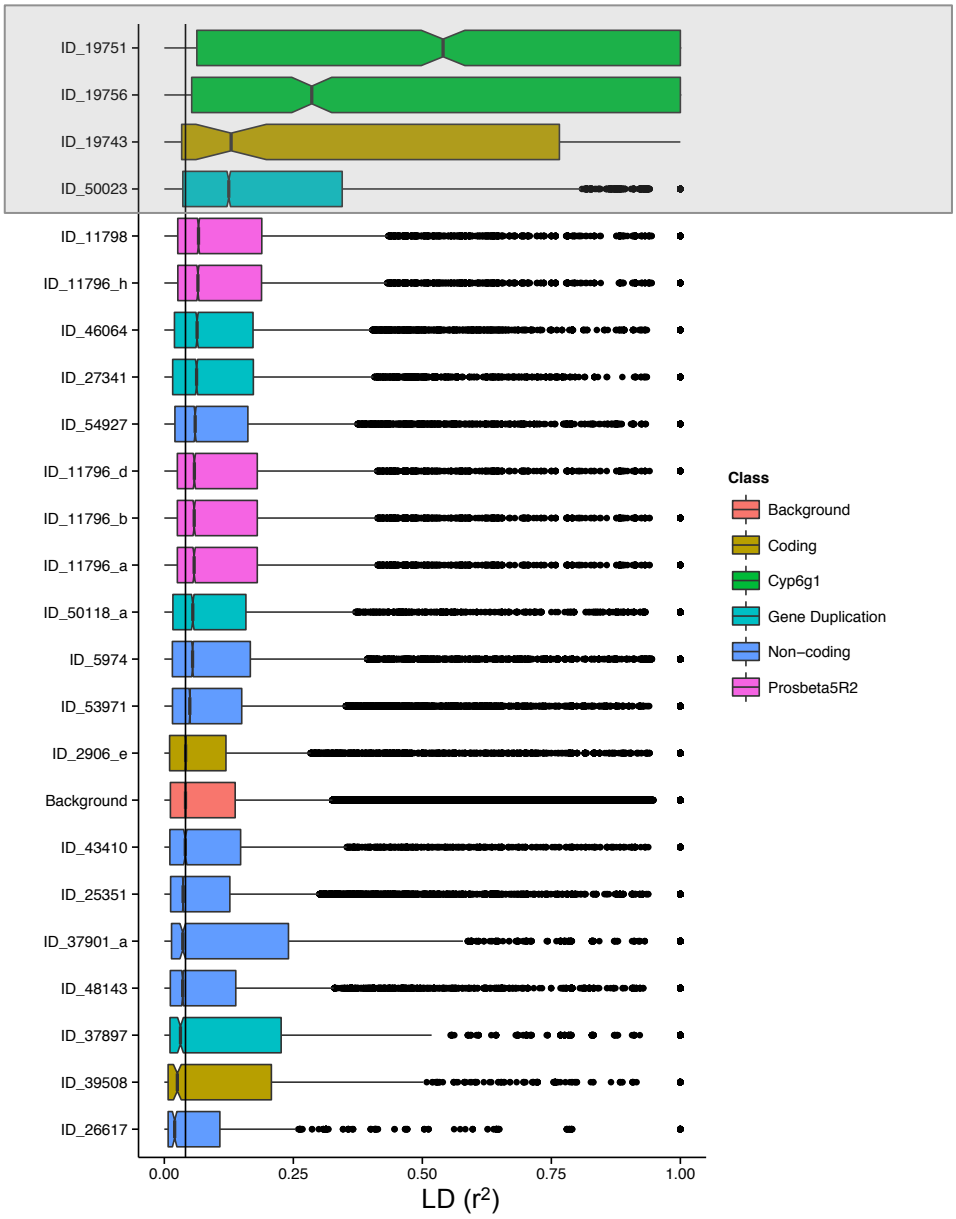
Supp. Figure 10. Extent of LD between SNPs located within 5 kb upstream and downstream the set of high-frequency duplications (and the *Prosbeta5R2* gene duplications). The data are from the **Beijing** population. The shaded box surrounds the set of high-frequency duplications showing the highest levels of LD. The vertical line corresponds to the median LD observed for all duplications segregating in the GDL (*i.e.* background expectation).



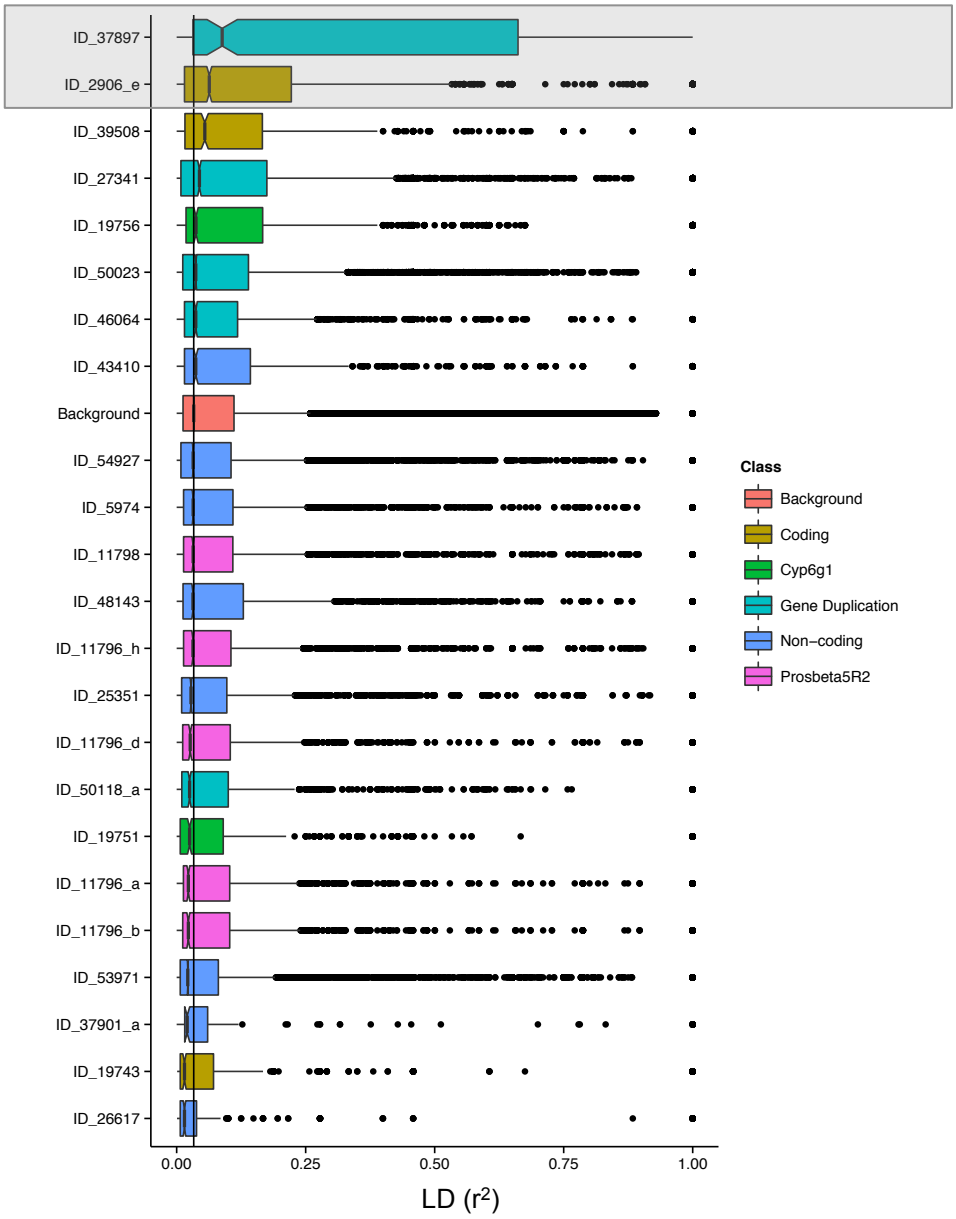
Supp. Figure 11. Extent of LD between SNPs located within 5 kb upstream and downstream the set of high-frequency duplications (and the *Prosbeta5R2* gene duplications). The data are from the **Ithaca** population. The shaded box surrounds the set of high-frequency duplications showing the highest levels of LD. The vertical line corresponds to the median LD observed for all duplications segregating in the GDL (*i.e.* background expectation).



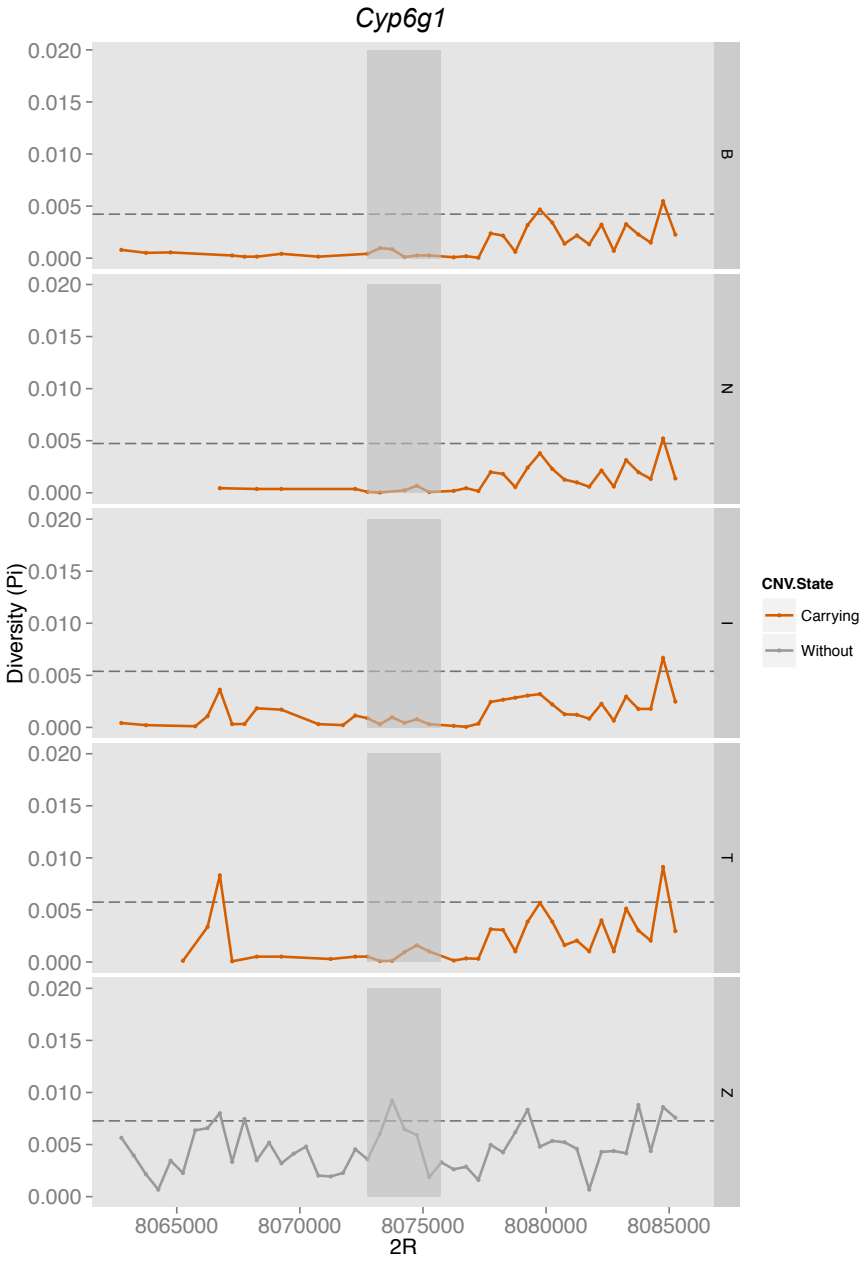
Supp. Figure 12. Extent of LD between SNPs located within 5 kb upstream and downstream the set of high-frequency duplications (and the *Prosbeta5R2* gene duplications). The data are from the **Netherlands** population. The shaded box surrounds the set of high-frequency duplications showing the highest levels of LD. The vertical line corresponds to the median LD observed for all duplications segregating in the GDL (*i.e.* background expectation).



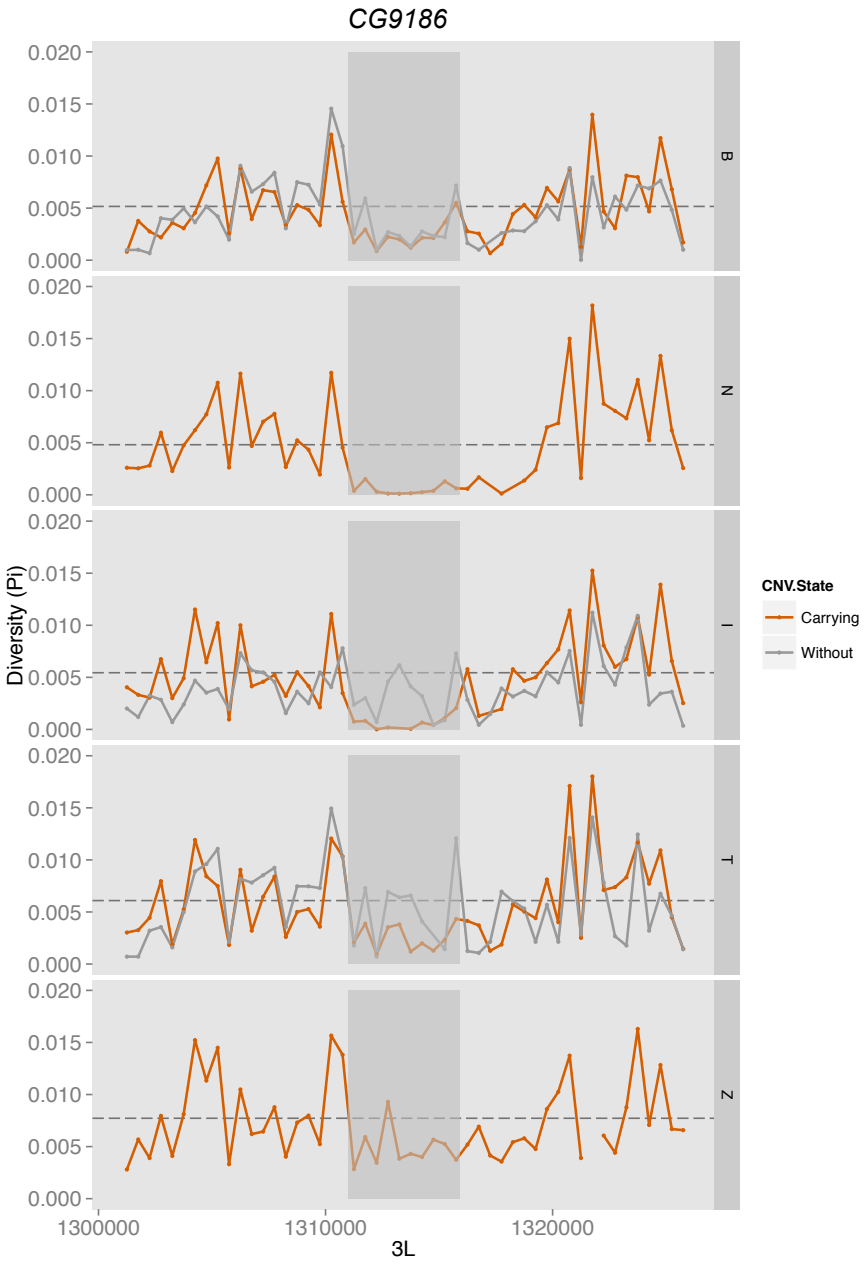
Supp. Figure 13. Extent of LD between SNPs located within 5 kb upstream and downstream the set of high-frequency duplications (and the *Prosbeta5R2* gene duplications). The data are from the **Tasmania** population. The shaded box surrounds the set of high-frequency duplications showing the highest levels of LD. The vertical line corresponds to the median LD observed for all duplications segregating in the GDL (*i.e.* background expectation).



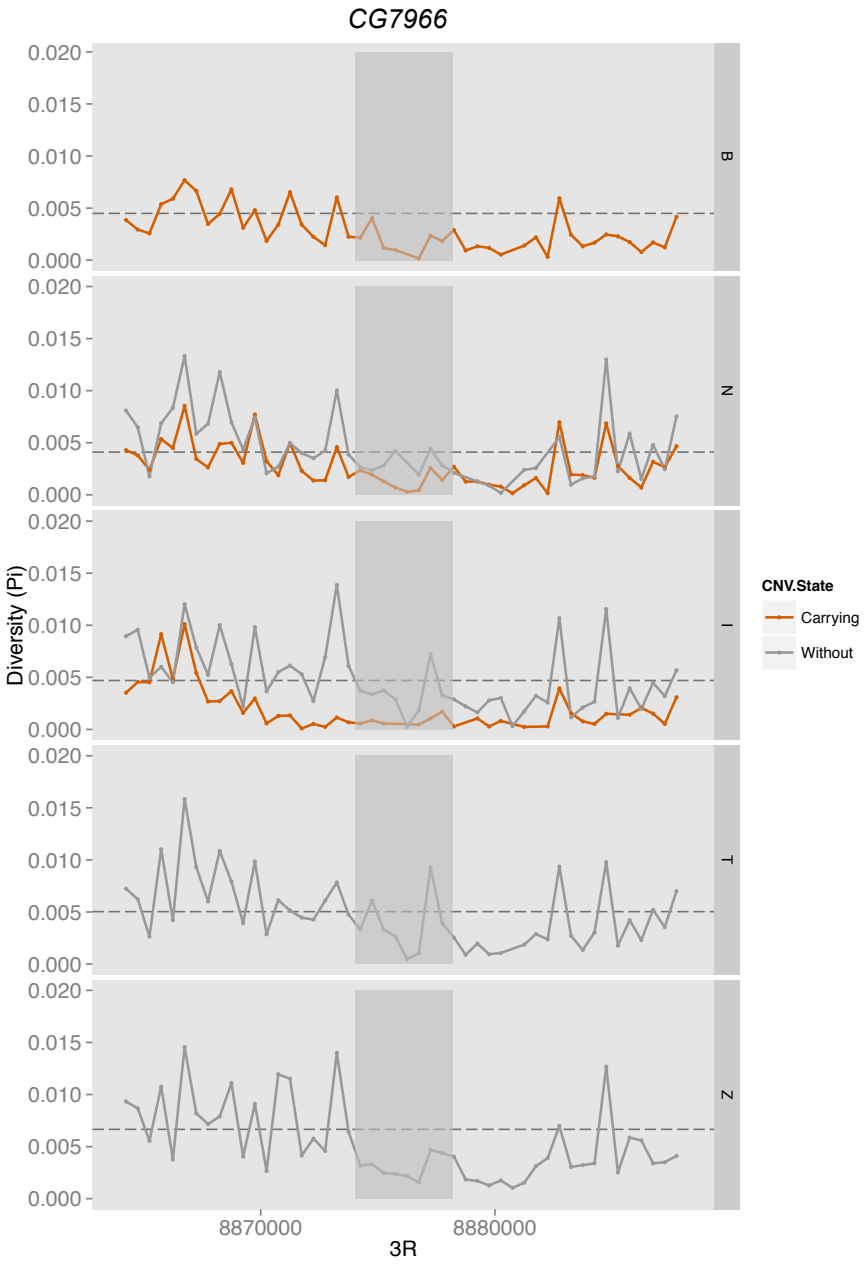
Supp. Figure 14. Extent of LD between SNPs located within 2.5 kb upstream and downstream the set of high-frequency duplications (and the *Prosbeta5R2* gene duplications). The data are from the **Zimbabwe** population. The shaded box surrounds the set of high-frequency duplications showing the highest levels of LD. The vertical line corresponds to the median LD observed for all duplications segregating in the GDL (*i.e.* background expectation).



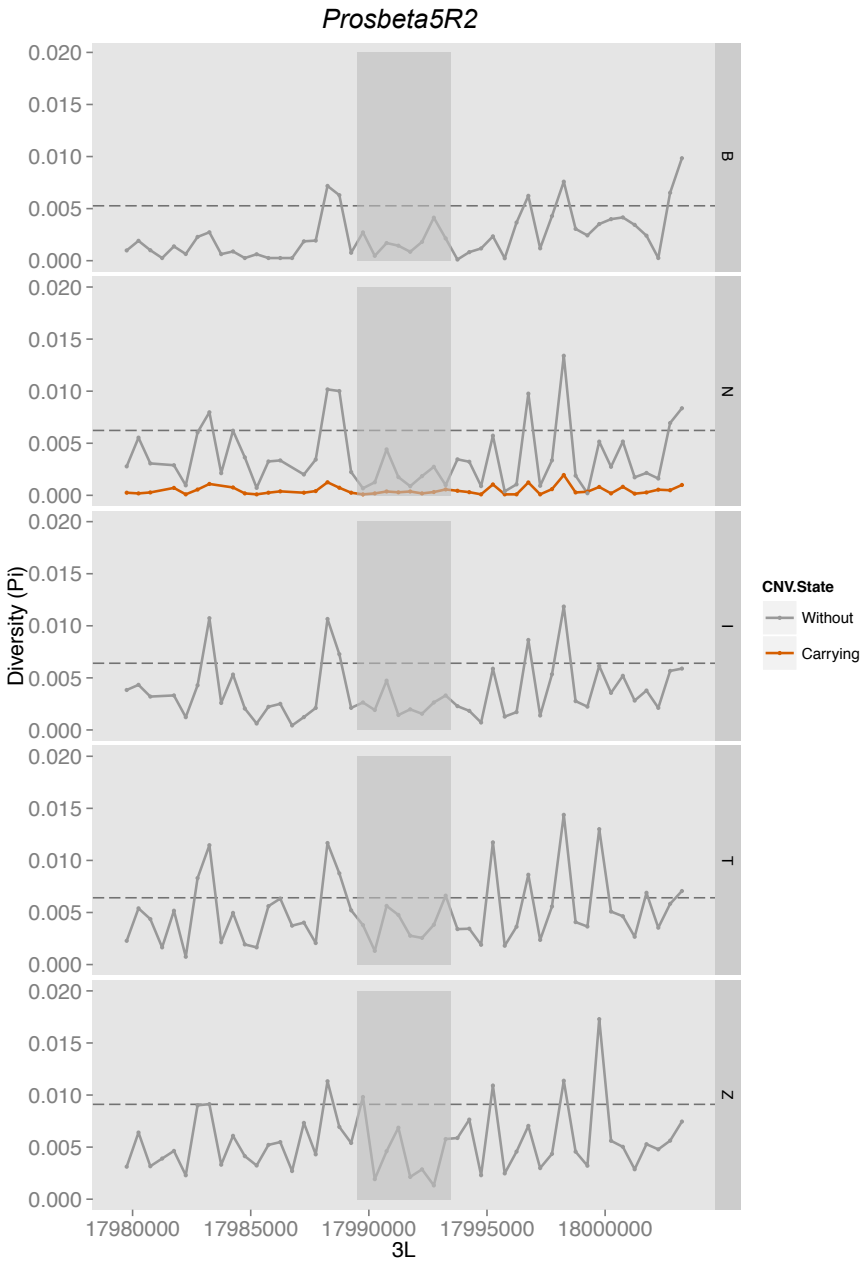
Supp. Figure 15. Reduced levels of nucleotide diversity flanking the *Cyp6g1* gene duplication in the non-African populations. The grey box marks the limits of the duplication and the dashed horizontal line indicates the median diversity levels for the chromosome in each population. A minimum of 5 lines carrying or not carrying the duplication was required to plot the diversity levels.



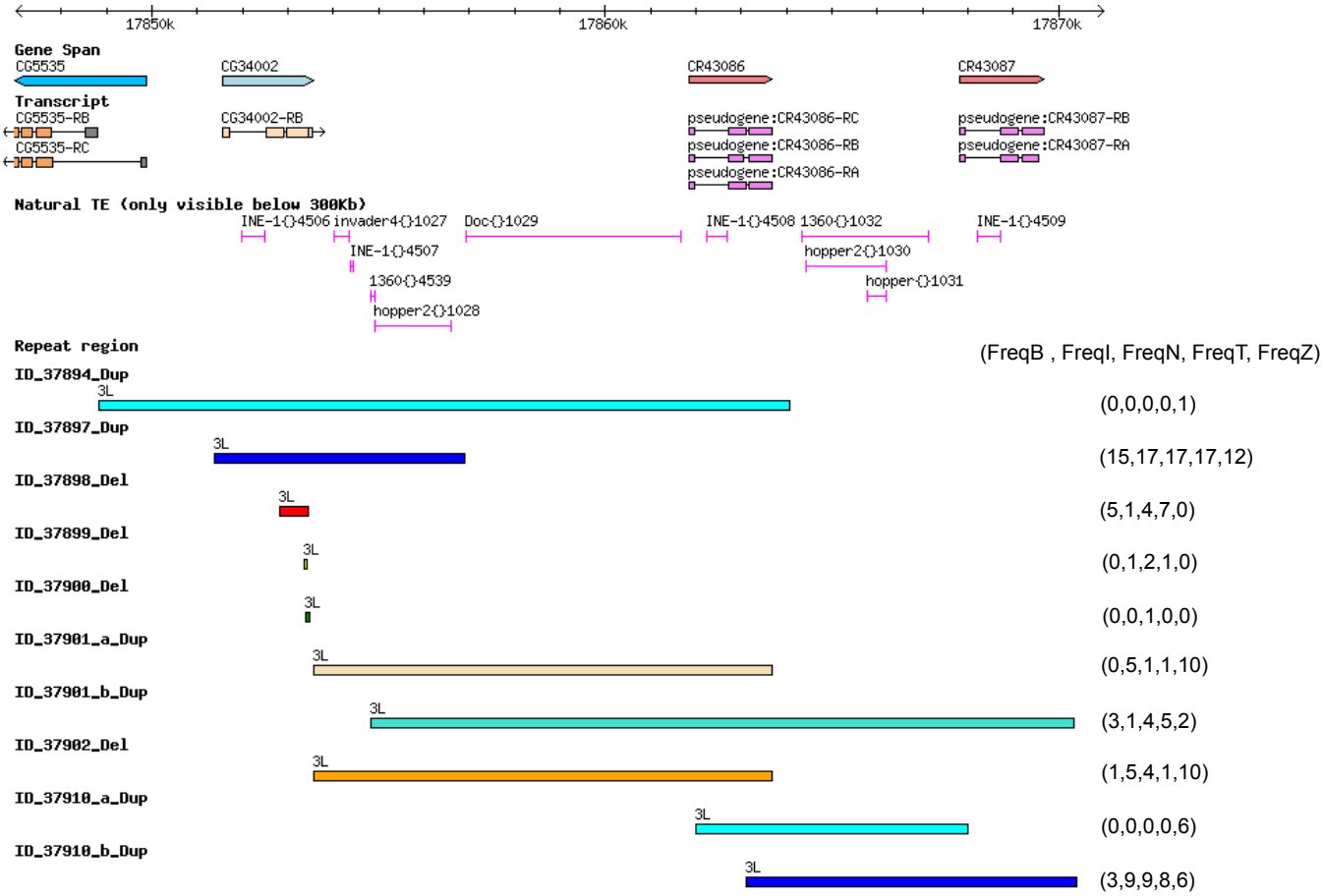
Supp. Figure 16. Nucleotide diversity levels flanking the *CG9186* gene duplication in all populations. There is only reduced diversity in the Netherlands. The grey box marks the limits of the duplication and the dashed horizontal line indicates the median diversity levels for the chromosome in each population. A minimum of 5 lines carrying or not carrying the duplication was required to plot the diversity levels.



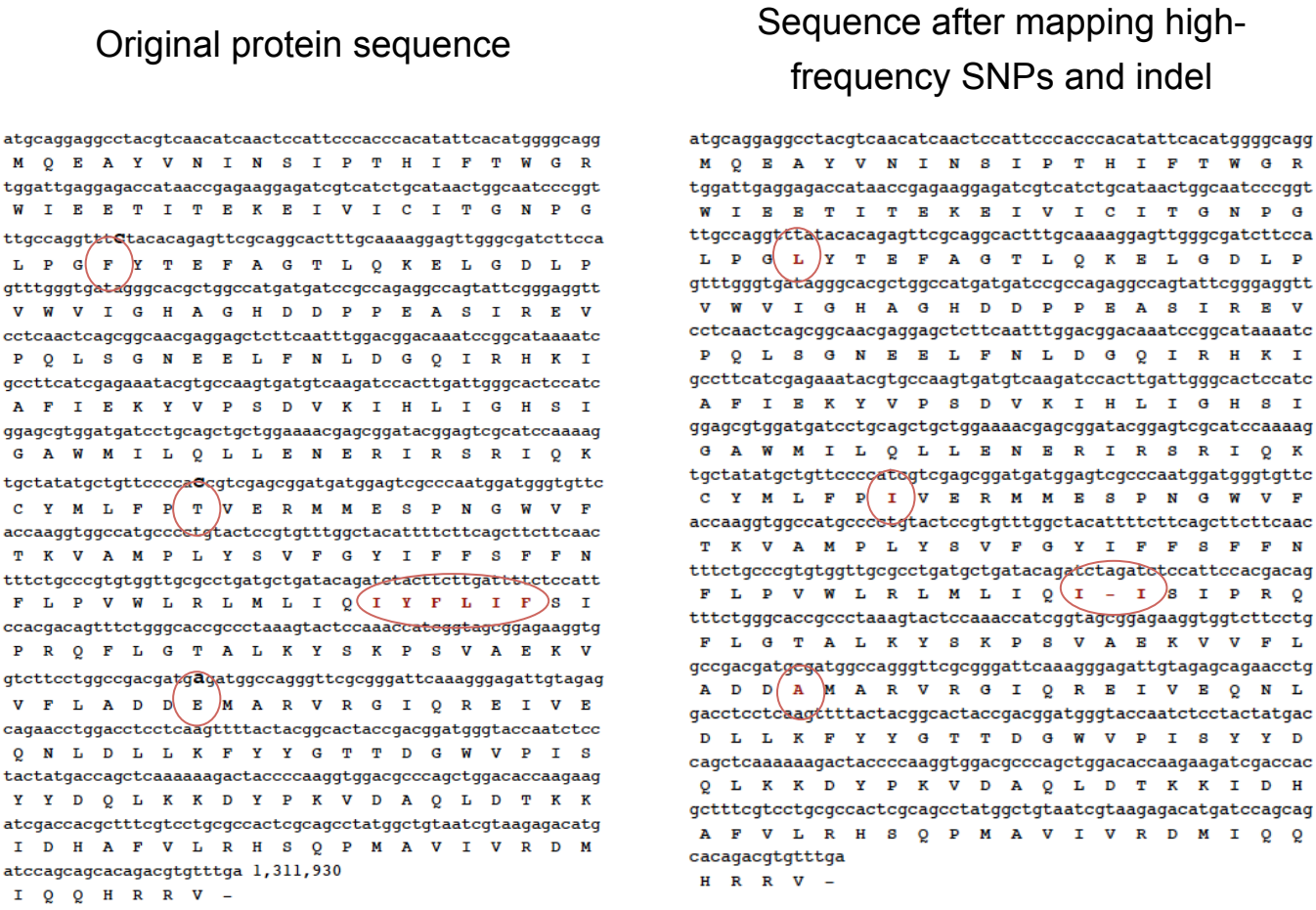
Supp. Figure 17. Nucleotide diversity levels flanking the *CG7966* gene duplication in all populations. There is only reduced diversity in Ithaca. The grey box marks the limits of the duplication and the dashed horizontal line indicates the median diversity for the chromosome in each population. A minimum of 5 lines carrying or not carrying the duplication was required to plot the diversity levels.



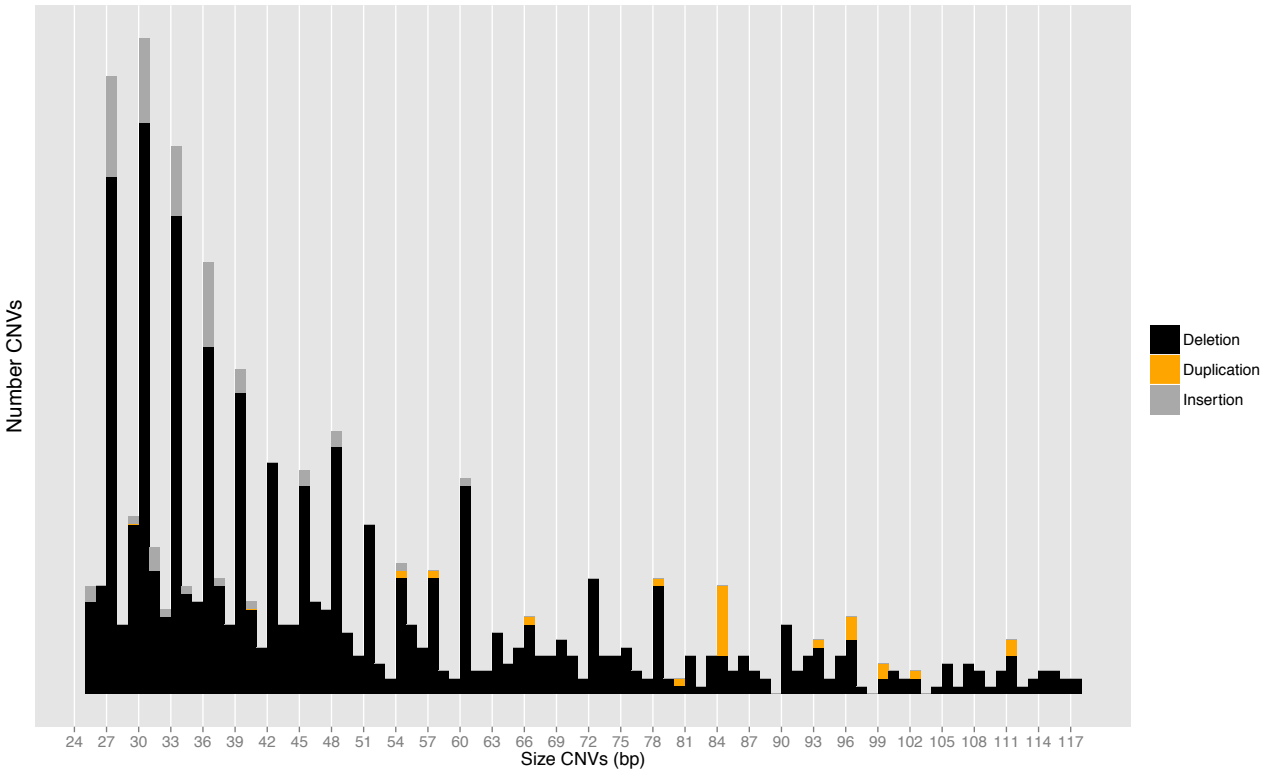
Supp. Figure 18. Nucleotide diversity levels flanking the highest-frequency *Prosbeta5R2* gene duplication in all populations. There is only reduced diversity in the Netherlands, which is the only population where this duplication is segregating in high-frequency. The grey box marks the limits of the duplication and the dashed horizontal line indicates the median diversity levels for the chromosome in each population. A minimum of 5 lines carrying or not carrying the duplication was required to plot the diversity levels.



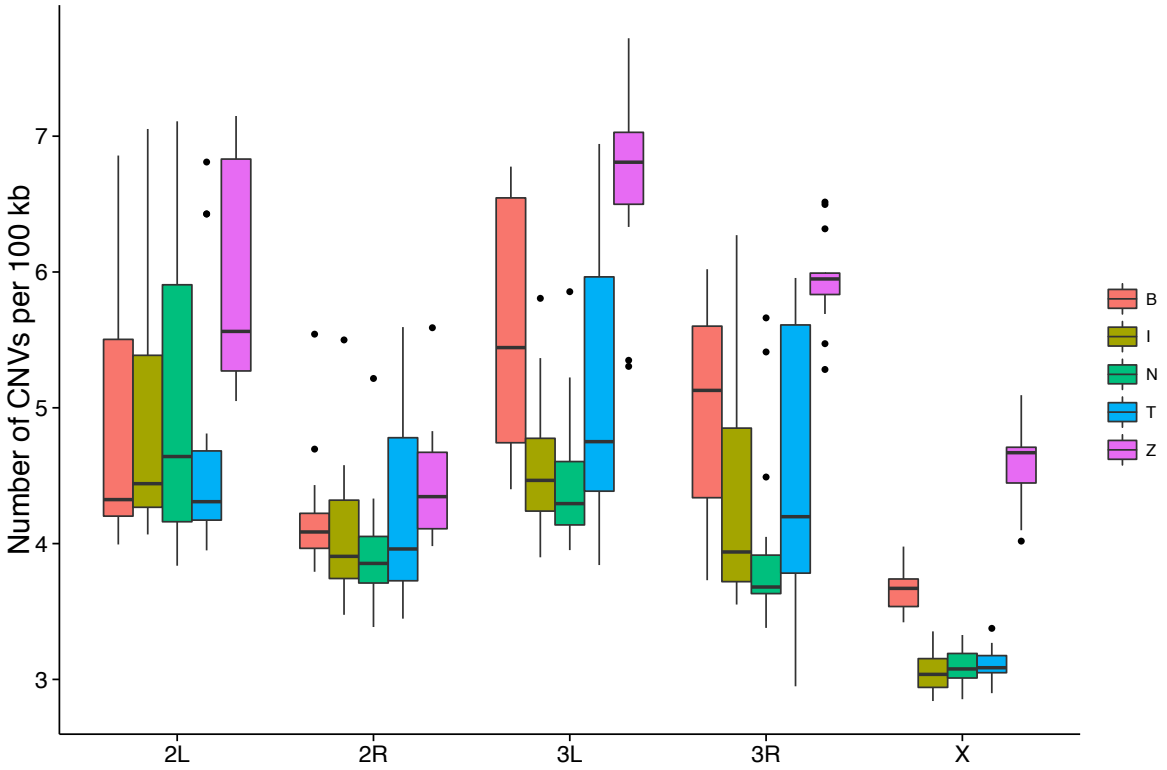
Supp. Figure 19. CNVs segregating in the *CG34002* locus. The numbers in parenthesis represent the frequency of each CNV in each of the 5 populations (*i.e.* Beijing, Ithaca, Netherlands, Tasmania, Zimbabwe).



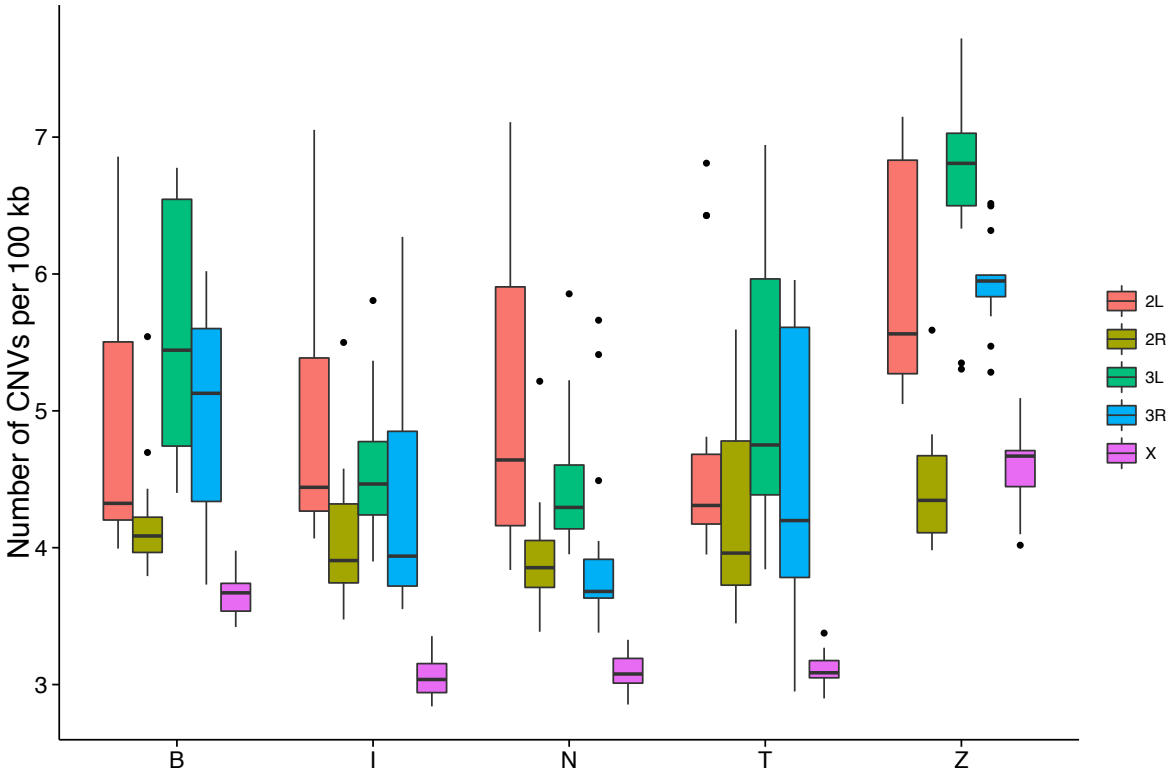
Supp. Figure 20. Additional mutations segregating with the *CG9186* gene duplication. The SNPs occur in positions conserved between flies and mammals and the indel in positions conserved within insects (conservation information was obtained fromThiel *et al.* 2013). Note that we mapped all substitutions to the protein sequence but that we actually do not know how these mutations are distributed between the two duplicated genes.



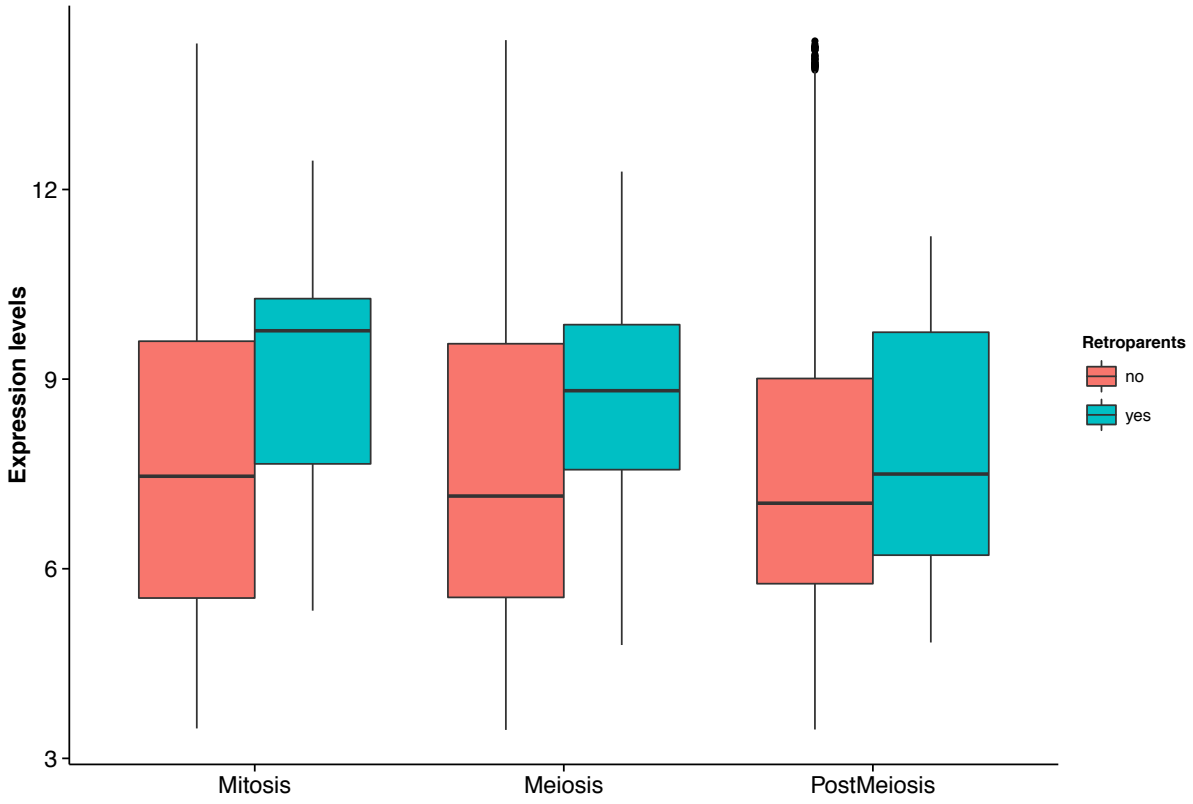
Supp. Figure 21. Size distribution of CNVs that occur within a coding exon. The vertical white bars mark the sizes where the CNVs are multiples of 3 and therefore less likely to severely affect the protein sequence.



Supp. Figure 22. CNV density per chromosome (per population). The boxplots summarize the estimated CNV density for each individual line within each of the populations (*n* = 13-19). Each population is represented by a different color and marked by the population's initial.



Supp. Figure 23. CNV density per population (per chromosome). The boxplots summarize the estimated CNV density for each individual line within each of the populations (*n* = 13-19). Each chromosomal arm is represented by a different color. The populations are marked by the population's initial.



Supp. Figure 24. Expression levels of genes with and without retrogenes throughout sperm development. The data was retrieved from the SpPress database (Vibranovski *et al.* 2009).

chr	CNV start	CNV end	CNV ID	CNV Size	CNV Type	Pipeline	Coverage Support	forward primer start	reverse primer end	PCR result	Primer ID	size PCR band if CNV absent	size PCR band if CNV present	Sanger sequencing of the PCR?	forward primer sequence	reverse primer sequence
2L	1020326	1020516	ID_504	190	Deletion	Pindel_mySR	Yes	1020241	1020640	Confirmed	A1	399	209	No	CCGCCTGAATGCGCAGTAATTAAG	ATGCTGAGGATGGCAGAAAGGATG
2L	1402392	1402786	ID_756	394	Deletion	mySR_Delly	Yes	1402238	1402979	Confirmed	B5	741	347	No	AAAGAAAGCAGAGTGGGTCGTGGA	GGTTCTAAGGAACCAGGAAAGTTGC
2L	3130852	3131061	ID_1868	209	Deletion	Pindel_mySR	Yes	3130733	3131258	Confirmed	A2	525	316	No	GTTAAGCGCAAACGCATGTCCTGT	TGGTCCGTGCATAATGGGTTATCT
2L	4164497	4167300	ID_2661	2803	Duplication	Pindel_mySR	Yes	4167030	4164733	Confirmed	Dup1	No band	508	No	TCTTAGAACTGCAGGGACTTTAC	CGCTTCACTGGATCTCTGATT
2L	4926263	4926309	ID_3109	46	Deletion	mySR_Delly	Yes	4926161	4926343	Confirmed	A4	182	136	No	AGACATCTAGACGCGATGTGTG	ACACTGCGTATGCGTGACTCGTTT
2L	5324893	5330212	ID_3293	5319	Duplication	Pindel_mySR	No	5329965	5325022	Confirmed	Dup2	No band	378	No	TTAGCTGCCTTATGGTCTTCTC	CATCGCCAGCAATTTCACTATC
2L	5326095	5329983	ID_3294	3888	Deletion	Pindel_mySR	No	5325808	5330204	Confirmed	Del1	4396	508	No	TGGGTGACGCAAGTGTAAA	CCTTGTAACGCCTCAAGAGAT
2L	11651186	11654233	ID_7386	3047	Deletion	Pindel_mySR	No	11651146	11654444	Confirmed	Del2	3298	251	No	TACCACTAAATGACGACCAACC	GCAGCTGGGTAACTACAAACTA
2L	12171337	12171526	ID_7733	189	Deletion	Pindel_Delly	Yes	12171267	12171681	Confirmed	A9	414	225	Yes	GGTAGTGGTTATTCGGATATGGCAGGT	ATCCTCACCACAAACGAAACACGC
2L	12171383	12171471	ID_7735	88	Deletion	mySR_Delly	Yes	12171267	12171681	Confirmed	A9	414	326	Yes	GGTAGTGGTTATTCGGATATGGCAGGT	ATCCTCACCACAAACGAAACACGC
2L	12276192	12276224	ID_7859	32	Deletion	Pindel_mySR	Yes	12276107	12276289	Confirmed	A10	182	150	Yes	CAGGTTCCGAAGAAGTAAATTTCGAG	ATGATGACTAGATTTCTCTATGCTCGT
2L	16741768	16744279	ID_10906	2511	Duplication	Pindel_mySR	No	16743991	16742003	Confirmed	Dup3	No band	528	No	GGATCACGAACGTCAGAAGAA	CTGCTGGTTAGTCACGAAGAA
2L	17332984	17333171	ID_11324	187	Deletion	Pindel_Delly	Yes	17332764	17333261	False Positive	A12	497	310	Yes	CTACCAGCAAACAACATCGGCAAC	TCTCTGATTTGGCTCGCCTGCATA
2L	17406158	17406416	ID_11394	258	Deletion	Pindel_mySR	No	17406032	17406563	Confirmed	B19	531	273	No	AGCGGATAGTTCTGCTCTGGATGA	TTCAGCTGACTTGCTTTGGAGCTG
2L	17496033	17501678	ID_11449	5645	Duplication	Pindel_mySR	Yes	17501538	17496290	Confirmed	Dup4	No band	403	No	CCTATGCGAGAAGTCGAAGATG	GTGAGCTGGGACAAACAGATAA
2L	17989516	17993458	ID_11798	3940	Duplication	mySR_Delly	Yes	17993324	17989727	Confirmed	Dup5	No band	347	No	AGTGACTTGGTGACTGAAGTG	TGTTTCCCTTTCTAGTGTAGG
2L	18022535	18022825	ID_11821	290	Deletion	Pindel_mySR	Yes	18022443	18023092	Confirmed	B18	649	359	No	TTTCTGTGCACTTCTGAGGGCGAT	AGTGTTGCCAAGGACTCAAGGACA
2L	18338661	18338987	ID_12058	326	Deletion	Pindel_mySR	Yes	18338480	18339111	Confirmed	B17	631	305	No	AACCCGGCAAACGGGTAAGAAA	TACTGTGCCGTCGCAAAGCACAAA
2L	18826632	18826820	ID_12345	188	Duplication	Pindel_mySR	No	18826441	18827092	Confirmed	Dup6	651	839	No	ATGAACAACTGCAAGGAACAAC	GCGGAAACAGACCCAGAAT
2L	19560028	19560256	ID_12712	228	Deletion	Pindel_mySR	Yes	19559982	19560342	Confirmed	A19	360	132	No	TCGGGACTTTACGTATTCCAGGCA	TGGGCTCGTCGAACTCAACTCTTT
2L	19856444	19856636	ID_12916	192	Deletion	Pindel_mySR	Yes	19856347	19856772	Confirmed	B16	425	233	No	ATGTTACGCTGCGAGTGTGGAAAG	TTCTGCTATCTTCGCCAGCCTTCT
2L	20509967	20510587	ID_13420	620	Deletion	Pindel_mySR	Yes	20509887	20510731	Confirmed	A13	844	224	No	TCACAGGATAAGCAACCAGGAGCA	ACACTCACACACACACACCGAA
2L	21094034	21096524	ID_13762	2490	Duplication	Pindel_mySR	Yes	21096307	21094236	Confirmed	Dup7	No band	437	No	GCACAATGGCACAGGAATTAG	ATCCAGTCCACGCCTAAATG
2L	21260553	21264957	ID_13832	4404	Duplication	Pindel_mySR	Yes	21264705	21260810	Confirmed	Dup8	No band	517	No	CTGACTTCGTCTGCTTCTGTAA	CAGTTGTGGGCGTTAGTATGA
2R	1243759	1243785	ID_16117	26	Deletion	Pindel_Delly	Yes	1243684	1243875	Confirmed	A40	191	165	Yes	ACGTAGCGATAAGTTCGGGTGCAT	ATTCCACCACTCCGTCATTCTGGT
2R	3807503	3812056	ID_17642	4553	Duplication	Pindel_mySR	Yes	3811865	3807631	Confirmed	Dup9	No band	345	No	CTGCATACAGTGGCGATATGT	GATGTTAATTCGGCGCTTCTTG
2R	5100681	5103740	ID_18237	3059	Deletion	Pindel_mySR	No	5100038	5103902	False Positive	Del5	3864	805	No	CTTCTGTGTTTCTGCCTCTCTC	GTGGCTATTTGTGGCCAATTC
2R	8250939	8251320	ID_19831	381	Deletion	Pindel_mySR	Yes	8250790	8251388	Confirmed	A20	598	217	No	GGCTTTAATGCATTGCCTTTGCGG	AGACTCACTTTGCGACTCCTTCGT
2R	12569636	12569766	ID_22030	130	Deletion	Pindel_mySR	No	12569550	12569960	Confirmed	B25	410	280	Yes	GCAGTCTGAGTAGTCATTGTCTCG	TTAGCTGTCTTCCACTCATCCTGG
2R	13363929	13372411	ID_22495	8482	Duplication	Pindel_mySR	Yes	13372212	13364060	Confirmed	Dup10	No band	332	No	TAACGGAACACCTCCAGAGT	GGTAACTTGGGAATCGGCTATG
2R	13886796	13886964	ID_22777	168	Deletion	Pindel_mySR	Yes	13886626	13887106	Confirmed	B24	480	312	No	ACTGAGCCACCGAATTACCCAGAT	GCTTTAACCCTTGTTTGCCCTGCT
2R	14403274	14403377	ID_23069	103	Deletion	Pindel_mySR	Yes	14403189	14403519	Confirmed	B23	330	227	Yes	CTCAACAGGCCAACAGTTGATGCT	AAATGTTTCACTCCCAGCCCAAGC
2R	14501211	14504070	ID_23117	2859	Duplication	Pindel_mySR	Yes	14503795	14501407	Confirmed	Dup11	No band	493	No	CTGATGGATGCCCAGGATTT	CTCCGAATGTCAGGTCTTTGT
2R	14761796	14761972	ID_23248	176	Deletion	Pindel_Delly	Yes	14761530	14762250	Confirmed	A22	720	544	No	ACAAGGTGAGGTGAGACACACACA	ACCCAACTCCAGACTAATTGCCCA
2R	17428222	17428455	ID_24736	233	Deletion	Pindel_mySR	Yes	17428115	17428606	Confirmed	A36	491	258	No	GTTTGCCACTGGGAGGTTTCGTTT	ATCTCCAGCGAAATTCCTCGCAAC
2R	17623237	17623358	ID_24840	121	Deletion	Pindel_mySR	Yes	17623088	17623508	Confirmed	B21	420	299	Yes	AAATCCCAGAAGCACGTTCATCCC	GCCTTACAGCGCCAAACATCAACTT
2R	19482411	19482575	ID_25937	164	Deletion	Pindel_mySR	Yes	19482242	19482742	Confirmed	B20	500	336	No	AATCGCTGGGTTCGAGTAGAGCAT	AGCCCATGAGATGCCAGAGTTTCT
2R	20152663	20158229	ID_26160	5566	Duplication	Pindel_Delly	No	20157957	20152770	Confirmed	Dup12	No band	381	No	CAGGCATTCCTAGGGCTTAAT	ACAAACGCCAAAGAGATGAAAC
2R	20153629	20157099	ID_26163	3470	Deletion	Pindel_mySR	No	20153340	20157224	Confirmed	Del6	3884	414	No	GAATTCCTGCCCTGCAGTATAA	GCAATTGACAGCCACTGAAAG
3L	95019	96235	ID_26642	1216	Deletion	mySR_Delly	Yes	94832	97045	Confirmed	A23	2213	997	No	TCGACACGTGTAGGAAGTAGCCTT	TGGCCACACGCCCTTTAAGATAGA
3L	95019	96953	ID_26643	1934	Deletion	Pindel_Delly	Yes	94832	97045	Confirmed	A23	2213	279	No	TCGACACGTGTAGGAAGTAGCCTT	TGGCCACACGCCCTTTAAGATAGA
3L	1250975	1252279	ID_27299	1304	Duplication	Pindel_mySR	No	1252096	1251230	False Positive	Dup13	No band	465	No	GCGAGTGTTCGGGCATATAA	CCTATCGAAGATCGGCTTCATC
3L	2073470	2073902	ID_27751	432	Deletion	Pindel_Delly	Yes	2073301	2073996	Confirmed	A26	695	263	No	AGGCTTTAGATCCTCTGTGCGTGT	AAAGTTCAGACCACGACGGTGACT
3L	2204024	2204356	ID_27823	332	Deletion	Pindel_mySR	Yes	2203879	2204524	Confirmed	B13	645	313	No	GCAGGAACAGATTATGCAGGGCAA	GATGTGGCAATTGAGACAGGCA
3L	2690145	2690247	ID_28170	102	Deletion	Pindel_mySR	Yes	2689998	2690303	Confirmed	B3	305	203	No	TTTGGCAGTGGGAGAAAGTTGCAC	AACCCACTTAGCAGGGCCAACATA
3L	3267197	3271645	ID_28521	4448	Duplication	Pindel_Delly	Yes	3271429	3267334	Confirmed	Dup14	No band	355	No	GCGCCACTTATTGTGTTGATAG	TTTCCTGAGAGAGAGAGAGAGAG
3L	4443461	4443461	ID_29203	30	Insertion	Pindel_mySR	No Info	4443399	4443545	Confirmed	A28	146	176	Yes	ACGTTCTTCTTCGGCAGCTTTCAC	CGAAGAACTCTCTGGAATGGCA
3L	4601122	4601310	ID_29318	188	Deletion	Pindel_mySR	Yes	4600950	4601444	Confirmed	B15	494	306	No	TGGACTGGCGATGGGATAAAGTGA	ACCCGAGAATTGGTAACTGGCAAC
3L	9352739	9352783	ID_32186	44	Deletion	Pindel_mySR	Yes	9352598	9352902	Confirmed	A30	304	260	Yes	TGAACATCCATCCTGGTGTGGCTT	TGACCTTGTCGATCCAATCCTTCT
3L	9366577	9366613	ID_32191	36	Deletion	Pindel_Delly	No	9366507	9366700	Confirmed	A31	193	157	Yes	GCTCCTGGAATGAATGGGCTCAAA	AGCTTGCCATGGGTGTAGTCAAGA
3L	9497630	9500563	ID_32255	2933	Duplication	Pindel_mySR	No	9500372	9497832	Confirmed	Dup15	No band	398	No	GAGCTGAGTGATTGAGTGTAGTT	TTGCACCTGTCCCGTTATTT
3L	10276638	10276748	ID_32776	110	Deletion	mySR_Delly	No	10276565	10276946	Confirmed	B31	381	271	No	CAGGACAACAGAAGGCGACTATCA	TTCGACTGCTGAGTAGCGAGATTG
3L	10620363	10620841	ID_33017	478	Duplication	Pindel_mySR	No	10620684	10620430	Confirmed	Dup16	No band	226	No	TGTGTGAGTGGCTTGTGTAG	CATTGGTTGTGACGGAGAGT
3L	11155671	11155673	ID_33362	33	Insertion	Pindel_mySR	No Info	11155572	11155734	Confirmed	A32	162	195	Yes	ACGACTAGTTCCGATCCCTACGAT	AATTGCGGGTCAGTGTGCTGT
3L	11223941	11224346	ID_33404	405	Deletion	Pindel_mySR	Yes	11223539	11224617	Confirmed	B30	1078	673	No	GCCCAAGTCAGCTTAAGCAATAGAC	AGTGCGACGCCGATATGAAGAACT
3L	11283776	11284049	ID_33453	273	Deletion	Pindel_mySR	No	11283602	11284213	Confirmed	B29	611	338	No	AACACCCAATGACCTTCAGGGACT	GCAGACACGTTCCTCTTCCGATTT
3L	11572702	11572778	ID_33675	76	Deletion	Pindel_Delly	Yes	11572668	11572808	False Positive	A33	140	64	Yes	TGCGATTCCACACCTGGACAATCT	TGAGGCTCCAATTTAAGTTTCTGAAC
3L	11572719	11572763	ID_33676	44	Deletion	mySR_Delly	Yes	11572668	11572808	False Positive	A33	140	96	Yes	TGCGATTCCACACCTGGACAATCT	TGAGGCTCCAATTTAAGTTTCTGAAC
3L	12044632	12044766	ID_33962	134	Deletion	Pindel_mySR	No	12044517	12044957	False Positive	B27	440	306	Yes	TGCCCAATTGCGGTAGCCAACAA	ACATGTCCGCAGTTGGTGGAGAAA
3L	13456269	13463139	ID_35022	6870	Duplication	Pindel_mySR	Yes	13462862	13456378	Confirmed	Dup17	No band	388	No	TGTCCTCGTTCTTACCTTGTTC	GGGTCTTTCAGGGAACTGATT
3L	15545075	15545105	ID_36475	30	Deletion	Pindel_mySR	No	15544861	15545190	Confirmed	A35	329	299	No	ACGAGAAGCTCAAATGCAGAAGCC	ATTGAGCTCCGGGAGTCTTTCGTT
3L	18084392	18084392	ID_38066	60	Insertion	Pindel_mySR	No Info	18084334	18084567	Confirmed	A37	233	293	Yes	TGAGCTATTAAGGACCGCGTGCAA	GTGGTATTGCAGAACAAAGAACTTCCG
3L	19209989	19210037	ID_38983	48	Deletion	Pindel_mySR	Yes	19209909	19210121	Confirmed	A25	212	164	No	AGAGCGCTTCGTAAACGCCAGTAA	AAATGGGATACGAGAAGGCTCTGC
3L	20386325	20386378	ID_39567	53	Deletion	Pindel_mySR	No	Designed based on insertion site of retrogene		Confirmed	R1	no band	band	Yes	AATCGTAGGCCAGTGGGTAGAAGT	TTCCGTGGCGGTTAGAGAAA
3L	20819474	20819650	ID_39755	176	Deletion	Pindel_mySR	No	20819122	20819777	False Positive	B26	655	479	Yes	TAAAGTGACAACGAGCCACAACGG	AGTTGCATTTCGAGTGGGTGGTTC
3L	21427987	21434135	ID_40031	6148	Duplication	Pindel_Delly	Yes	21433918	21428116	Confirmed	Dup19	No band	348	No	GATTCAGCGGAGCAGGATATT	GGAGAAGGGAGATTGGGATTAG
3R	5399434	5399528	ID_44244	94	Deletion	Pindel_mySR	Yes	5399306	5399626	Confirmed	A43	320	226	No	CAAACTCGTTGGCTGGCAATCACA	CTCAAGGTCAAGGATGTTGTGCGA
3R	5406836	5411316	ID_44249	4480	Duplication	Pindel_mySR	Yes	5411193	5407071	Confirmed	Dup20	No band	371	No	TCAACATGAAGTCCGCTCTG	GTTTAACCCACTCGGATGGAA
3R	6477446	6477618	ID_44669	172	Deletion	Pindel_Delly	Yes	6477248	6477687	False Positive	B10	439	267	Yes	CGGCGTTTGAGTTCTTCATGCCTT	ACGCAGTCGTTGACATCAGAGGAA
3R	7330476	7330742	ID_45148	266	Deletion	Pindel_mySR	Yes	7330194	7331010	Confirmed	B11	816	550	No	AGATGAAGACCAGCAACCTGGTGA	AGACGCTGTCAGAATCTGTTGGGA
3R	7978798	7979022	ID_45530	224	Deletion	Pindel_Delly	Yes	7978624	7979180	Confirmed	B12	556	332	No	AGATGCTTCAACTCATTTGCCCGC	TGCTCGACTTGATTTCCCTTCCCA
3R	9629654	9629837	ID_46507	183	Deletion	Pindel_mySR	No Info	9629530	9629995	Confirmed	B6	465	282	Yes	CCGCCACCCACAAATCTCAAACAT	TAAGCGCTTTGGCGACCTGAATGA
3R	9629765	9629955	ID_46508	190	Deletion	Pindel_mySR	No Info	9629530	9629995	Confirmed	B6	465	275	No	CCGCCACCCACAAATCTCAAACAT	TAAGCGCTTTGGCGACCTGAATGA
3R	10547433	10547504	ID_47025	71	Deletion	Pindel_mySR	Yes	10547367	10547566	Confirmed	A45	199	128	Yes	GGTATCCCTCTTCATAATCTGATTGGT	TCTGCGCATTTGCTAGGGTTATGC
3R	12903705	12903705	ID_48324	30	Insertion	Pindel_mySR	No Info	12903605	12903776	Confirmed	A46	171	201	Yes	GGCGAGTGCATTTGGGTTTGCATT	AGCGGATTTGCGGAGAAGCA
3R	14620350	14625772	ID_49417	5422	Duplication	Pindel_mySR	Yes	14625385	14620722	False Positive	Dup21	No band	761	No	CAACTGGTGTGGTGGTAGAA	AACGGAAGCCAACTACTATCG
3R	14959649	14962218	ID_49660	2569	Deletion	Pindel_mySR	No	14959470	14962263	Confirmed	Del7	2793	224	No	CTTGGAGAAGGAAGCCAAAGA	TGGCAAGGCACAGCAATA
3R	15437714	15441800	ID_50023	4086	Duplication	Pindel_mySR	Yes	15441597	15437879	Confirmed	Dup22	No band	370	No	GTCGACAATCTCAAGGGAAATCT	GCACAAACAAGCGCCTATAATG
3R	15618542	15618542	ID_50140	39	Insertion	Pindel_mySR	No Info	15618436	15618603	Confirmed	A47	167	206	Yes	GGTCGTAAGATGATGACGGCCAAA	ATTATCATATGGCCACGACCAGCC
3R	16309164	16313649	ID_50621	4485	Duplication	Pindel_mySR	No	16313528	16309398	Confirmed	Dup23	No band	363	No	GCTCAGCTCATGAAGGACTTT	TGCGATGCTTCTCTGATACAC
3R	16634889	16638826	ID_50833	3937	Duplication	Pindel_mySR	No	16638569	16635201	Confirmed	Dup24	No band	571	No	GGCTTCATTCTCTGCTCGATAC	TGCGCTCCTCTTCACAATAC
3R	16980330	16980507	ID_51030	177	Deletion	Pindel_mySR	Yes	16980141	16980580	Confirmed	A48	439	262	No	AGTCAACTTTCCGACTGCCTCTGA	TATTGCGCACAGGTGGCAAAGTTC
3R	18164644	18164771	ID_51795	127	Deletion	Pindel_mySR	Yes	18164460	18164873	Confirmed	B33	413	286	No	CACACAAACACGCACACACTCTCT	GCACTGTCCTGATTGGCATTCACA
3R	18896411	18896459	ID_52222	48	Deletion	Pindel_mySR	Yes	18896328	18896546	Confirmed	A49	218	170	No	CCGTACATAAACGAATCGTAGTTGG	ACCCAATTTACAGACCTTTGGCAC
3R	18896440	18896488	ID_52223	48	Deletion	Pindel_Delly	No Info	18896328	18896546	Confirmed	A49	218	170	No	CCGTACATAAACGAATCGTAGTTGG	ACCCAATTTACAGACCTTTGGCAC
3R	19758343	19758374	ID_52712	31	Deletion	Pindel_mySR	Yes	19758251	19758425	Confirmed	A50	174	143	No	TAGCCAGACCAACGAAGAGCATGA	GACACAAACACGCCGAATAGCAAG
3R	21129505	21131422	ID_53383	1917	Duplication	Pindel_mySR	Yes	21131163	21129746	Confirmed	Dup25	No band	506	No	TGTGGGCATTGAGGGATAAG	GTCGGAGTAACTCAAGAGAAGAG
3R	24178286	24178573	ID_55306	287	Deletion	mySR_Delly	Yes	24178161	24178718	Confirmed	A51	557	270	No	ACTCTGCTCGGTTCTCGTGGATTT	TGGCAATCTGGAATAGGGCACAGA
3R	25062302	25067776	ID_55800	5474	Deletion	Pindel_mySR	No	25062136	25068011	False Positive	Del9	5875	401	No	CTGTTGACCACATGGATGGA	CCCGGGCACTATGACATAAATA
3R	27677327	27677357	ID_57364	30	Deletion	Pindel_mySR	No	27677265	27677404	Confirmed	A52	139	109	Yes	TGTTTGTGTCTTTCCCACTAAGGG	AGAAGCACTACCCACCCAATCCAA
X	674664	674998	ID_58056	334	Deletion	Pindel_mySR	Yes	674441	675246	Confirmed	B7	805	471	No	TCGTCATCGTCATCAAAGTCGGCA	TGCAGTTTGCCTACACCGGCTAAT
X	1865884	1866119	ID_58457	235	Deletion	Pindel_mySR	Yes	1865710	1866224	Confirmed	B1	514	279	No	AACAACTCTTCGACTCTTCGCCCT	ACGATGCCAGCGATGTTTGAACTG
X	5647023	5650073	ID_60055	3050	Duplication	Pindel_mySR	Yes	5649888	5647208	Confirmed	Dup26	No band	372	No	GCAGCTCGTACATTTCGTTTC	TCATGTTGCTGGCGAGTT
X	5725925	5727365	ID_60098	1440	Duplication	Pindel_mySR	No	5727091	5726174	False Positive	Dup27	No band	570	No	GTTAATGAAATGCCTGCCACTC	TTCGAATCCCGGTGAAGTTATC
X	7689128	7689298	ID_60883	170	Deletion	Pindel_mySR	Yes	7688991	7689339	Confirmed	A55	348	178	No	ATCTGTATCTGCGGCTGTGTGTGA	GTTTCTCGTGCCTTGTCATGGCTT
X	8671819	8671819	ID_61362	33	Insertion	Pindel_mySR	No Info	8671723	8671866	Confirmed	A57	143	176	Yes	GGATAGATGGATGGGTTGAGTACG	GCTTGTTGATTGAAGTGAAGGAGCG
X	13689566	13693814	ID_63785	4248	Deletion	Pindel_mySR	No	13689038	13694420	False Positive	Del12	5382	1134	No	TGCAGCCAGCAGACATAAA	GTGCCTCGACTAGAACTGAAATC
X	14130351	14133714	ID_63969	3363	Duplication	Pindel_Delly	Yes	14133565	14130565	Confirmed	Dup29	No band	365	No	CATTGCCATAACGGACGAGTA	CTGAAGAAGACCACGATGCTAA
X	20191355	20191403	ID_66792	48	Deletion	Pindel_Delly	No	20191265	20191461	Confirmed	A54	196	148	Yes	TTTGGTTGCGCATTCAGTCGTGTC	CCAGAGTAGAAGCACATGATACGC

Supp. Table 1. Description of the CNVs empirically evaluated by PCR and Sanger sequencing. The table details the CNVs tested, the sequences of the primers used, the expected and observed PCR results and whether additional Sanger sequencing was performed.

Supp. Table 2. Coordinates and characteristics of the CNVs identified in this study. Columns 1-3 have the CNV coordinates; 'CNV_ID' is a unique identifier for each CNV; 'size' refers to the size of the CNV in base pairs; 'type' to whether it is an insertion, deletion or duplication; 'pipeline' identity of the 2 pipelines that support the call (no distinction made between calls supported by the 3 pipelines or only by Pindel and the in-house pipeline; 'ntlen' number of additional nucleotides inserted/deleted at the breakpoint (0 if none is observed), the information is only available for calls made by Pindel; 'homely' length of the stretch of microhomology present at the breakpoint (0 if none is observed), the information is only available for calls made by Pindel; 'Annotation.r5.52' refers to the genomic region the CNV overlaps, 'Intergenic'/'Intronic' means the whole sequence is non-coding, 'CodingExon' means that at least 1 bp of a coding exon is included in the CNV, '3Exon'/'5Exon' means that at least 1 bp of an UTR exon is included in the CNV (and 0 bp for a coding exon); 'CompGeneDups' 1 if the CNV encompasses a complete gene, by combining this column with the previous it is possible to distinguish protein-coding genes (i.e. marked as 'CodingExon') from non-coding genes which appear as 1 in the 'CompGeneDups' column but as something other than 'CodingExon' in the 'Annotation.r5.52' column; 'Chimera1' 1 if the CNV forms a new chimeric gene structure between genes in the same strand; 'Chimera0' 1 if the CNV forms a new chimeric gene structure between genes in the opposite strand; 'CompGeneDels' 1 if at least one gene is completely deleted, as for 'CompGeneDups' by combining this column with the 'Annotation.r5.52' column it is possible to distinguish protein-coding genes (i.e. marked as 'CodingExon') from non-coding genes (i.e. marked as something other than 'CodingExon'); 'Fusion1' 1 if the CNV forms a new fusion gene between genes in the same strand; 'Fusion0' 1 if the CNV forms a new fusion gene between genes in the opposite strand; 'Freq84' frequency of the CNV in the whole dataset; 'FreqB' frequency of the CNV in Beijing; 'FreqI' frequency of the CNV in Ithaca; 'FreqN' frequency of the CNV in the Netherlands; 'FreqT' frequency of the CNV in Tasmania; 'FreqZ' frequency of the CNV in Zimbabwe; the next 84 columns refer to the presence (1) or absence (0) of each CNV in the 84 lines; 'CoverageSupport' refers to whether or not the call is further supported by read depth; the final 10 columns list for each CNV the 10 possible pairwise Fst comparisons between the five populations.

This table is provided as a flat text file

		Observed number of events								% of observed events						
		Coding Exon	5' UTR Exon	3' UTR Exon	Intronic	Intergenic	Total	Coding Exon except if complete gene duplication / deletion**	Complete gene duplications / deletions**	Coding Exon	5' UTR Exon	3' UTR Exon	Intronic	Intergenic	Coding Exon except if complete gene duplication / deletion**	Complete gene duplications / deletions**
Empirical data	Duplications	948	96	95	609	473	2221	643	305	43%	4%	4%	27%	21%	29%	14%
	Deletions	2227	1180	2403	30248	20504	56562	2159	68	4%	2%	4%	53%	36%	4%	0.1%
	Insertions	65	68	173	2118	1426	3850	65	0	2%	2%	4%	55%	37%	2%	0.0%
Simulated data	Duplications - Controls	91834	10360	6160	42820	54723	205897	80932	10902	45%	5%	3%	21%	27%	39%	5%
	Deletions - Controls	1279784	205255	276724	2106800	1471671	5340234	1274002	5782	24%	4%	5%	39%	28%	24%	0.1%
	Insertions - Controls	73342	12916	20131	155231	103249	364869	73342	0	20%	4%	6%	43%	28%	20%	0.0%

** These classes are a subset of the 'CodingExon' calls

Supp. Table 3. Purifying selection eliminates a significant fraction of CNVs that overlap protein-coding genes. Numbers of CNVs overlapping different genomic contexts vs. the expectation based on shuffling the CNV coordinates (but keeping the size) within chromosomal arms.

Parental Gene name	Parental Gene FB id	Number coding introns deleted	Number UTR introns deleted	Extra deletions from lower quality calls?	Complete*?	Freq 84	B	I	N	T	Z	chr parental gene	Expression in testis**?	Expression in ovaries**?
Adf1	FBgn0000054	2	1	In-house piepline (1) + Hydra (1)	Complete	1	0	0	1	0	0	2R	moderate	moderately high
Pcl	FBgn0003044	1	0	No	Partial	1	1	0	0	0	0	2R	moderate	high
Adam	FBgn0027619	4	0	In-house piepline (1)	Complete	1	0	0	0	1	0	2R	high	very high
l(1)G0320	FBgn0028327	1	0	No	Complete	1	0	0	0	0	1	X	high	high
Kr-h2	FBgn0028419	2	1	In-house piepline (1)	Complete	1	0	1	0	0	0	2L	high	high
CG32113	FBgn0052113	3	0	Pindel (1)	Partial	1	1	0	0	0	0	3L	low	moderate
CG3631	FBgn0038268	3	0	No	Complete	1	0	1	0	0	0	3R	low	moderate
CG9914	FBgn0030737	1	0	No	Complete	2	0	2	0	0	0	X	high	moderate
CG33969	FBgn0053969	1	0	No	Partial	2	0	0	0	2	0	3L	moderate	moderately high
CG32082	FBgn0052082	5	0	In-house piepline (2)	Partial	3	0	1	0	1	1	3L	very low	very low
CLIP-190	FBgn0020503	1	0	No	Partial	5	1	0	2	2	0	2L	moderately high	moderately high
Bsg	FBgn0261822	4	0	Hydra (2)	Complete	5	2	3	0	0	0	2L	low	moderately high
eIF-4E	FBgn0015218	4	1	No	Complete	8	0	0	0	8	0	3L	very high	very high
Cf2 + Pen	FBgn0000286 FBgn0267727	2	0	No	Partial	8	1	1	1	3	2	2L	moderate	moderately high
cp309	FBgn0086690	1	0	No	Partial	12	1	5	0	5	1	3L	very low	moderate
SMC2	FBgn0027783	1	0	No	Partial	14	4	4	1	4	1	2R	moderate	high
c(3)G	FBgn0000246	3	0	In-house piepline (1)	Partial	22	2	7	3	5	5	3R	low	moderately high

* Complete only refers to coding exons

** Expression retrieved from modENCODE (through Flybase)

Supp. Table 4. Description of the polymorphic retrogenes identified in this study. The table includes information on the structure of retrogenes, their frequency in the five populations and the expression profiles of the parental genes in the germlines.

Supp. Table 5. List of genes with independent complete gene duplications. The column ‘Annotation.r5.52’ refers to the genomic region the CNV overlaps, when not ‘CodingExon’ means the gene duplicated is not a protein-coding gene.

This table is provided as a flat text file

Supp. Table 6. Expression values of the genes represented in the microarray that are duplicated in the GDL. Each column represents the median expression across replicates for each line.

This table is provided as a flat text file