Supplemental Figures:

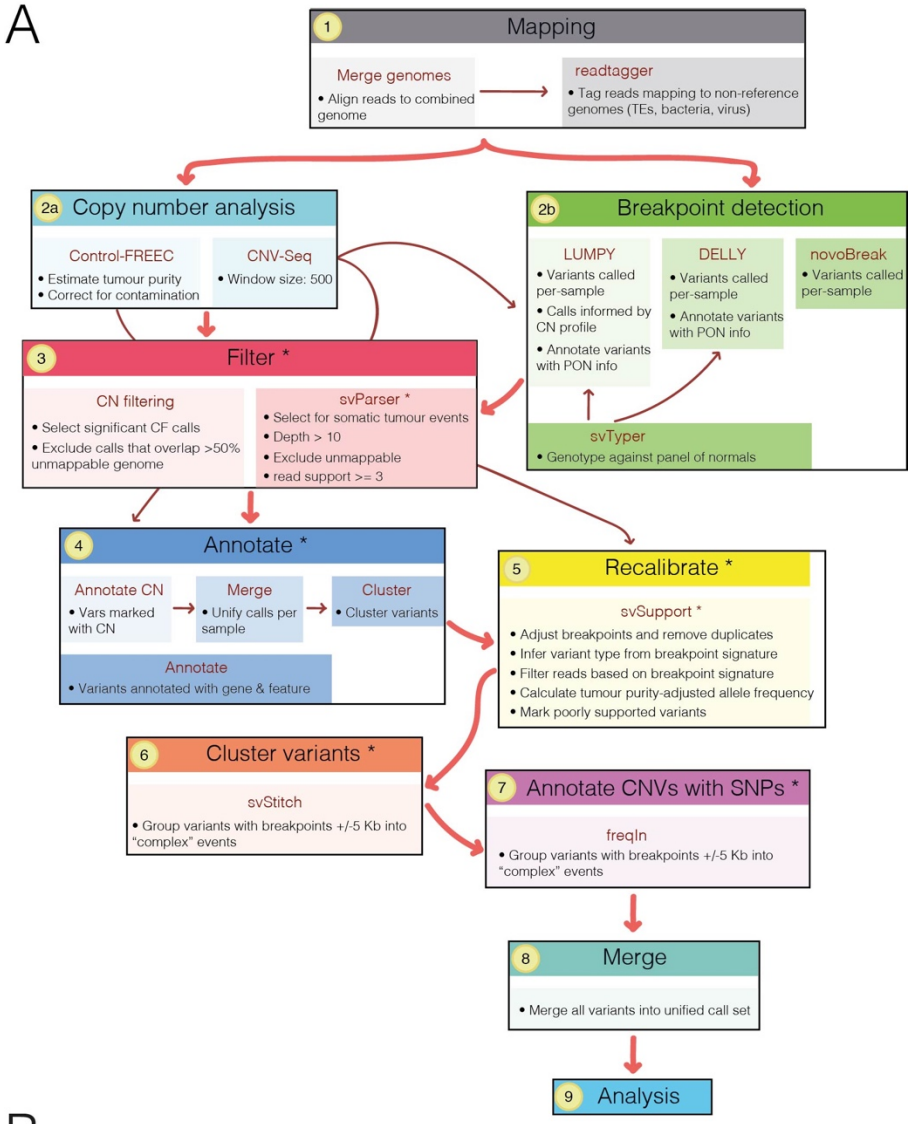
Evolution and genomic signatures of spontaneous somatic mutation in *Drosophila* intestinal stem cells

Nick Riddiford[1], Katarzyna Siudeja[1], Marius van den Beek[1], Benjamin Boumard[1], Allison J. Bardin[1] *

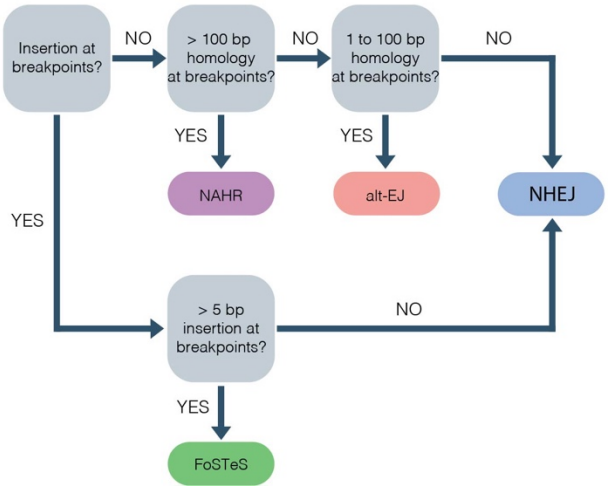
[1] Institut Curie, PSL Research University, CNRS UMR 3215, INSERM U934, Stem Cells and Tissue Homeostasis Group, Paris, France.
* author for correspondence

Fig. S1

A


1 Mapping
Merge genomes
• Align reads to combined genome
readtagger
• Tag reads mapping to non-reference genomes (TEs, bacteria, virus)
2a Copy number analysis
Control-FREEC
• Estimate tumour purity
• Correct for contamination
CNV-Seq
• Window size: 500
2b Breakpoint detection
LUMPY
• Variants called per-sample
• Calls informed by CN profile
• Annotate variants with PON info
DELLY
• Variants called per-sample
• Annotate variants with PON info
novoBreak
• Variants called per-sample
svTyper
• Genotype against panel of normals
3 Filter *
CN filtering
• Select significant CF calls
• Exclude calls that overlap >50% unmappable genome
svParser *
• Select for somatic tumour events
• Depth > 10
• Exclude unmappable
• read support >= 3
4 Annotate *
Annotate CN
• Vars marked with CN
Merge
• Unify calls per sample
Cluster
• Cluster variants
Annotate
• Variants annotated with gene & feature
5 Recalibrate *
svSupport *
• Adjust breakpoints and remove duplicates
• Infer variant type from breakpoint signature
• Filter reads based on breakpoint signature
• Calculate tumour purity-adjusted allele frequency
• Mark poorly supported variants
6 Cluster variants *
svStitch
• Group variants with breakpoints +/-5 Kb into "complex" events
7 Annotate CNVs with SNPs *
freqIn
• Group variants with breakpoints +/-5 Kb into "complex" events
8 Merge
• Merge all variants into unified call set
9 Analysis

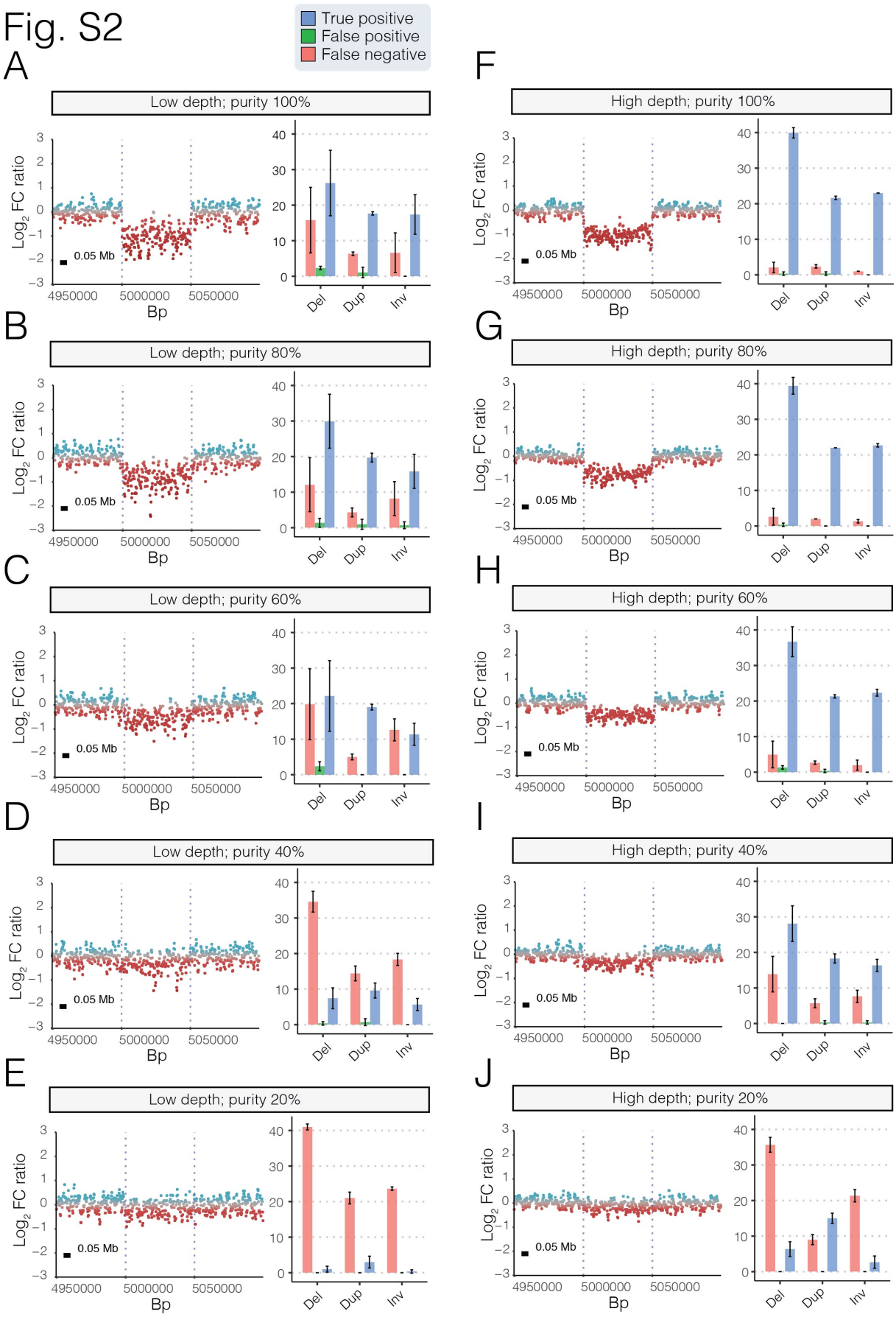
B


Insertion at breakpoints?
NO
> 100 bp homology at breakpoints?
NO
1 to 100 bp homology at breakpoints?
NO
YES
YES
NAHR
alt-EJ
NHEJ
YES
> 5 bp insertion at breakpoints?
NO
YES
FoSTeS

Supplemental Fig. S1. Schematic showing a bioinformatic pipeline for identifying structural variants

(**A**) Reads were mapped to *Drosophila* transposable element sequences as well as multiple genomes found as contaminants in our sequencing data (Methods; Supplemental Methods). Structural variant discovery was then performed in two complementary but distinct steps. First, read-depth-based approaches (CNV-Seq (Xie and Tammi 2009); Control-FREEC (Boeva et al. 2012)) were used to detect copy number variants (CNVs). Secondly, we used read mapping-based approaches, utilising aberrantly mapped and/or split-reads (LUMPY (Layer et al. 2014); DELLY (Rausch et al. 2012); novobreak (Chong et al. 2017)) to detect precise breakpoints of multiple classes of structural variant. To ensure only true somatic events were considered, a panel of normals (PON) was constructed by combining variant calls from all sequenced normal samples, and used to select for somatic (tumour-only) calls. We then combined variants into unified per-sample call sets, before merging variants that were found by multiple approaches, and then annotated calls with the location of the breakpoints with respect to gene features (Methods). Finally, breakpoints within close proximity (+/- 5 kb) were clustered into "complex" events, and CNVs were annotated with the frequency of heterozygous SNPs, in order to discern false positives (Methods; Supplemental Methods). (**B**) Assigning putative underlying mechanisms of structural variants by analysis of breakpoint junctions. We annotated breakpoints for microhomology and inserted sequences, and classified each breakpoint junction using criteria adapted from (Yang et al. 2013; Kidd et al. 2010).

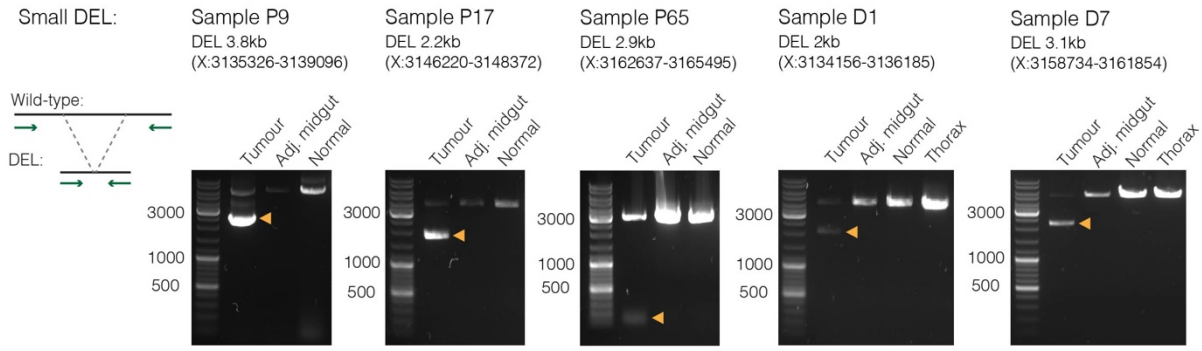
Fig. S2


True positive
False positive
False negative
A
Low depth; purity 100%
B
Low depth; purity 80%
C
Low depth; purity 60%
D
Low depth; purity 40%
E
Low depth; purity 20%
F
High depth; purity 100%
G
High depth; purity 80%
H
High depth; purity 60%
I
High depth; purity 40%
J
High depth; purity 20%
Log2 FC ratio
0.05 Mb
4950000
5000000
5050000
Bp
Del
Dup
Inv

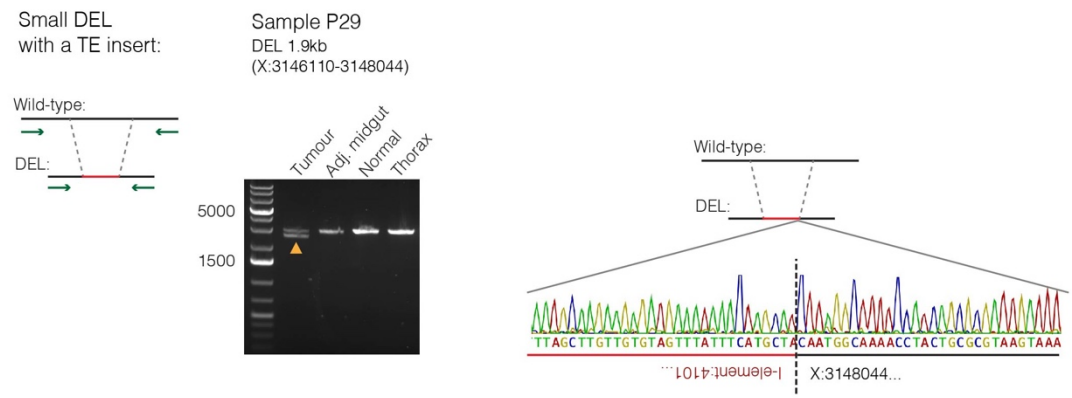
Supplemental Fig. S2. Error rates of simulated structural variants
We simulated 90 structural variants (see Supplemental Tables S2 and S3), distributed the mappable genome in two sequencing depth conditions representing a 'low' average depth (tumour: 10x, normal 30x; **A-E**) and a 'high' average depth (tumour: 30x, normal: 50x; **F-J**). For both conditions, reads were simulated at five different levels of normal-in-tumour contamination, representing tumour purity values of 100% (**A**, **F**), 80% (**B**, **G**), 60% (**C**, **H**), 40% (**D**, **I**) and 20% (**E**, **J**). Each panel shows a read-depth ratio plot over the same 50 kb simulated deletion on the X Chromosome, as well as the error rates detected in that sequencing condition. Bar plots show the number of variants detected in each error rate category for each class of structural variant simulated. Each point on the read-depth ratio plot represents the Log_2 ratio of read counts in 500 bps windows between the simulated tumour and normal sample, and dotted lines indicate breakpoints. Error bars show the standard deviation of the mean of three replicates.

Fig. S3

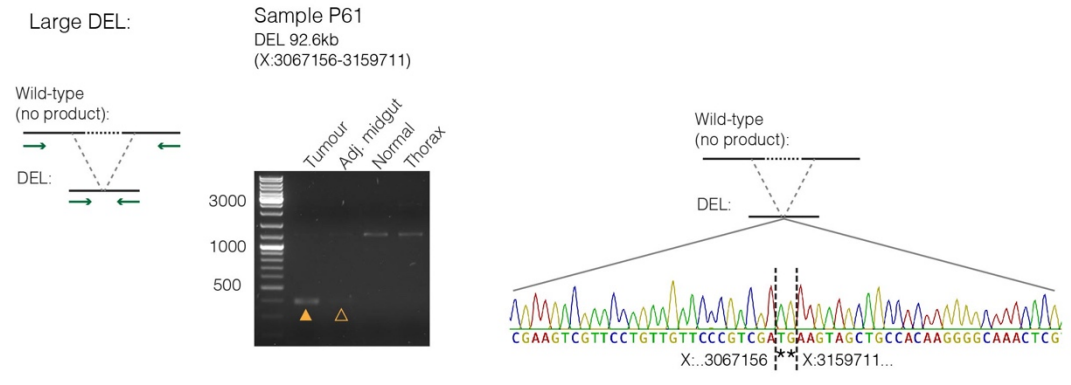
A


Small DEL:
Sample P9
DEL 3.8kb
(X:3135326-3139096)
Sample P17
DEL 2.2kb
(X:3146220-3148372)
Sample P65
DEL 2.9kb
(X:3162637-3165495)
Sample D1
DEL 2kb
(X:3134156-3136185)
Sample D7
DEL 3.1kb
(X:3158734-3161854)
Wild-type:
DEL:
Tumour
Adj. midgut
Normal
Thorax
3000
1000
500

B


Small DEL
with a TE insert:
Sample P29
DEL 1.9kb
(X:3146110-3148044)
Wild-type:
DEL:
Tumour
Adj. midgut
Normal
Thorax
5000
1500
TTAGCTTGTTGTGTAGTTTATTTCATGCTACAATGGCAAAACCTACTGCGCGTAAGTAAA
X:3148044...

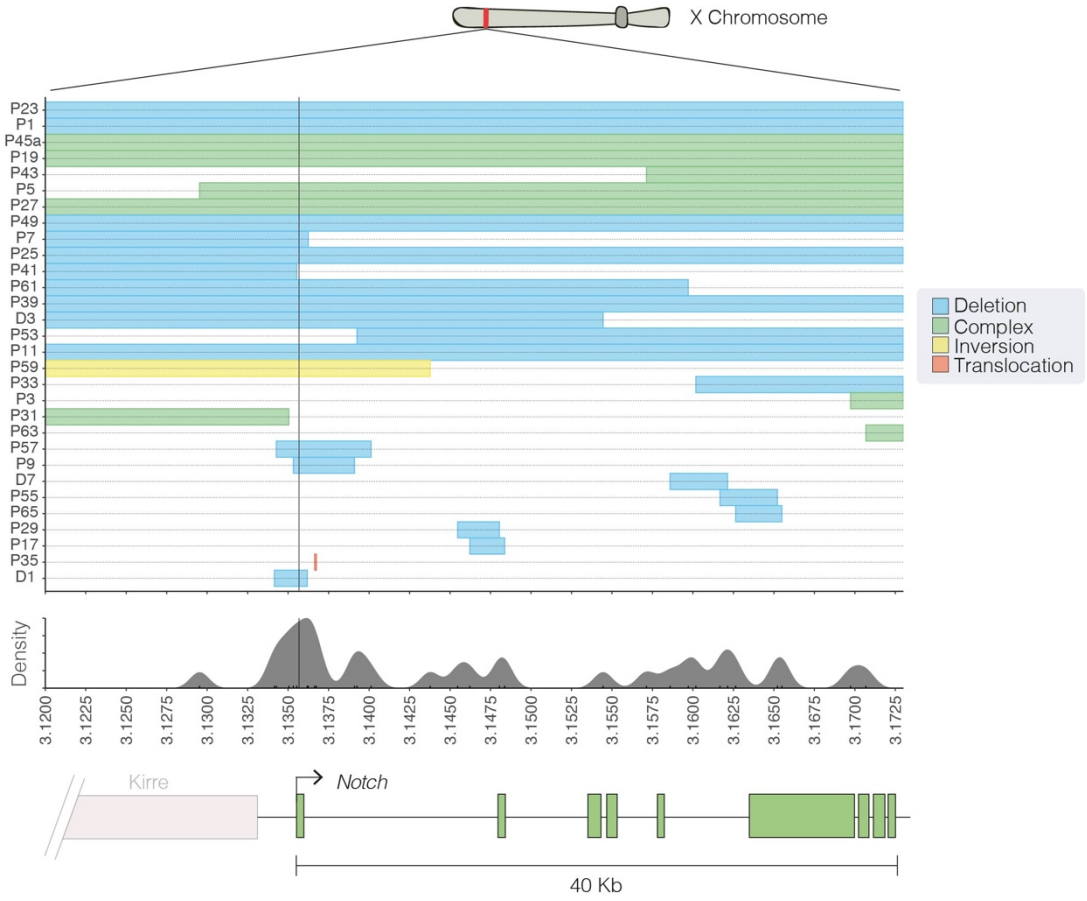
C


Large DEL:
Sample P61
DEL 92.6kb
(X:3067156-3159711)
Wild-type
(no product):
DEL:
Tumour
Adj. midgut
Normal
Thorax
3000
1000
500
CGAAGTCGTTCCTGTTGTTCCCGTCGATGAAGTAGCTGCCACAAGGGGCAAACTCG
X:...3067156
X:3159711...

Supplemental Fig. S3 Validation of *Notch* structural variants

Validation was performed using tumour DNA as well as matched control DNA from the adjacent gut, the head (normal) or the thorax of the same fly. Yellow arrowheads indicate tumour specific amplicons. (**A**) We validated five small deletions (2-4 kb) via PCR. Here, we designed primers downstream and upstream of a deletion to amplify either a wild-type *Notch* fragment or a shorter variant containing a deletion. Wild-type bands were detectable in all samples, whereas short deletion bands were present only in tumour samples and not in the controls. (**B**) Sample P29 contained a short deletion with a transposable element insert in the breakpoint. The 3' breakpoint was sequenced confirming the insertion of an internal fragment (at 4101 bps) of a *Drosophila* I-element in a reverse orientation. Dashed vertical line indicates the breakpoint, and the coordinates of each breakpoint are given below. (**C**) A large deletion (92.6 kb) from sample P61 was validated with primers amplifying the newly formed breakpoint in the tumour sample. We also detected a tumour-specific band in the adjacent gut control, suggesting contamination with tumour cells upon manual dissection. However, this amplicon was not present in head (normal) and thorax controls. The amplicon was sequenced (chromatogram). Dashed lines indicate putative breakpoints, which cannot be unambiguously assigned because of the 2 base-pair microhomology (indicated with stars).

Fig. S4

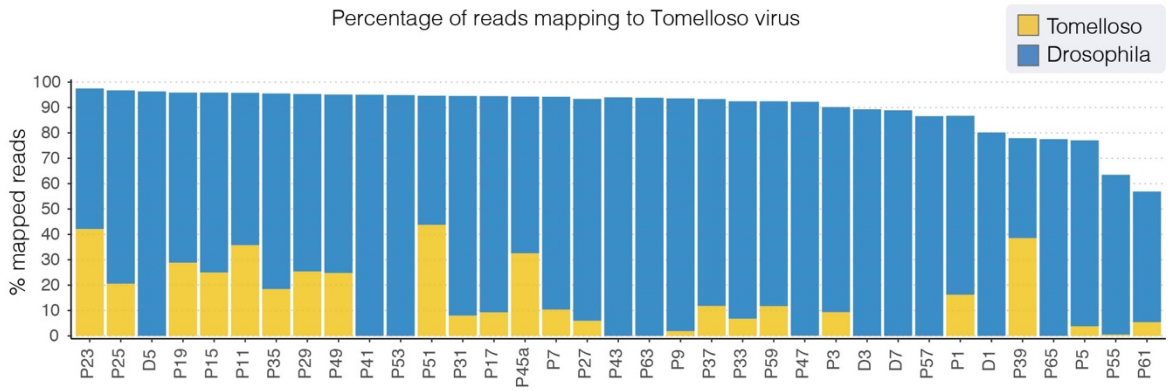


Supplemental Fig. S4. The distribution of structural variant breakpoints over the *Notch* locus

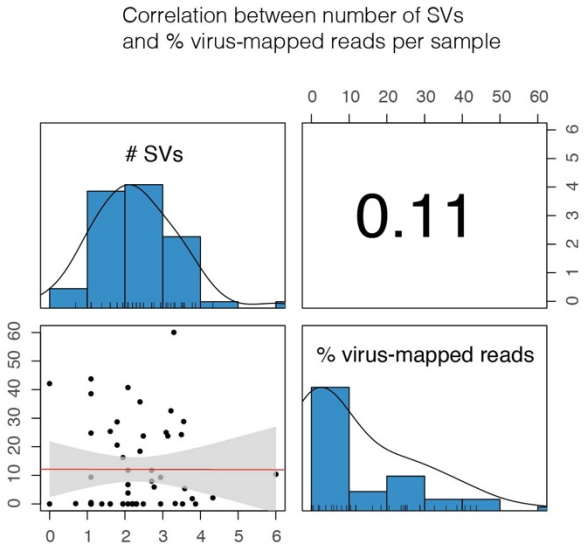
A close up of the region shown in Fig. 2B. Breakpoints over *Notch* did not occur at the same genomic locus, but clustering of breakpoints was observed around the TSS of *Notch* (black vertical line).

Fig. S5

A

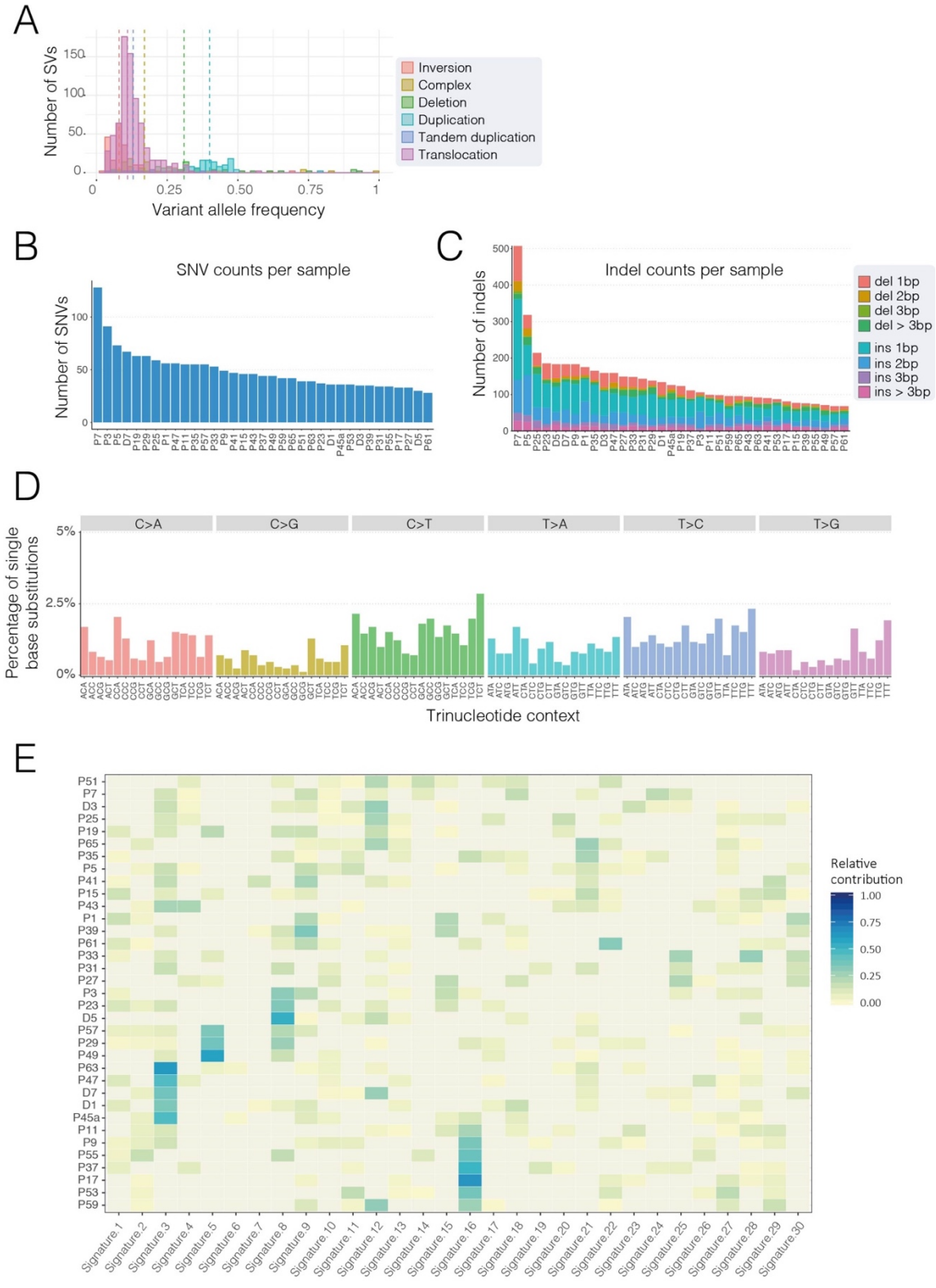


B



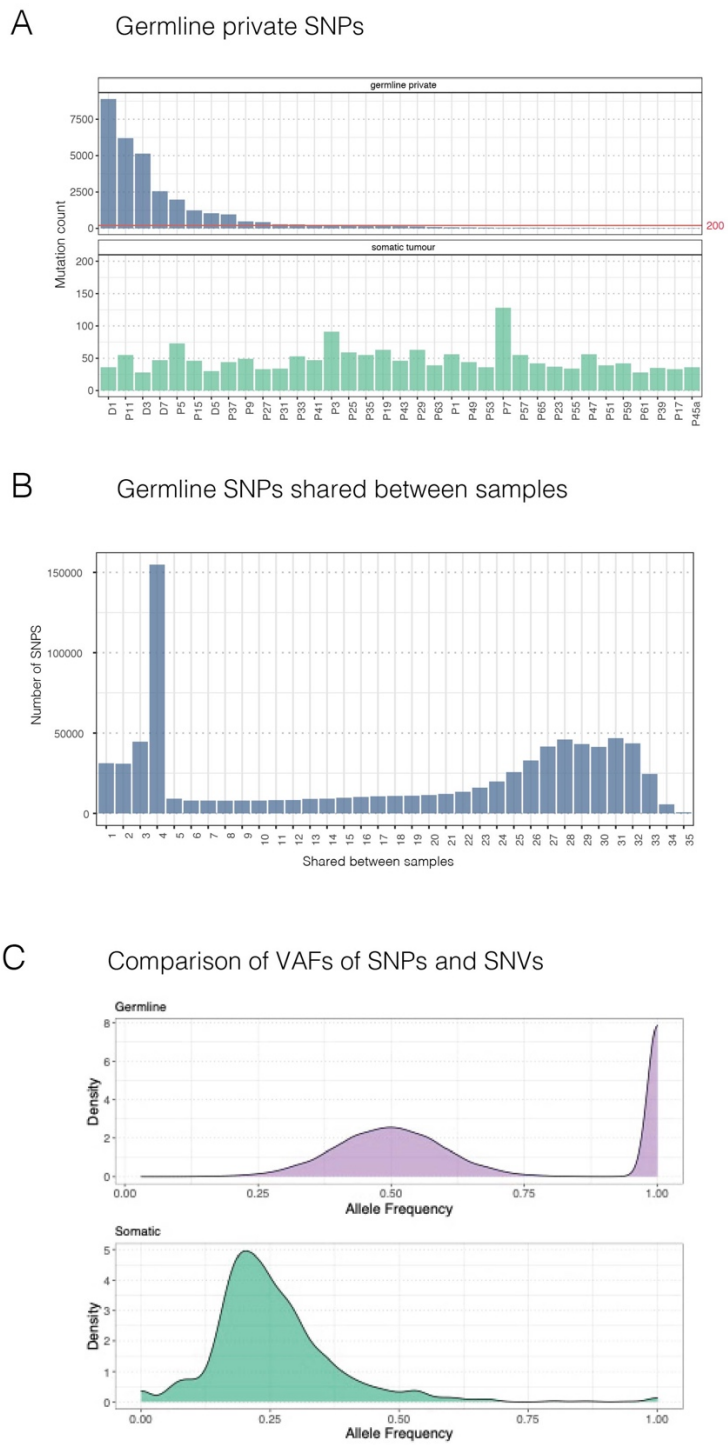
Supplemental Fig. S5. Viral contamination is not correlated with mutation count
(**A**) Percentage of reads mapping to the virus *Tomelloso* and *Drosophila* genome per tumour sample. (**B**) Pearson's correlation between the number of mutations detected per sample and the percentage of reads mapping to the *Tomelloso* genome.

Fig. S6



Supplemental Fig. S6. Genome-wide mutations
(**A**) The distribution of variant allele frequency of structural variants observed genome-wide, coloured by class. Number of SNVs (**B**) and indels (**C**) detected genome-wide. (**D**) Distribution of SNVs within a trinucleotide context across samples. (**E**) Per-sample cosine similarity (relative contribution) between observed mutational spectra and COSMIC mutational signatures.

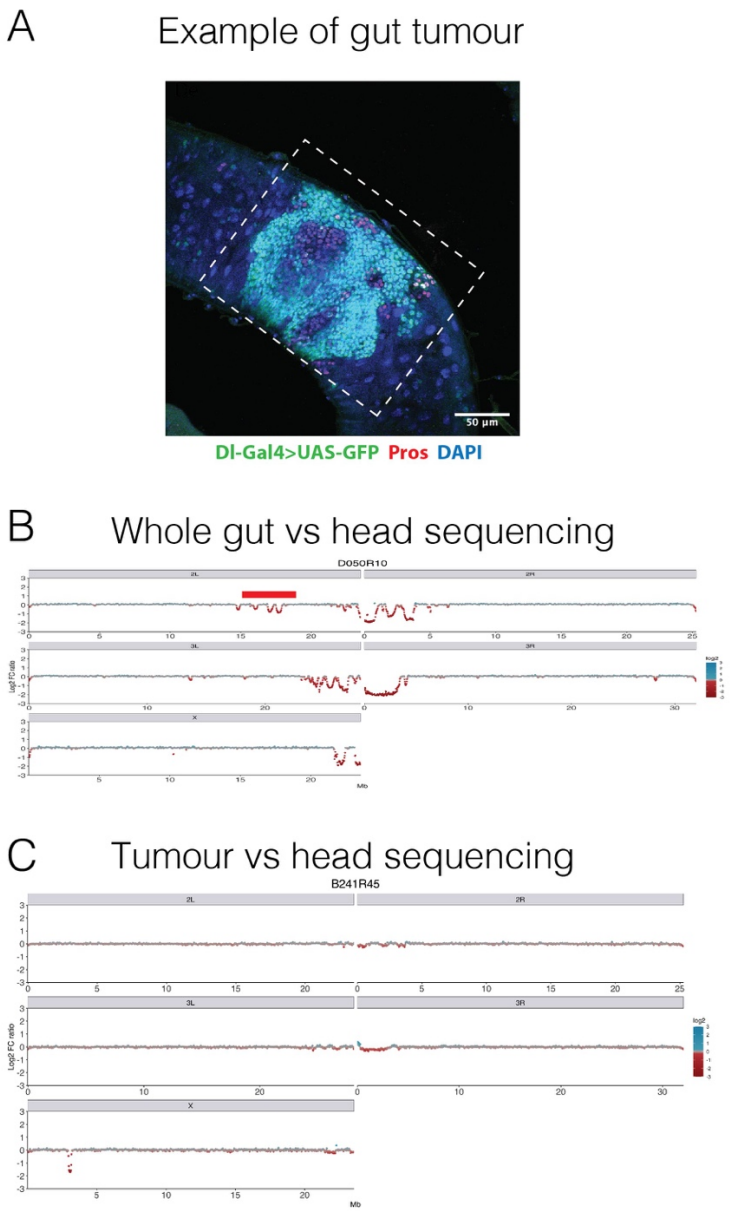
Fig. S7



Supplemental Fig. S7. Comparison of SNPs between samples and VAFS of SNPs and SNVs

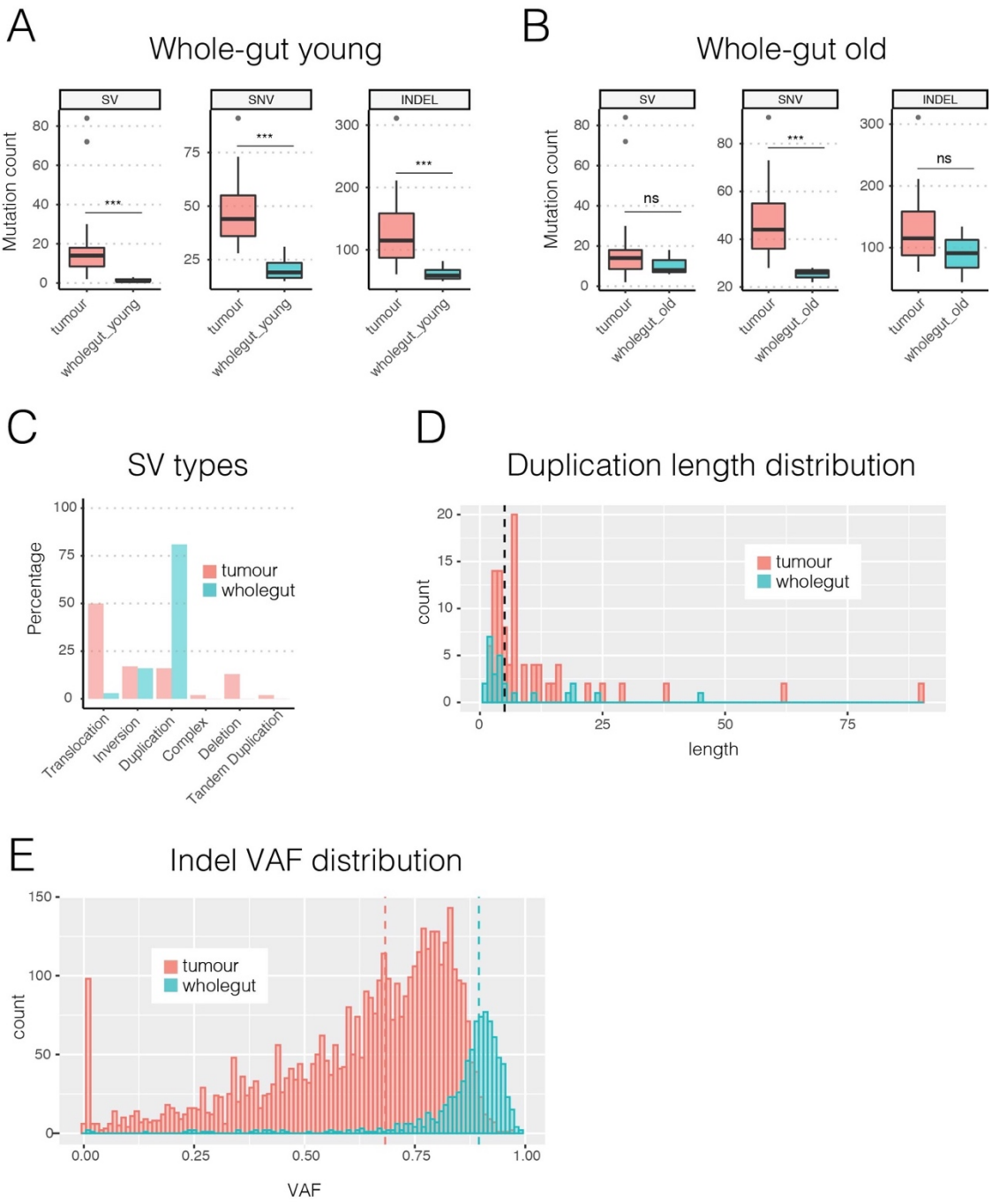
(**A**) Germline SNPs found exclusively in each sample compared to the number of SNVs detected per sample. 2/3 of the samples have fewer than 200 exclusive SNPs (red line in top plot). (**B**) Plot of number of samples sharing germline SNPs. (**C**) Top: Density plot of allele frequency of ~ 50,000 germline SNPs of from one of our sample pairs. Bottom: plot of combined allele frequencies of our called somatic SNVs. Note the much lower variant allele frequency of the SNVs compared to heterozygous SNPs, which center at 0.5, arguing against the called SNVs being contaminated by pre-existing germline SNPs.

Fig. S8



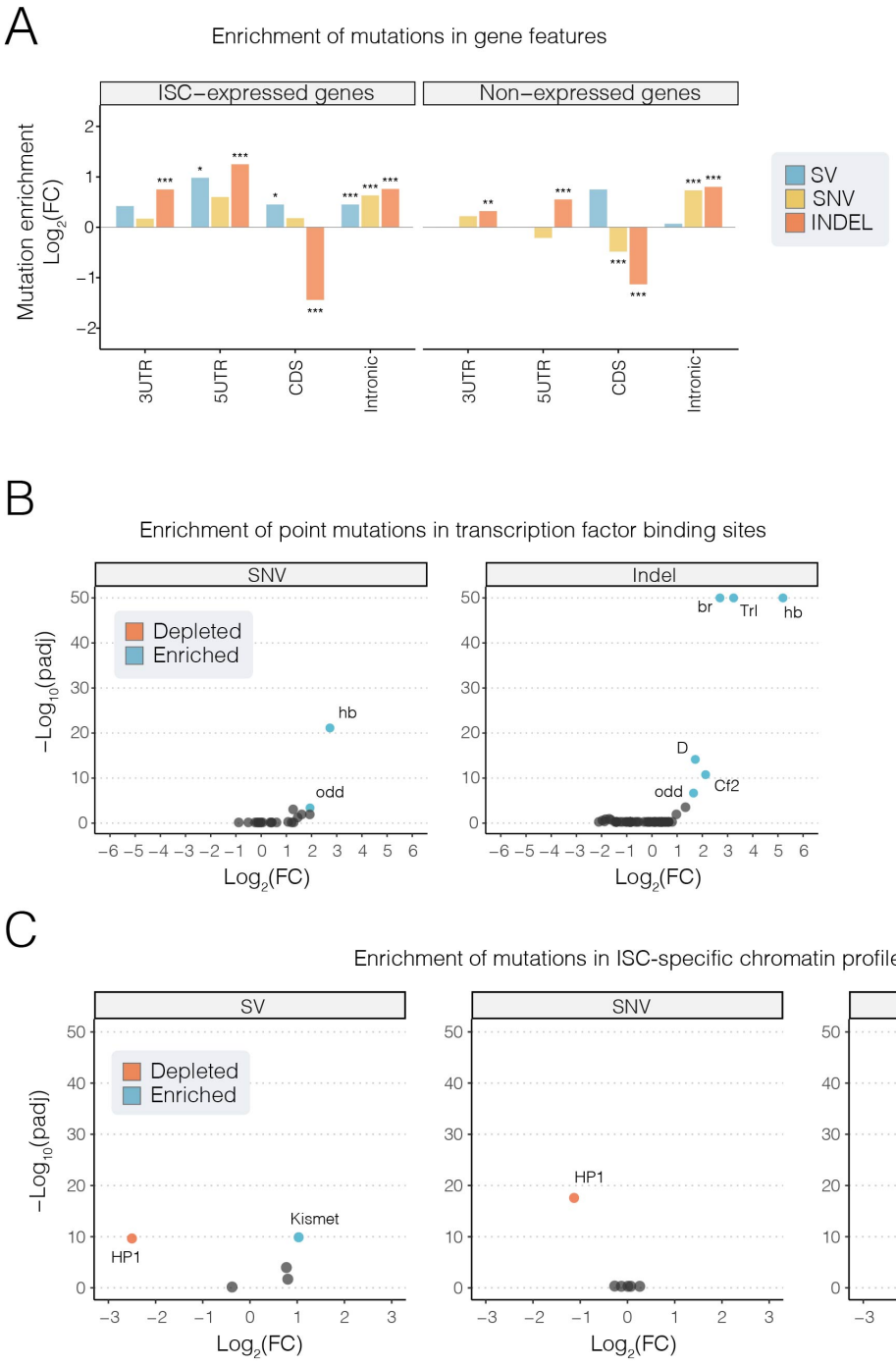
Supplemental Fig. S8. Differences in sequencing of polyploid whole-gut and diploid tumours.
(**A**) A representative example of a tumour dissected for sequencing, dissected area indicated in white box. A small amount of adjacent polyploid ECs are dissected with the diploid tumour reducing tumour purity. **(B)** An example of log2 fold change in coverage when whole gut genomic DNA is compared to head genomic DNA. Notice the drops in sequencing coverage due to the regions within polyploid cells that are not fully replicated during endoreplication. Red bar indicates underendoreplicated midgut regions of Ch 2 L, highlighted in Fig 4 of (Spradling, 2017). **(C)** A representative example of the log2 fold change between a dissected tumour (A573R29) and its corresponding head sample. Notice the diploid nature characterised by the absence of drops in coverage at underendoreplicated regions found in (B).

Fig. S9



Supplemental Fig. S9. Tumour vs whole-gut sequencing of SVs, SNVs, INDELs
Mutation counts compared between tumour and young (1 week) **(A)** and old (6 week) **(B)** whole gut samples. Significantly more SVs, SNVs, and INDELs are detected in the sequenced tumours than in young guts, consistent with our enhanced ability to detect variants due to the clonal expansion of this tissue. Similarly, we could detect significantly more SNVs in tumours than in in old guts. While SVs and INDELs were detected in both tumours and old guts, they differed in SV type and Indel VAF. **(C)** Most of the SVs detected in old whole gut samples were duplications occurring at low allele frequencies, whereas tumours showed more translocations. **(D)**. Number of duplications plotted by length in kb. **(E)** Indels detected in old whole gut samples were low allele frequency events. These data argue against either adjacent ECs or library preparation artefacts causing the observed SNVs, SVs and INDELs in our tumour data.

Fig. S10



Supplemental Fig. S10. Distribution of somatic mutations in genome features
(**A**) Enrichment of mutations in ISC-expressed vs non-expressed genes. Indels were strongly depleted in CDS regions. (**B**, **C**) Volcano plots showing enrichment or depletion of mutations in transcription factor binding sites (**B**) and chromatin landscape (**C**). Highlighted features represent those that with an E-score ($-Log_{10}(p)*Log_2(FC)$; Methods) > 5. The y axis of **B** and **C** are restricted to a maximum $-Log_{10}(padj)$ value of 50 Asterisks denote significance: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.5$. All p values shown have been generated from a two-sided binomial test, and adjusted for multiple comparisons using a Benjamini-Hochberg adjustment.