Big Data Class – 2024

Practical training – DNAseq:
Studying structural variation in genomes
Fanny Hartmann
(fanny.hartmann@universite-paris-saclay.fr)

1-Introduction about SVs:

What are structural variants (SVs)? What are the different types of SVs? What is the effect of SVs? How to detect SVs in genomes? This will be discussed in the introduction of the class.
Here are some references to go further:

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biology* 20: 246.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution* **35**: 561–572.

2-Introduction about the study case and the data:

We will study SVs affecting *Spok* genes in the fungus *Podospora anserina*. *Spok* genes are selfish genetic elements that undergo meiotic drive, ie. preferential transmission of a particular allele during sexual reproduction. They are called spore killer genes as they are responsible for certain spore killing sibling spores. Four *Spok* genes (*Spok1, Spok2, Spok3, Spok4*) were identified to undergo SVs in *Podospora anserina.*

The data used will be:
-alignment files (.bam format) from two *P. anserina* strains, PaWa63 and PaWa46. The alignment files were generated against the reference genome of *P. anserina* (the S strain).
-annotation (.gff format) of the *Spok2* gene in the S strain.
-Nucleotide sequences (.fasta format) of colinear regions in the PaWa28 and PaWa53 strains.
-Nucleotide sequences (.fasta format) of the Spok3 gene region in the PaWa28, PaWa53 and PaWa21 strain.

The data was extracted from this study:
Vogan AA, Ament-Velásquez SL, Granger-Farbos A, Svedberg J, Bastiaans E, Debets AJ, Coustou V, Yvanne H, Clavé C, Saupe SJ, et al. 2019. Combinations of Spok genes create multiple meiotic drivers in Podospora. eLife 8.

3. Studying gene deletions

Gene deletion can be identified by studying read coverage in the alignment file of one genome against another. *Spok2* gene is present in the S strain (used as reference genome). It is also present in other strains but missing in others; it is an example of gene deletion.

Task 1: Compute read coverage of the PaWa46 strain and the PaWa63 strain in the *Spok2* gene region using the tool bedtools genomecoverage (https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html) (use

preferentially the BEDGRAPH output format) and visualize the results using R Studio (a R script is provided). Conclude on the presence or absence of the *Spok2* gene in each strain.

Task 2: The *Spok3* gene is missing in the S strain used as reference genome; can we identified the gene deletion with this method? Can you think of another method to identify *Spok3* gene deletion in the S strain?

4. Studying other SVs.

The *Spok3* gene region is affected by multiple SVs (inversion, duplication, translocation…). We will perform pairwise sequence comparison in the *Spok3* gene region in three strains of P anserina and use dot plots to visualize SVs.

Dot plots are two-dimensional plots where the x-axis and y-axis each represents a sequence and the plot itself shows a comparison of these two sequences by a calculated score for each position of the sequence. If a window of fixed size on one sequence (one axis) match to the other sequence a dot is drawn at the plot. We can use the online ncbi blast tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi?BLAST_SPEC=blast2seq&LINK_LOC=align2seq&PAGE_TYPE=BlastSearch).

Task 3: Perform sequence comparison between strains PaWa53 and PaWa28 and visualize the dot plot in a genomic region that is colinear and in the Spok3 gene region. Is there SV(s) in the Spok3 gene region ? If yes, what is the SV(s) type(s) (inversion, deletion, duplication…)?

Task 4: Perform sequence comparison between strains *PaWa28* and *PaWa21* and visualize the dot plot in the Spok3 gene region. Is there SV(s) in the Spok3 gene region ? If yes, what is the SV(s) type(s) (inversion, deletion, duplication…)?

5. Report to send:

Send me your answers to Task 1 to 4 in PDF format by email (fanny.hartmann@universite-paris-saclay.fr) using as object 'UE big data DNAseq' before Friday 15.11 AM 9:30.