Supplemental Figures: Population genomics reveals mechanisms and dynamics of *de novo* expressed Open Reading Frame emergence in *Drosophila melanogaster*

Contents

S1 Exon overlap

S1.1 Percentage of overlap of the cumulated sequences of neORF and pre-dating existing exon

In this analysis, we investigated the coverage between de novo expressed ORFs and existing genes, in the category « exon longer ». The first figure represents the percentage of coverage of both genes (de novo and old one). The x axis represents he coverage (percentage of ovevrlap compared to the size of the 2 genes in overlap). The y axis represent the density, which relates to the number of overlap that show this percentge. We observe that in each line, the overlap is small, as most of them relate to less than 20 % of the cumulated overlap

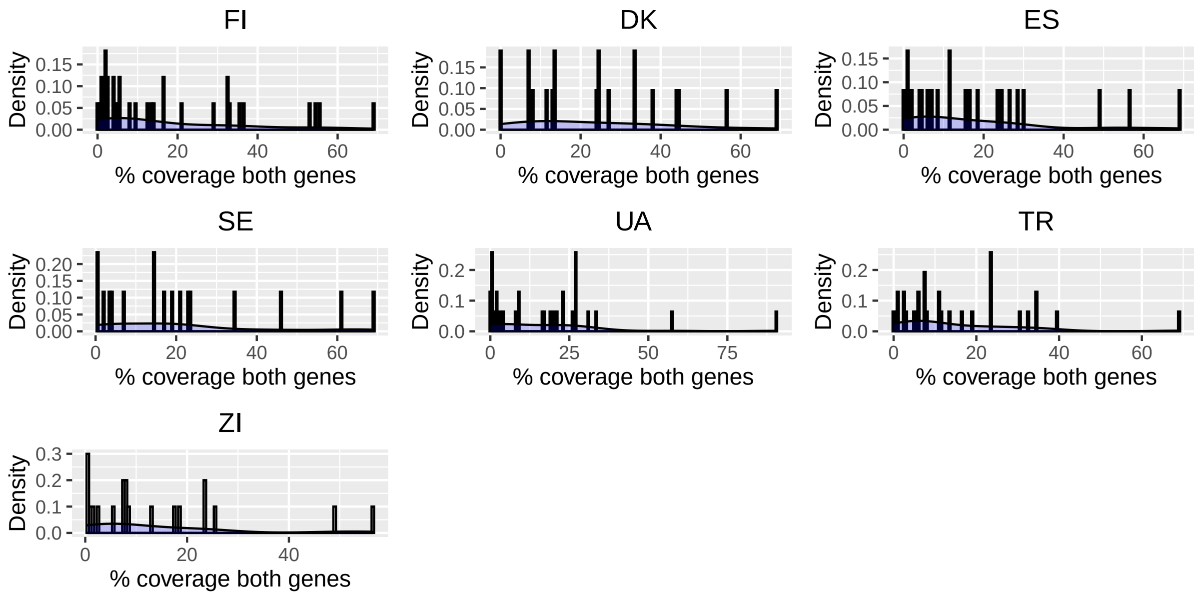

Figure 1: **Percentage of cumulated overlap**

S1.2 Percentage of overlap of the sequence of the neORF to the pre-dating exon

The first figure represents the percentage of neORF overlapping to the pre-existing exon. The x axis represents the coverage. The y axis represent the density. Here also,the overlap is mainly less than 20 of the sequence, except in the Zambian line where it is more homogenous. The second figure represents the percentage of pre-existing gene overlapping with the neORF.

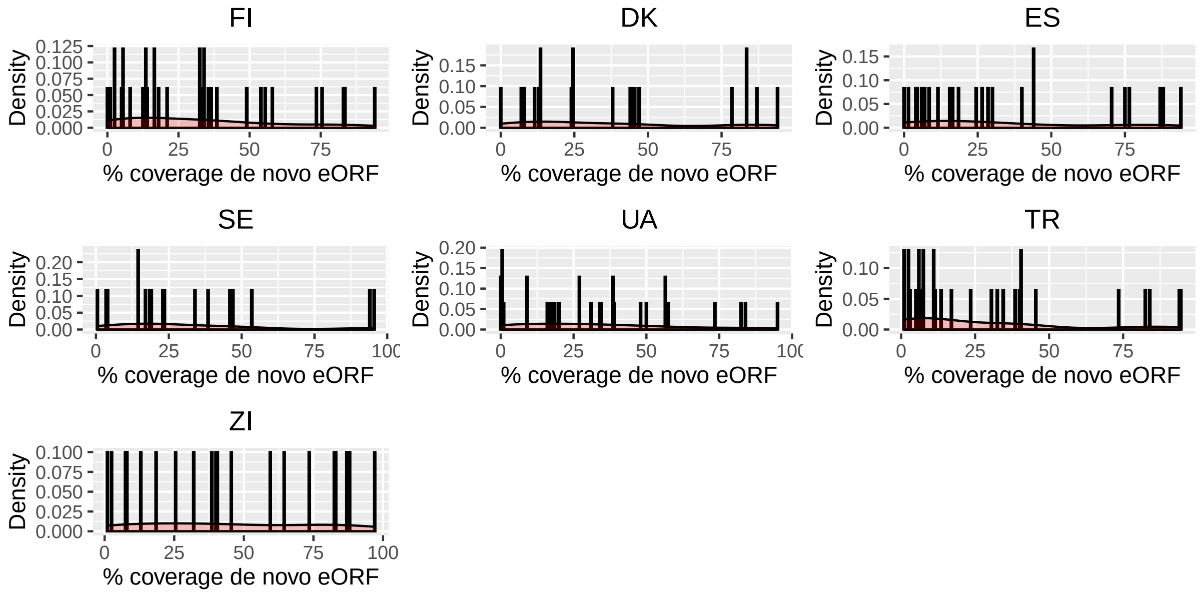

Figure 2: **Percentage of neORF overlap**

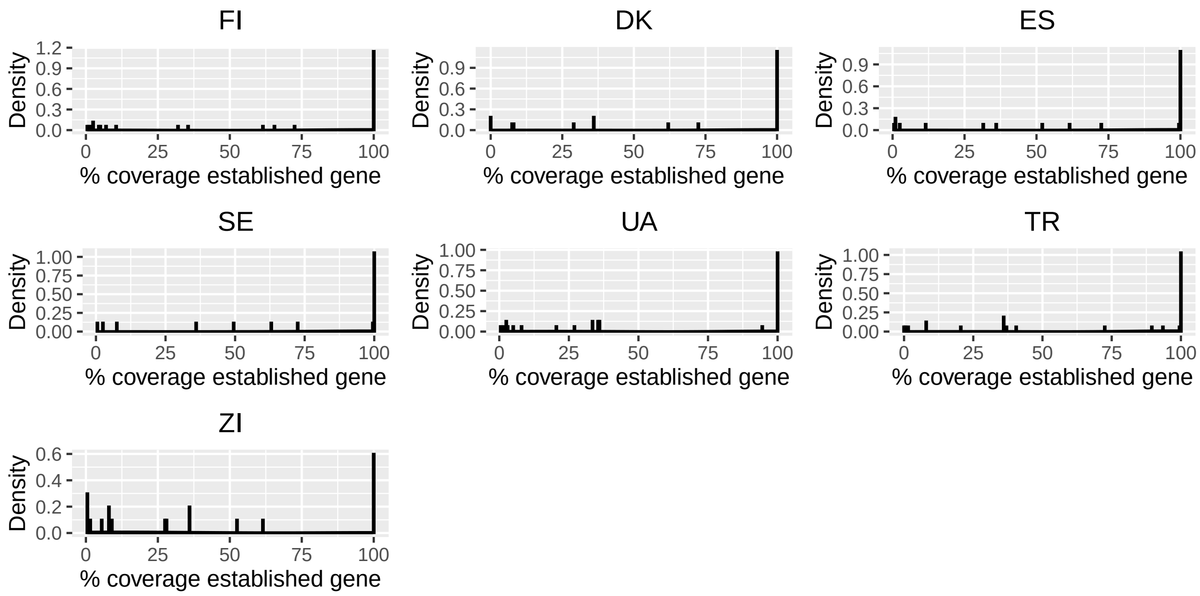

Figure 3: **Percentage of gene overlap**

S1.3 Percentage of overlap of the sequence neORF whose RNA sequence starts before the exon it overlaps, after, or both

In this analysis, we investigated which percentage of tghe transcript containing the neORFs (from ExonLonger) overlapped to an exon upstream (Start before), downstream (End after), or both, meanning that is fully covers the gene.

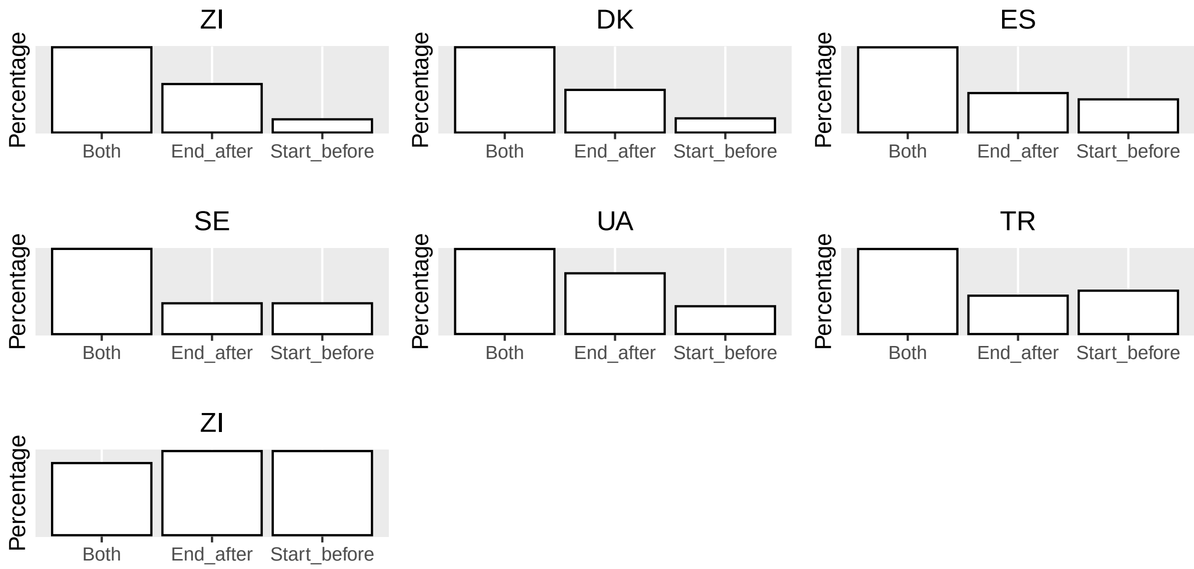

Figure 4: **neORF overlap**

S2 *de novo* genes properties

The figures represent 3 properties of de novo neORFs and *de novo* genes detected in (Heames et al., 2020).

S2.1 Length

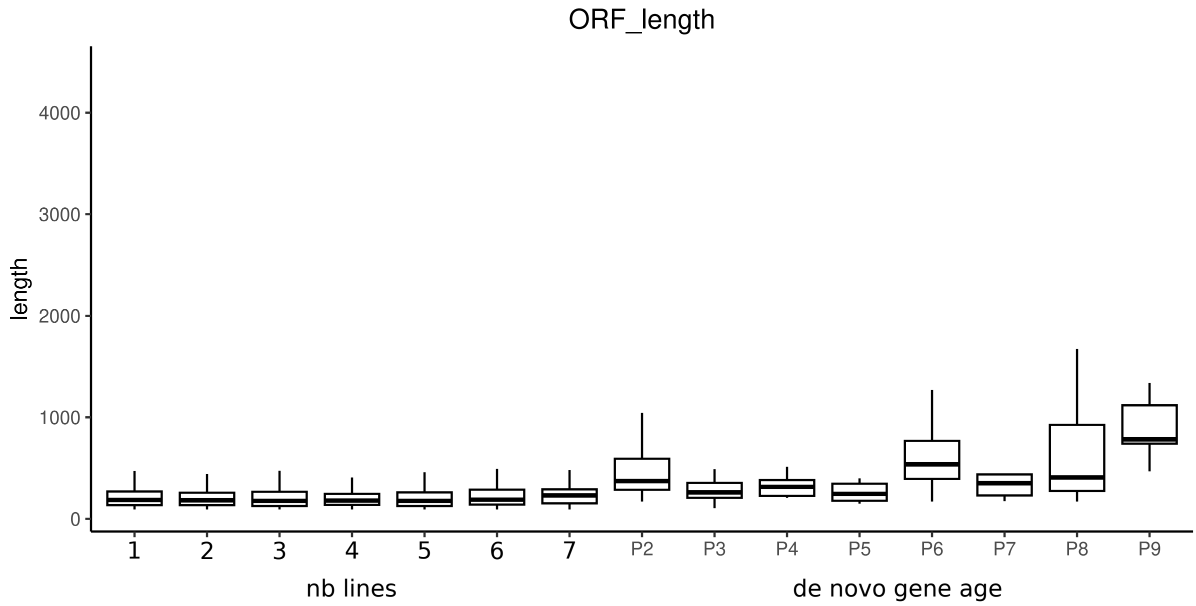

Figure 5: **Length**

S2.2 Aggregation propensity

S2.3 Intrinsic disorder

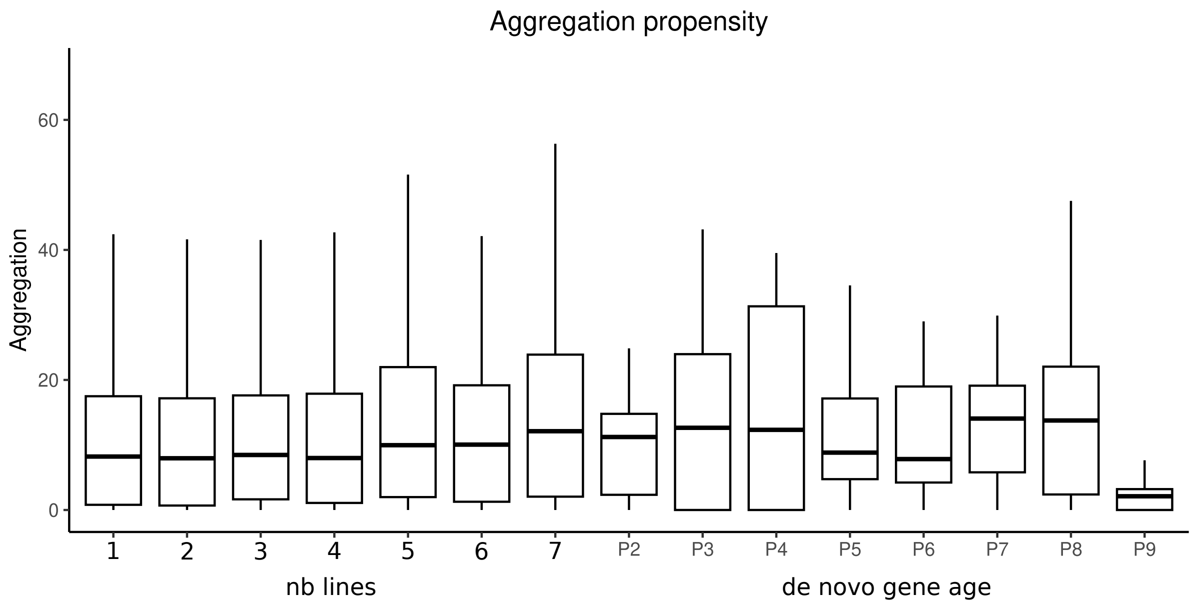

Figure 6: **Aggregation propensity**

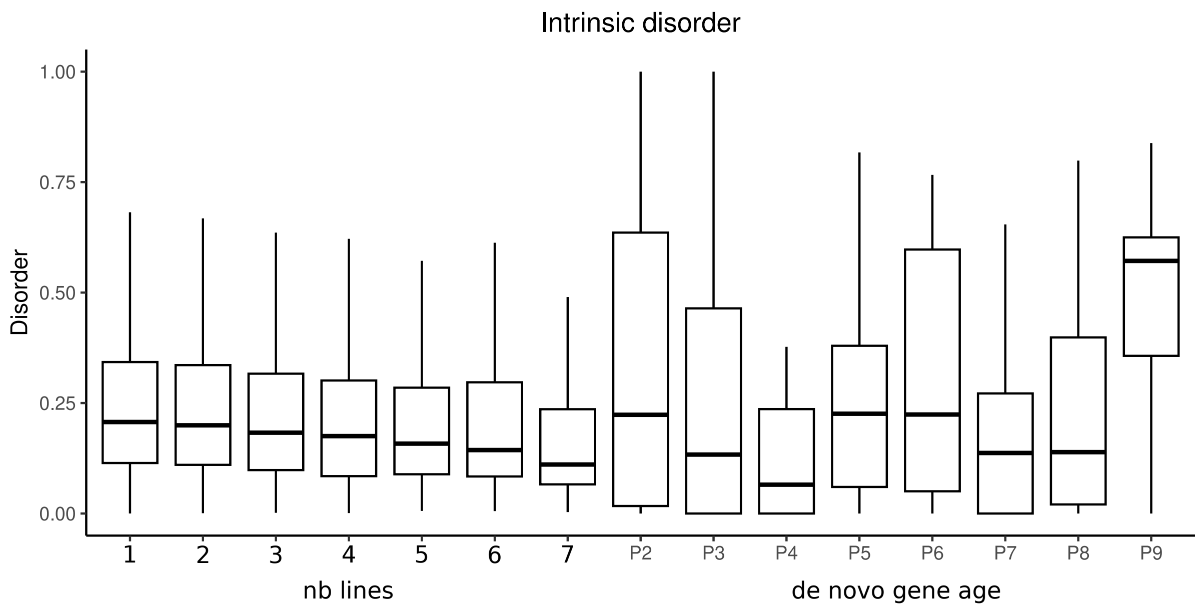

Figure 7: **Intrinsic disorder**

S3 Distance TE - neORFs

In this analysis, we studied neORFs that were not overlapping or inside a transposable elements. All lines taken together, we studied the distance from each neORF to a transposable element. The x axis represents the distance in nucleotide between neORFs and Tes. The y axis represents the density. We observe that neORFs that were not inside of a TE were rather close to one. However, Tes are distributed regularly in the genomes, which might also explain this result.

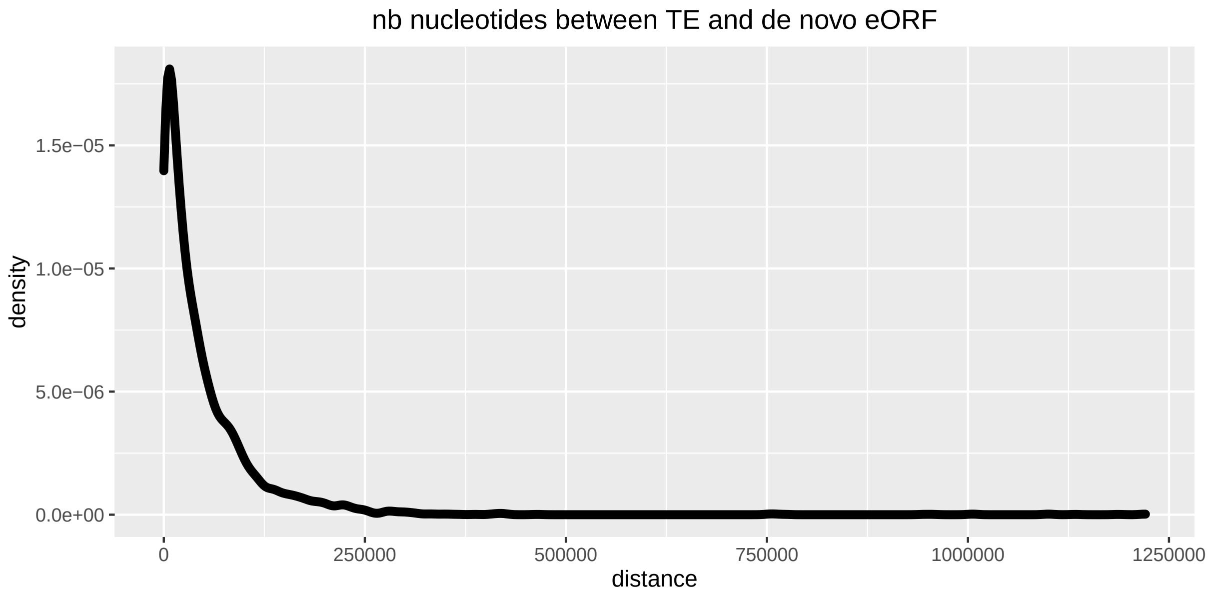

Figure 8: **Distance TE to closer neORF**

S4 TE overlap families

In this figure, we investigated which type of Tes were overlapping or containing a neORF. This result shows that neORFs overlap more often with retrotransposons. Yet, the ones that overlap with LINEs are expected to duplicate more often than the ones overlapping with retrotransposons.

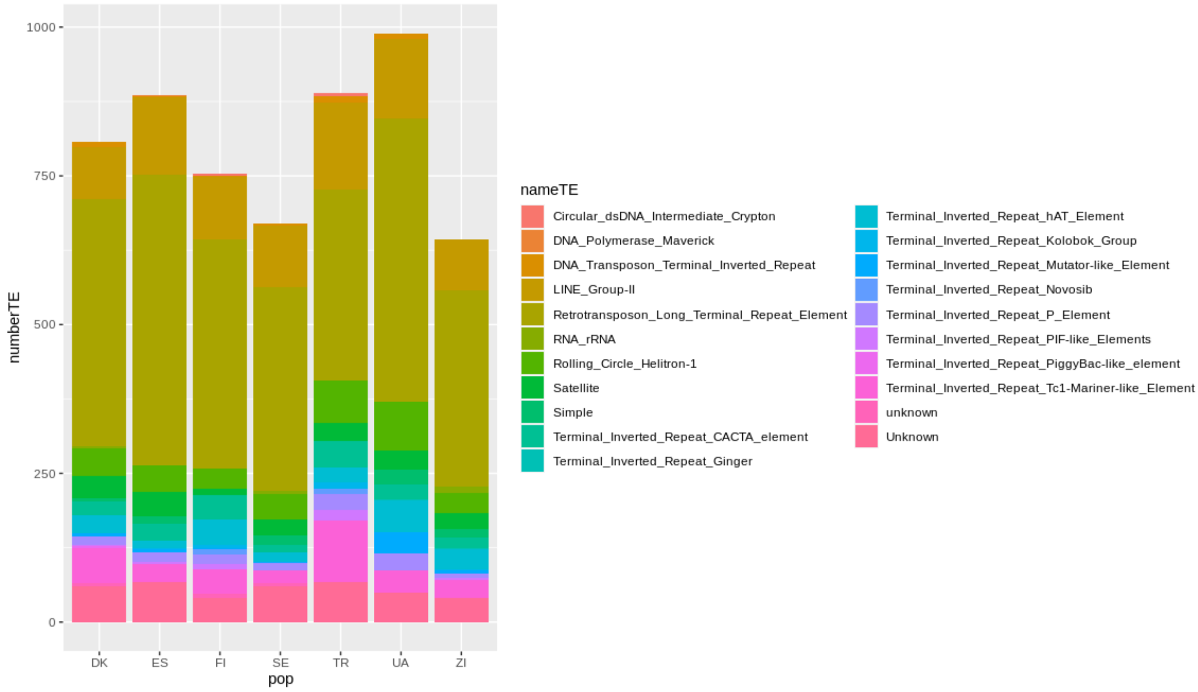

Figure 9: **TE families**

S5 TE overlap with neORF

S5.1 Percentage of neORF covered by a TE for neORFs overlapping with a TE

The x axis represents the percentage of neORFs overlapping to a TE, for the neORF that do overlap to a TE. The y axis represents the density.

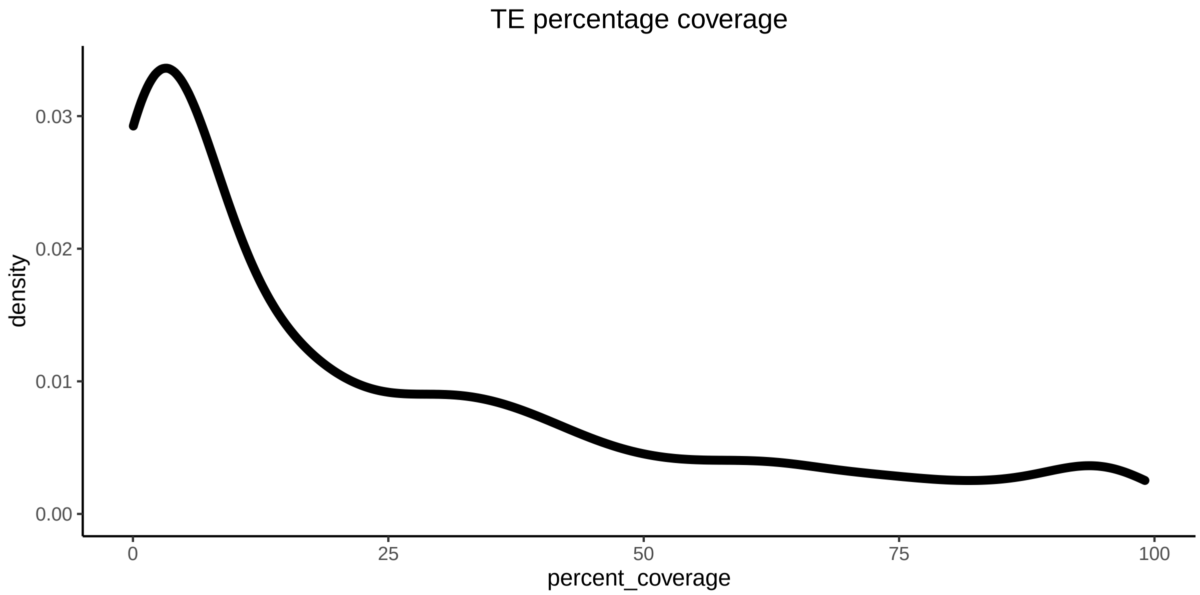

Figure 10: **neORF TE coverage**

S5.2 Percentage of neORF covered by a TE for neORFs overlapping with a TE per line

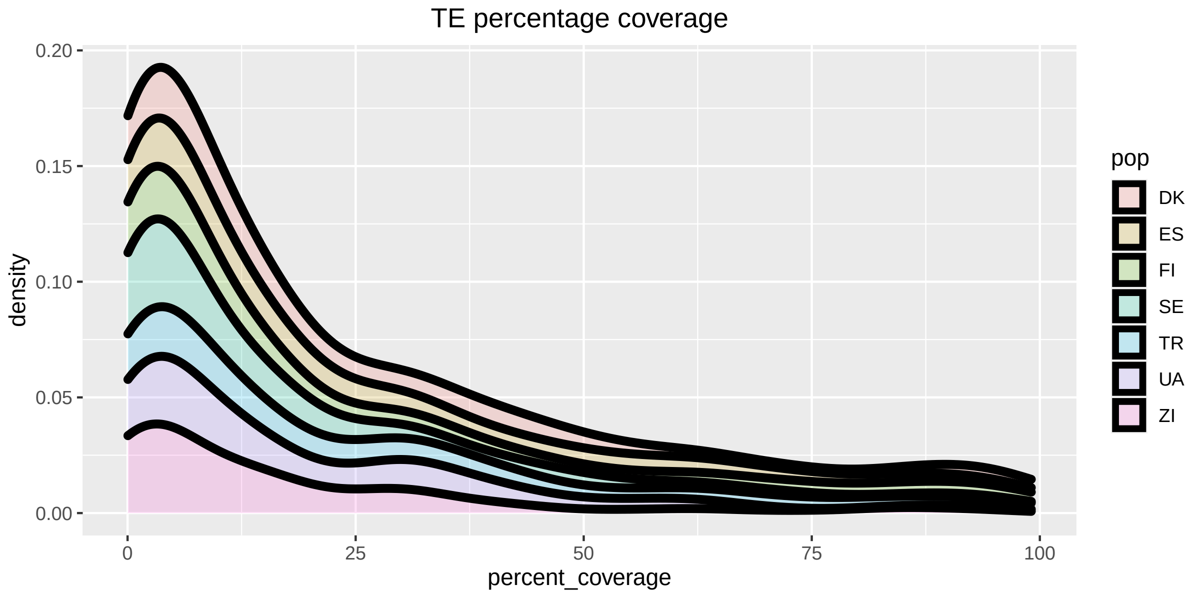

Figure 11: **neORF TE coverage**

S6 Distribution of TEs and neORFs

Distribution of TEs and neORF in the chromosomes arms of the 7 lines. The chromosomes are segmented by 100,000 nucleotides windows.

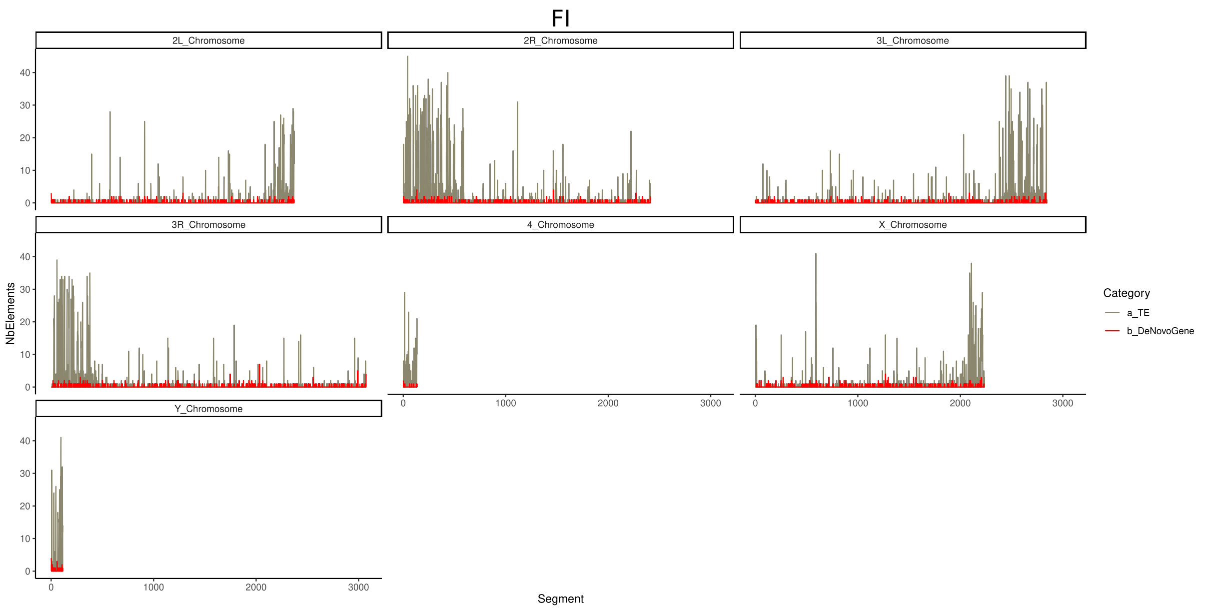

Figure 12: **TE and proto-genes distribution FI**

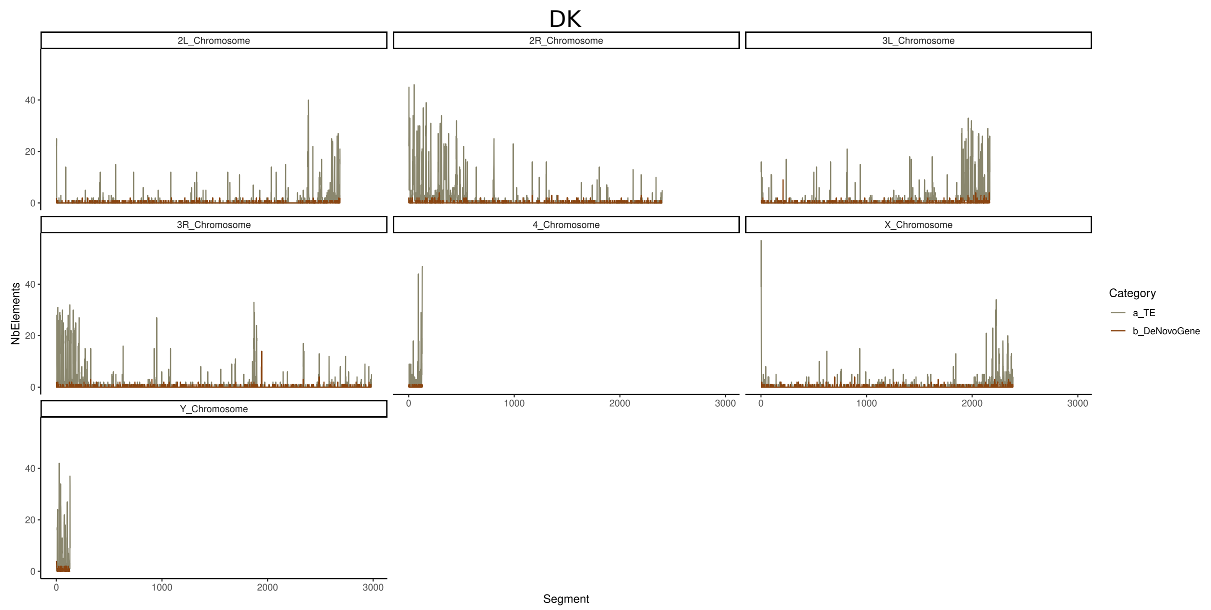

Figure 13: **TE and proto-genes distribution DK**

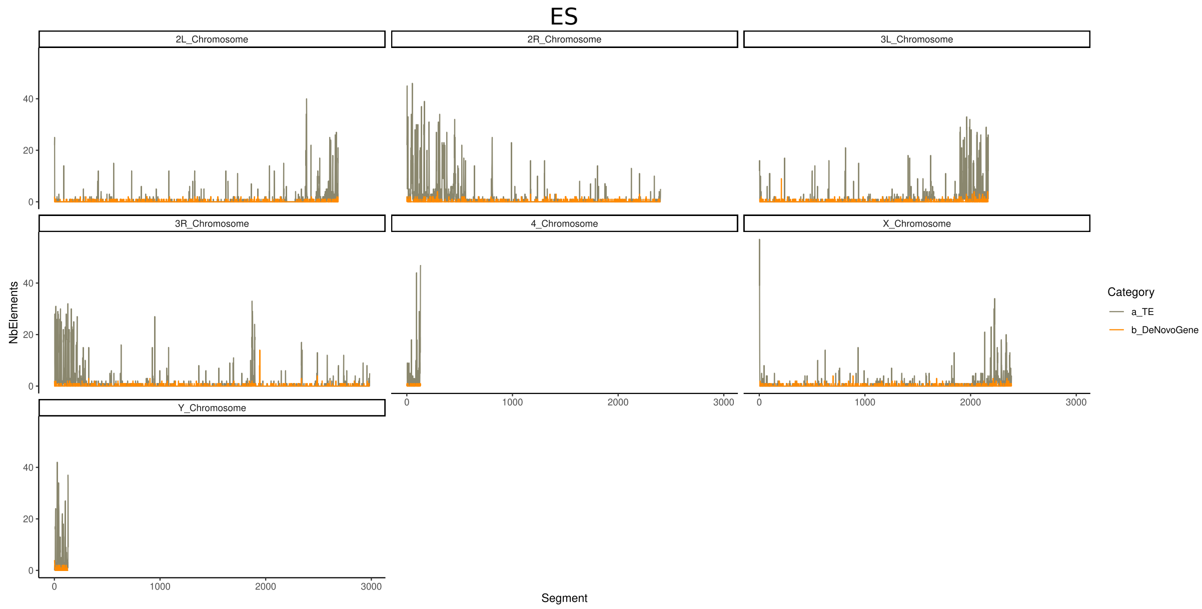

Figure 14: **TE and proto-genes distribution ES**

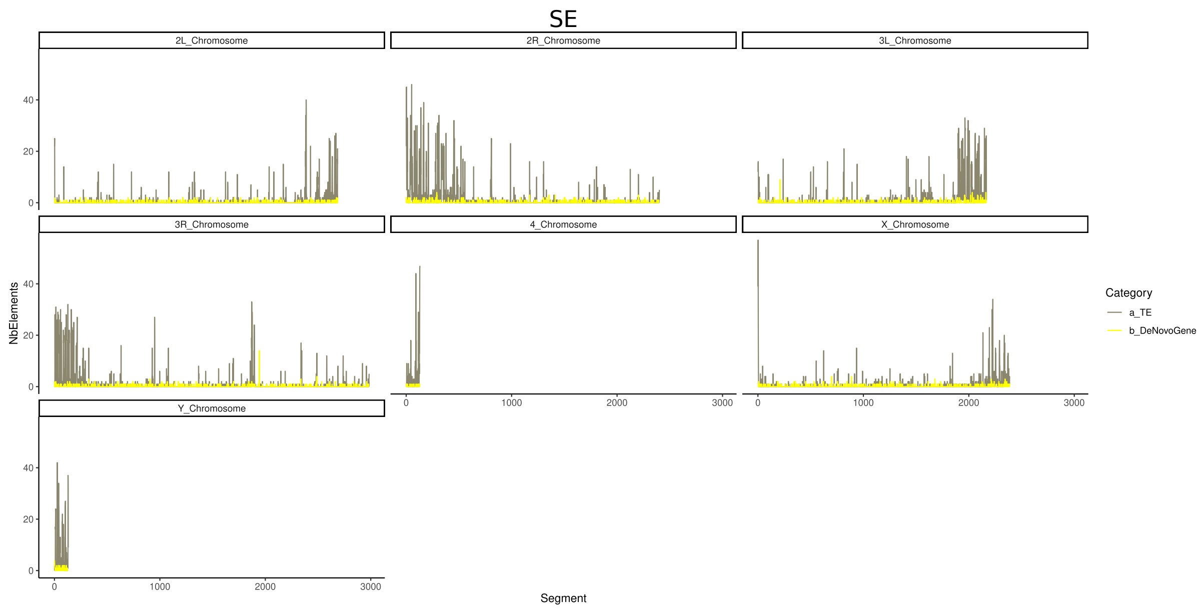

Figure 15: **TE and proto-genes distribution SE**

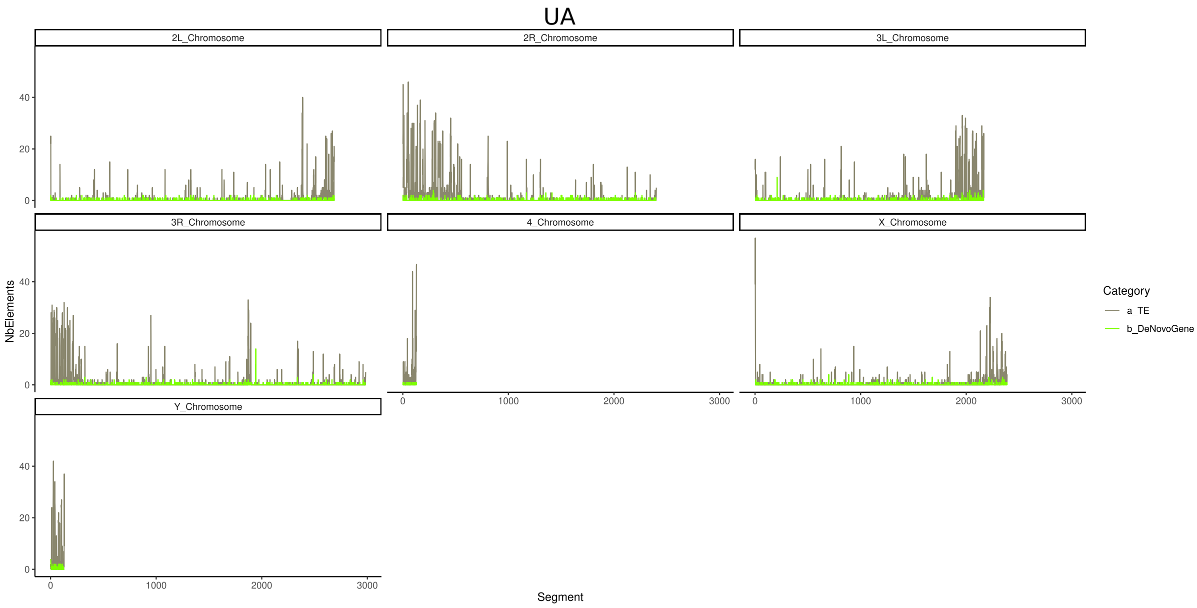

Figure 16: **TE and proto-genes distribution UA**

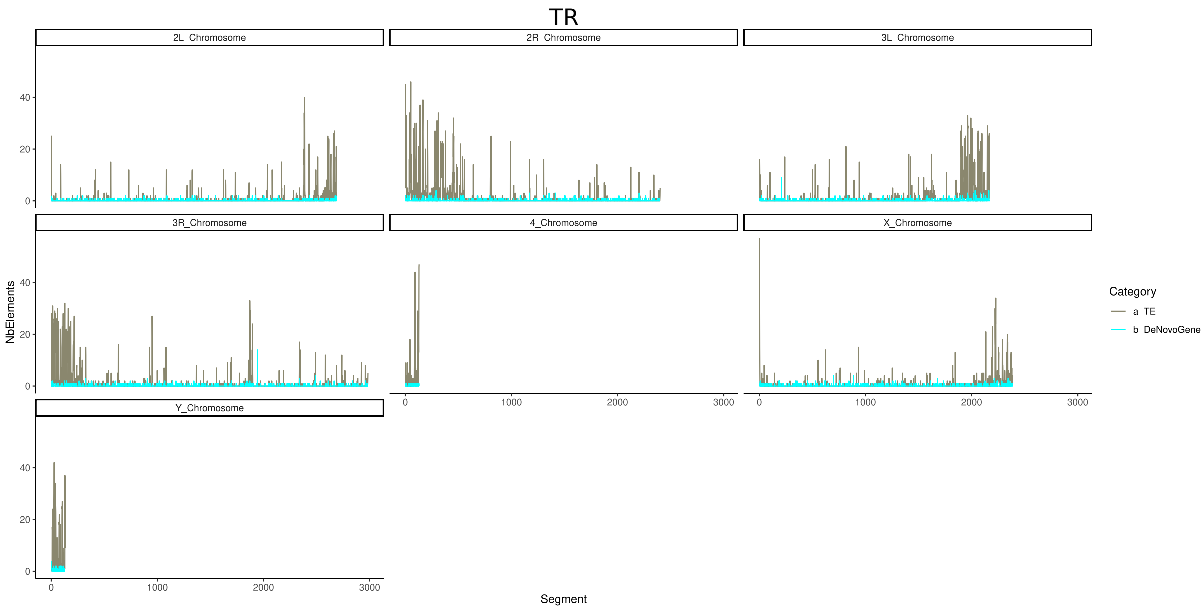

Figure 17: **TE and proto-genes distribution TR**

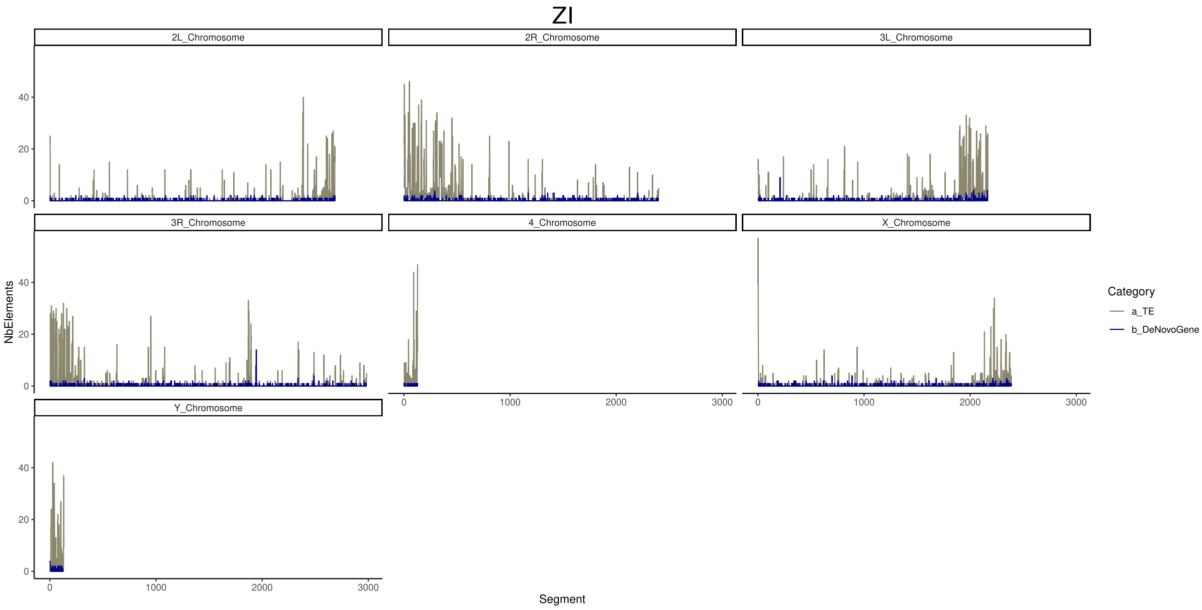

Figure 18: **TE and proto-genes distribution ZI**

S7 Details mutations

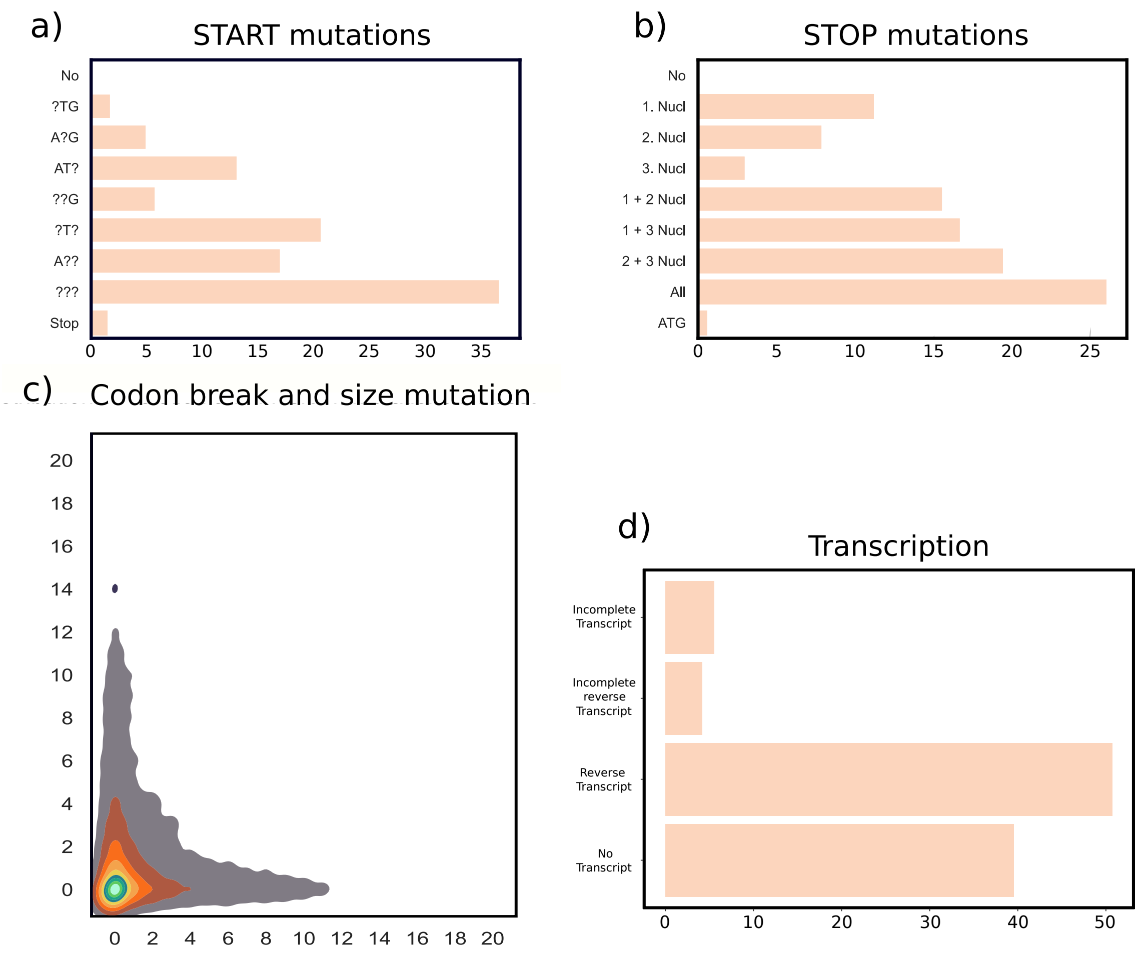

Figure 19: **Detail mutations**

S8 Position second ATG in homologous sequence

The x axis represents the position of the second START codon, showing at which percentage of the size of the homologous sequence it is found in.

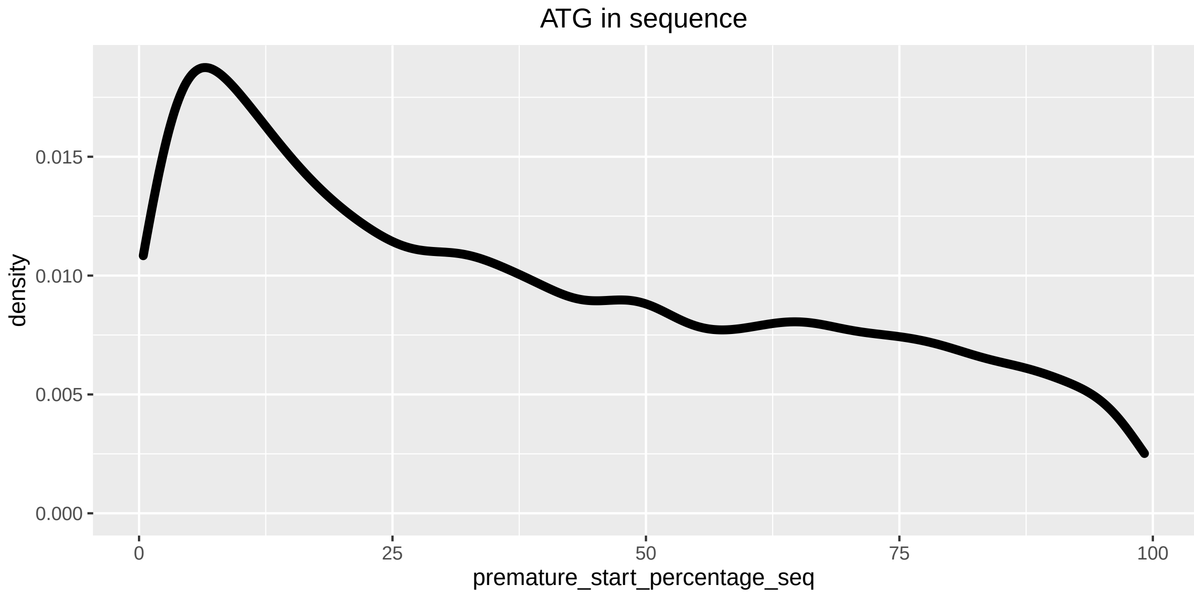

Figure 20: **Premature START position**

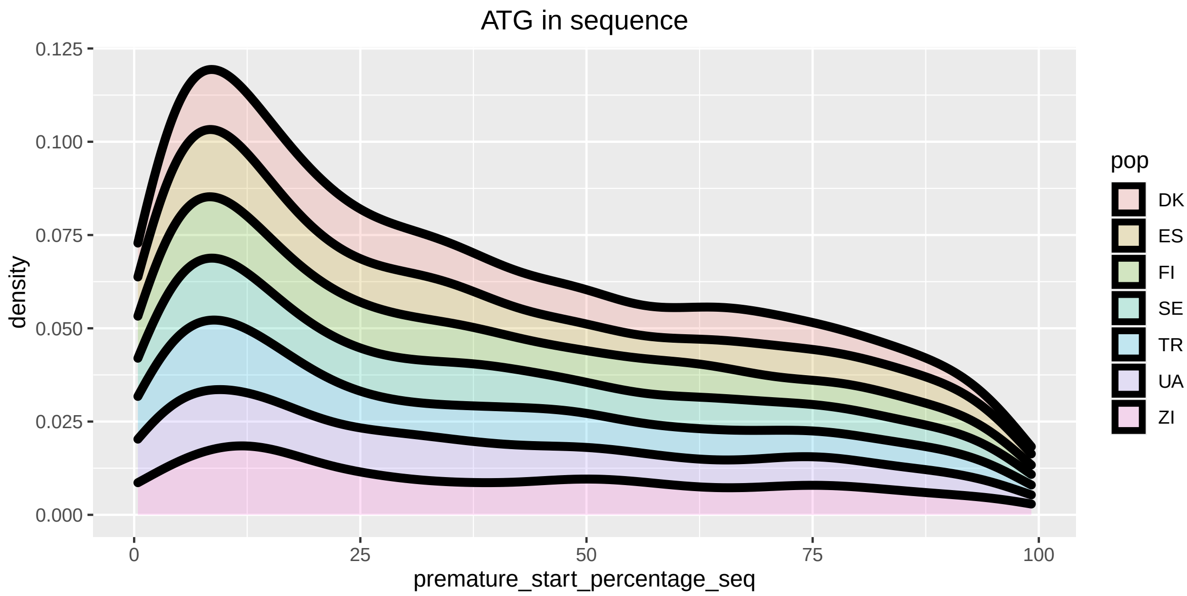

Figure 21: **Premature START position per line**

S9 Position premature stop codon in homologous sequence

Position of premature STOP codon in homologous sequence that have stop codons. The x axis represents the position of thepremature stop codon, showing at which percentage of the size of the homologous sequence it is found. The y axis represents the density.

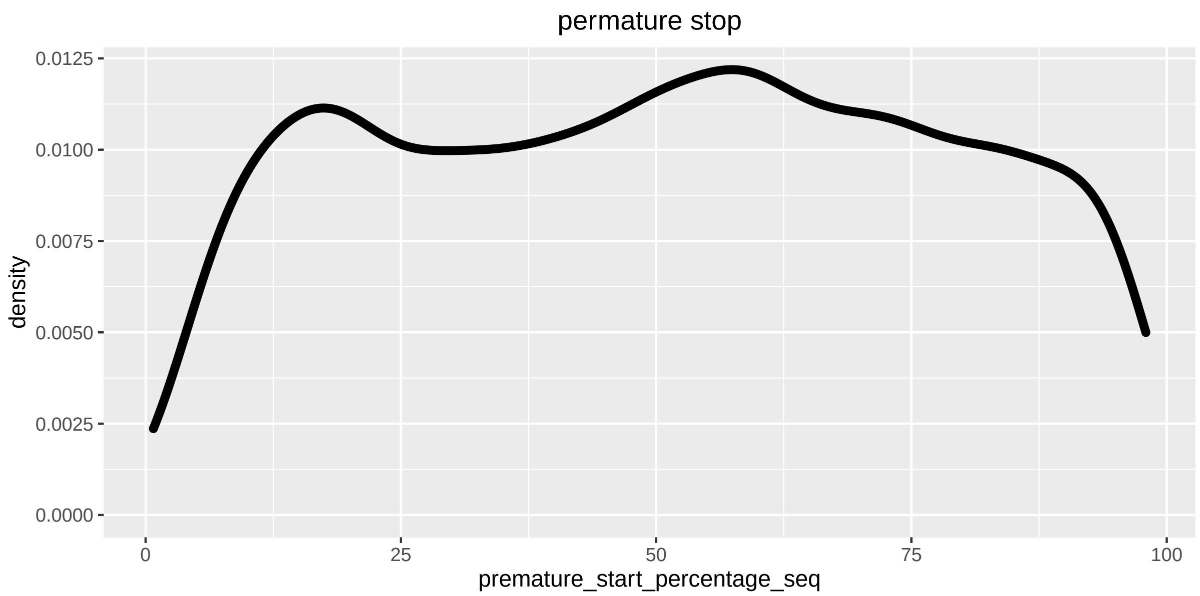

Figure 22: **Premature stop codon position**

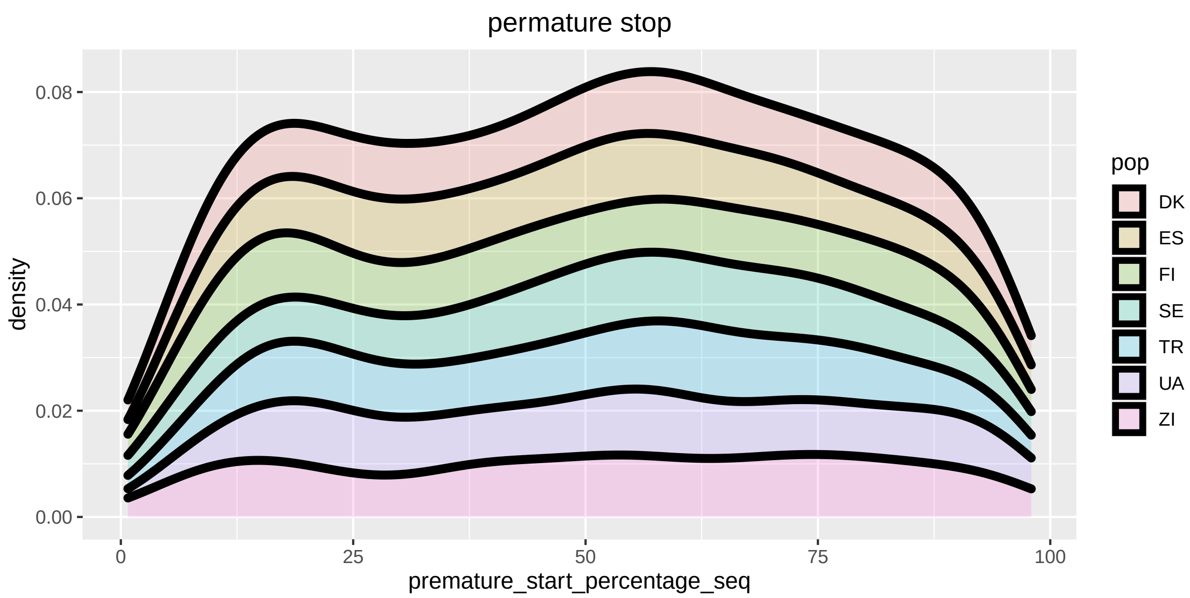

Figure 23: **Premature stop codon position per line**

S10 Length ORF vs transcript

The x axis represents the size of the unspliced transcript. The y axis represents the size of the ORF (in nucleotides), in the transcript. Interestingly, we observe that the longer a de novo transcript is, the longer is the ORF it contains.

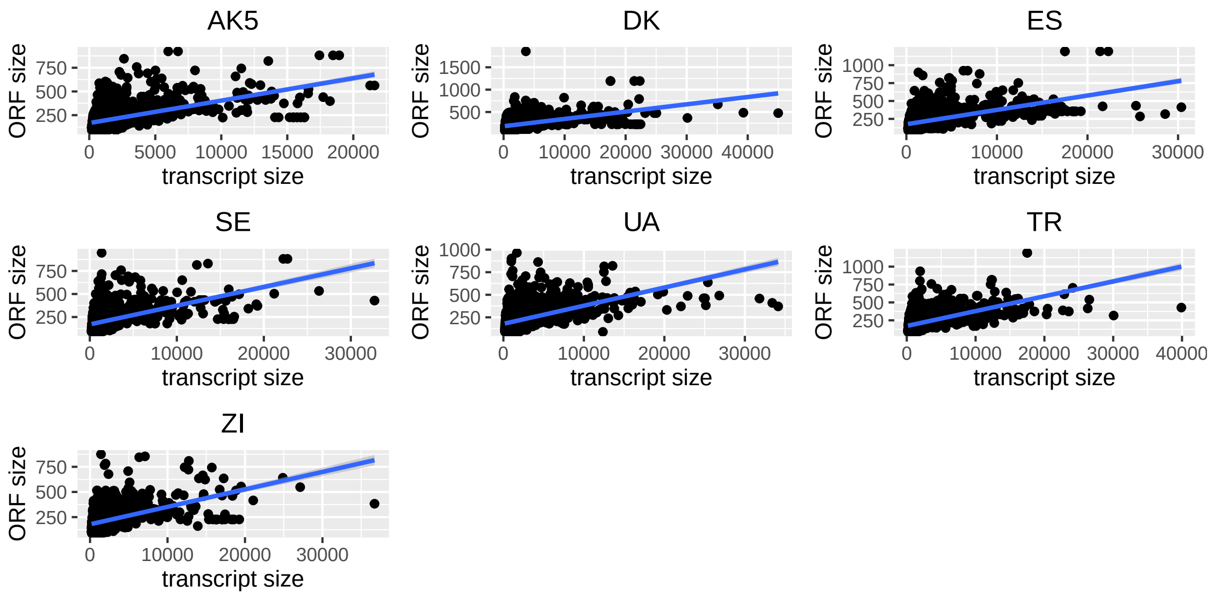

Figure 24: **Length Length ORF vs Transcript**

S11 Build orthogroup

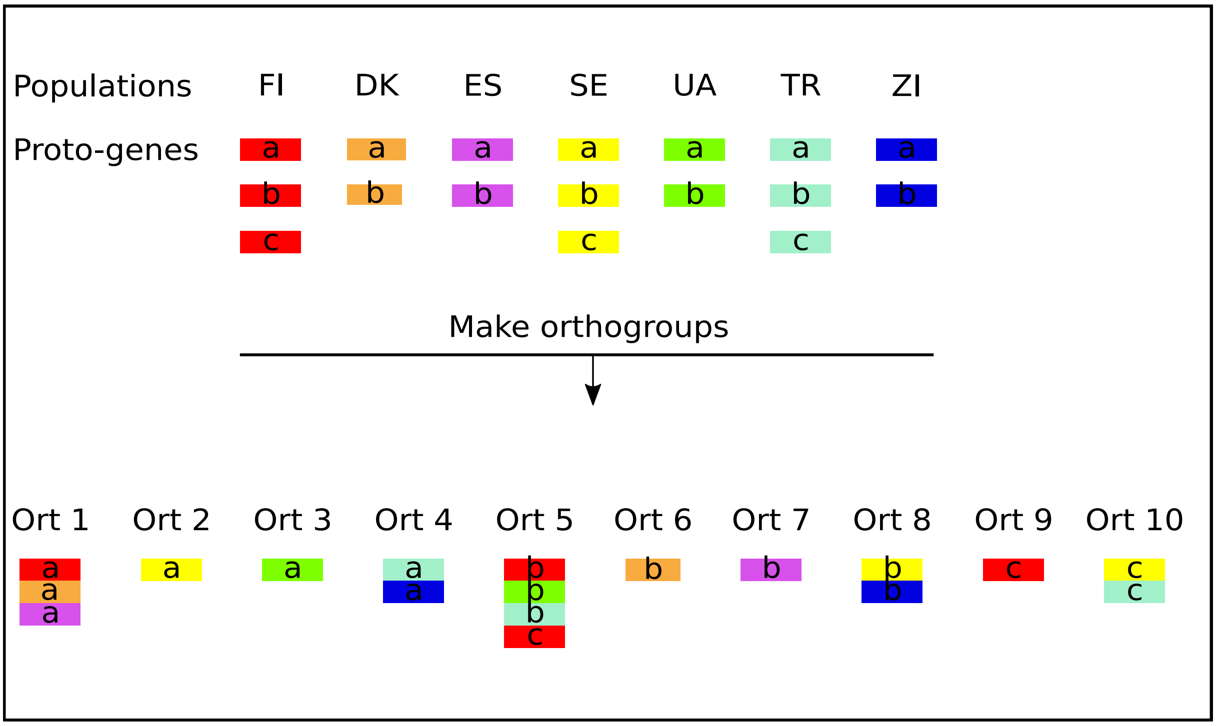

Figure 25: **Orthogroups**

S12 Flowchart detected homologous sequences

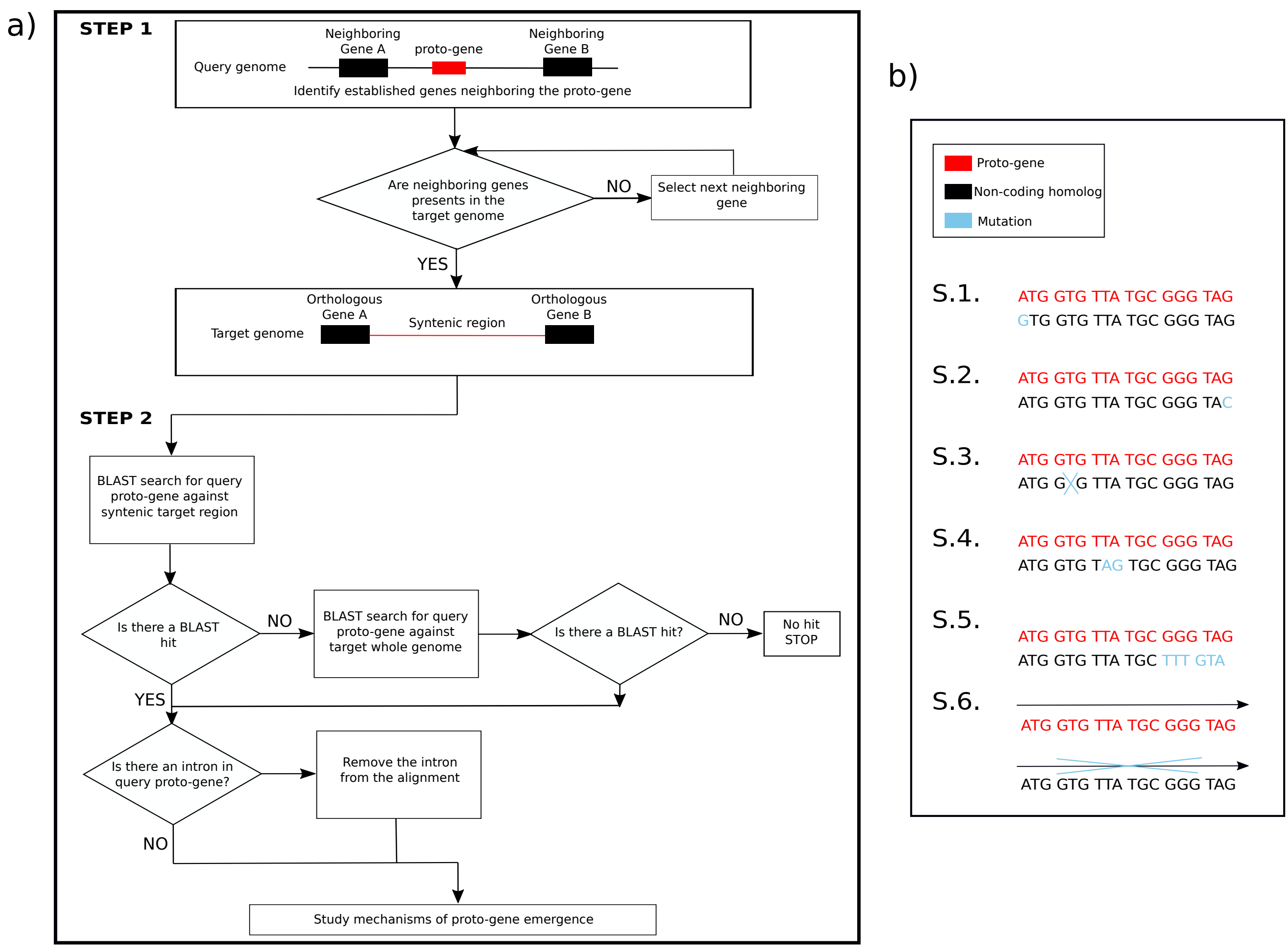

Figure 26: **Flowchart**

S13 Synteny interval between two genes

The figures represent the sizes of syntenic regions. We show the distribution of the sizes, in nucleotides, of the syntenic regions between two genes. The x axis represents the size of the syntenic region. The y axis represnts the density.

Figure 27: **Global gene interval**

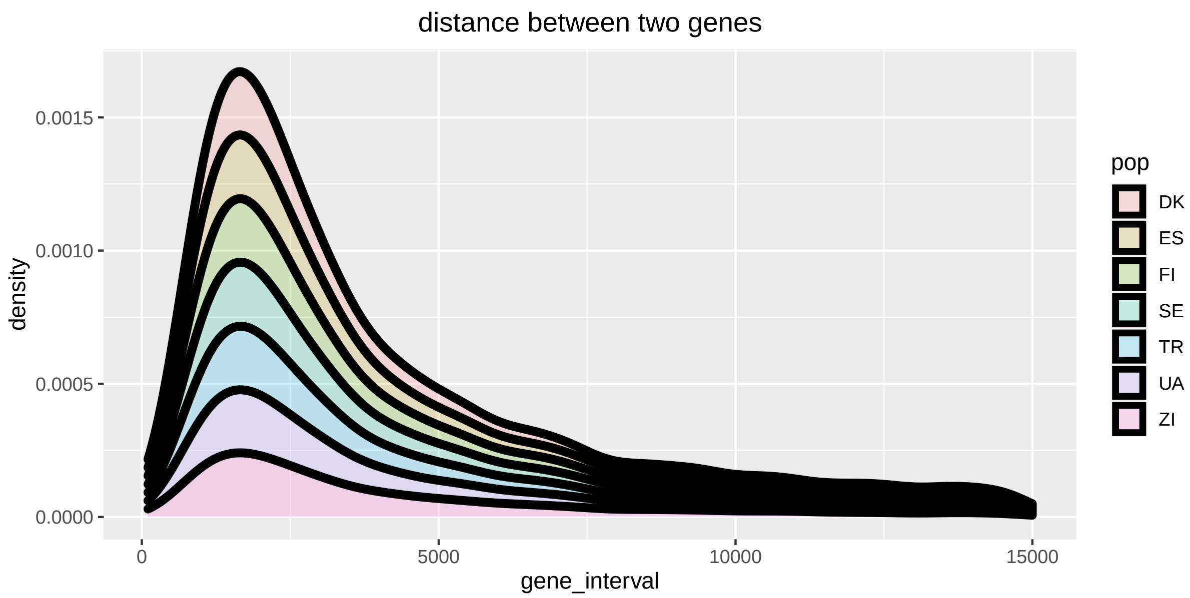

Figure 28: **Gene interval per line**

S14 HCA clusters

The figure represents the average number of hydrophobic clusters in neORFs (here called proto-genes). The x axis represents the number of line sharing the neORF. The y axis represents the average number of hydrophobic clusters in the proteins.

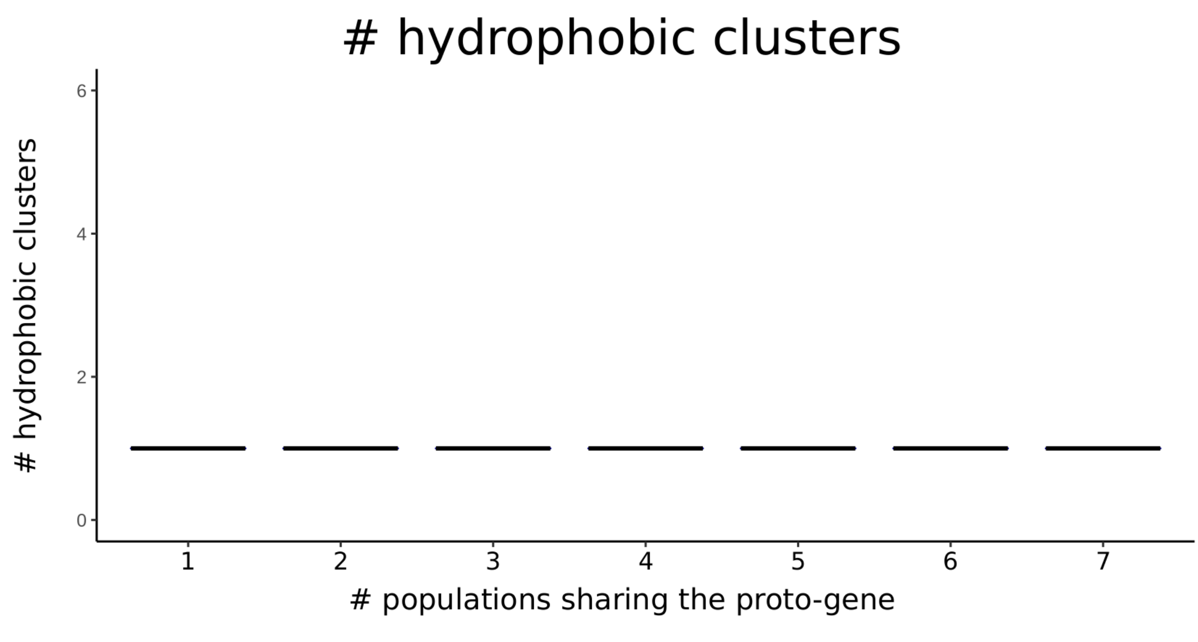

Figure 29: **HCA clusters in lines proto-genes**

S14.1 Average length of homologous sequences to neORFs

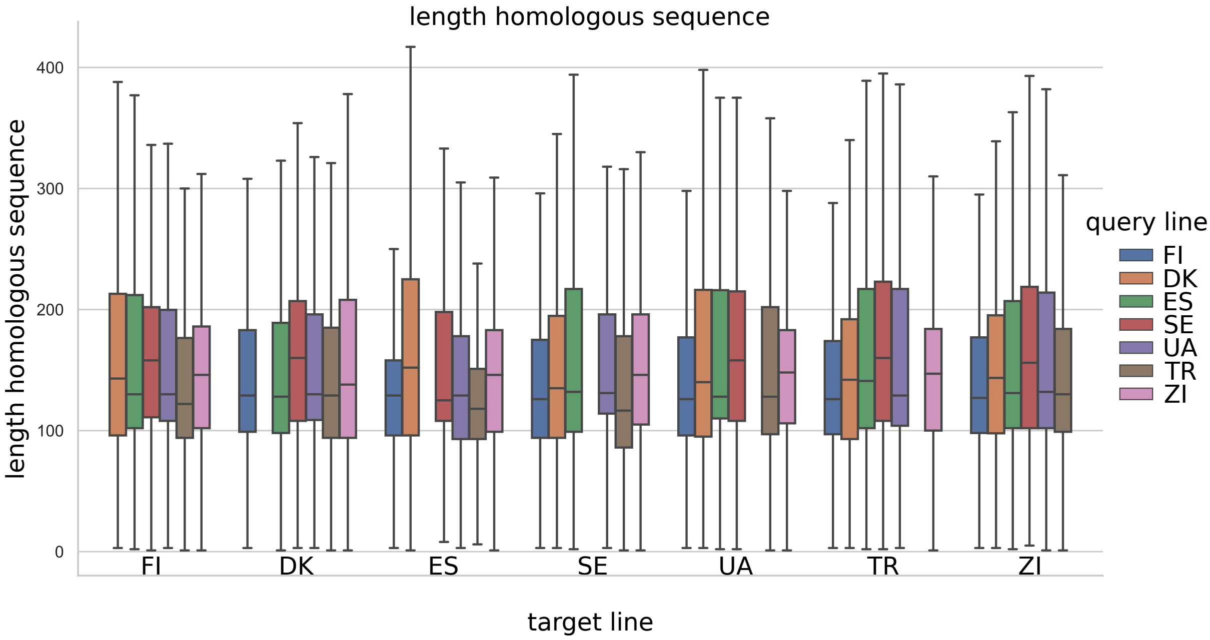

Figure 30: **Length homologous sequence**