Structural Variant detection

Olivier QUENEZ - INSERM

Guillaume Robert-Siegwald - Inovarion
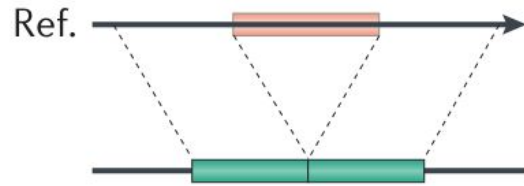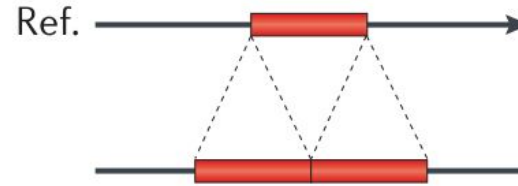*Mathieu Charles - INRAE*

Workflow
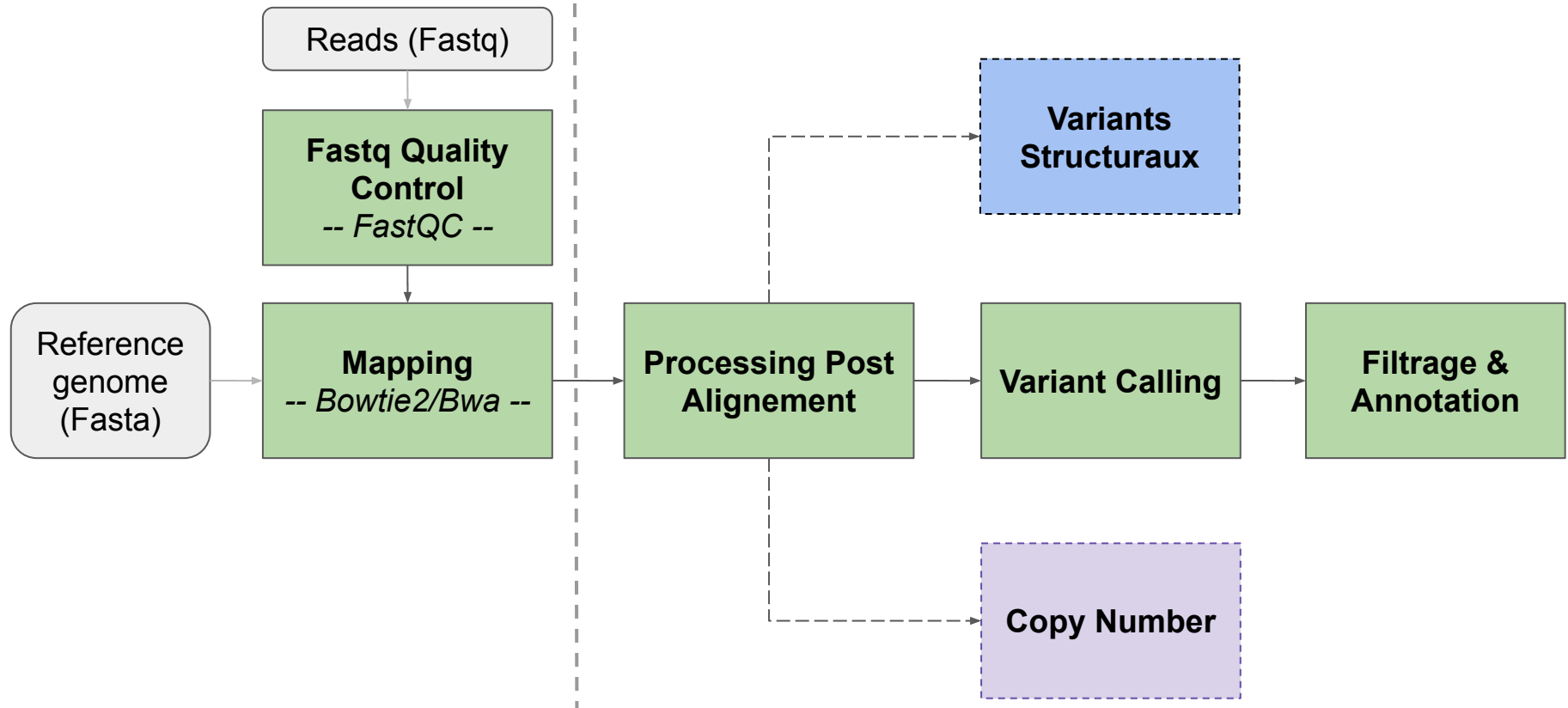
Définition

- Consensus actuel : Réarrangement génomique >50bp

- Différents types de variants structuraux :

→ Réarrangements déséquilibrés (variation du nombre de copie - CNV)

- Délétion
- Duplication

→ Réarrangements équilibrés

- Insertion
- Inversion
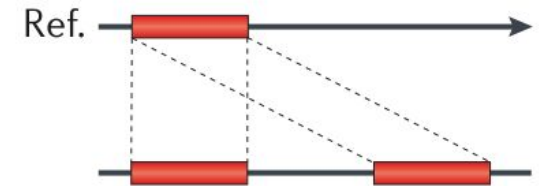- Translocation

Deletion
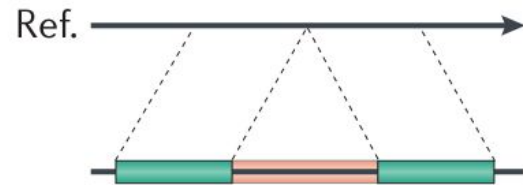

Tandem duplication

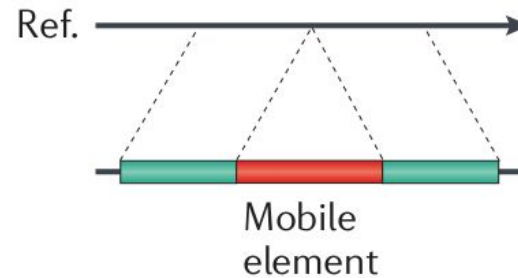# Interspersed duplication

Novel sequence insertion

Mobile-element insertion

Mobile
element

Inversion

Translocation

Alkan C, Coe BP, Eichler EEGenome structural variation discovery and genotyping. Nat Rev Genet 12:363–376
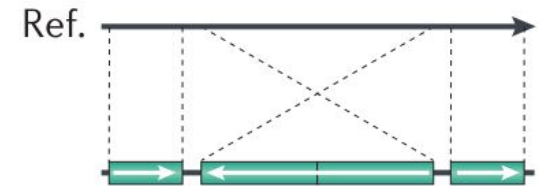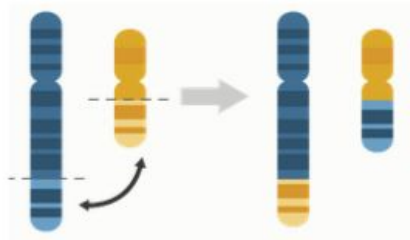
Principe de détection des SVs

A Read Depth (RD)
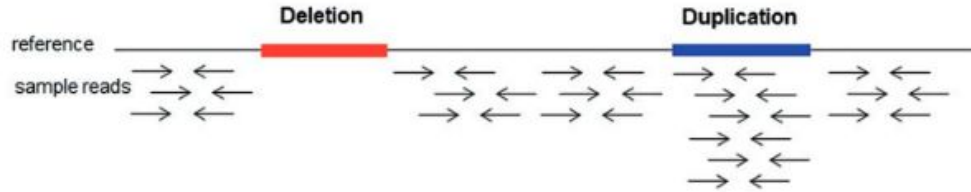
B Paired Reads (PR)
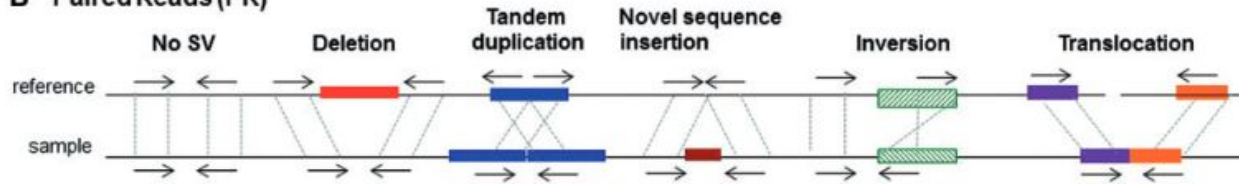
C Split Reads (SR)
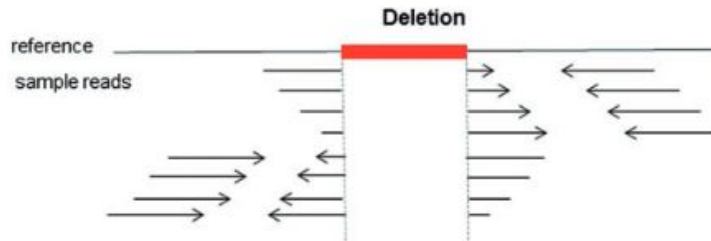
D. De Novo Assembly (AS)
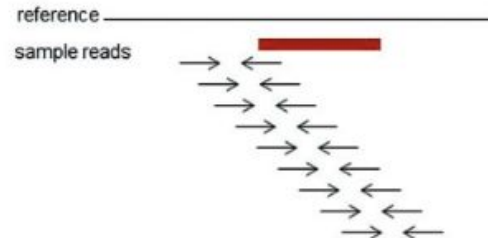
Short reads ou long reads?

Short reads (Illumina) : selon l'outil et la qualité des données

→ **faible recall** : 10 à 70% des SVs détectés

→ **faible précision** : jusqu'à 90% de Faux Positifs

→ Difficulté à caractériser des SVs complexes (alignement imprécis dans les régions répétées et faible résolution)

/!\ Un calling consensus avec plusieurs outils de détection peut être utile avec des données short reads /!\

Long reads (PacBio/MinION) :

→ Meilleure caractérisation des altérations des régions répétées

→ Une faible profondeur de couverture suffit (15-30x)

Quel outil choisir ?

Critères de choix :

- Ai-je des données short reads ou long reads ?

- Ai-je de nombreux échantillons ?

- Quel type de SV est-ce que je recherche ?

- Est-ce que la profondeur de couverture est suffisante ?

- Que privilégier : sensibilité et / ou spécificité

- Quel est le format de sortie de l'outil ?

Détection de SV pour données short reads

	SV Callers	CNV	INS	DEL	DUP	INV	TRA	Data	Discovery RD	Discovery SC	Discovery PR	Discovery OEA	Discovery UM	Validation RD	Validation SC	Validation PR	Validation OEA	Validation UM	CL	SA	CA	ST	BP Resolution (Y/N)	References
CNV	BIC-seq	x						PE;SE	x					x								x	N	[110]
	cn.MOPS	x						PE;SE	x					x								x	N	[44]
	cnD	x						PE	x					x								x	N	[88]
	CNVeM	x						PE	x					x								x	N	[105]
	CNVnator	x						PE;SE	x					x								x	N	[3]
	CNV-seq	x						PE;SE	x					x								x	N	[111]
	JointSLM	x						PE;SE	x					x								x	N	[59]
	RDXplorer	x						SE	x					x								x	N	[115]
	SegSeq	x						PE;SE	x					x								x	N	[15]
	CNVer	x						PE			x			x					x			x	N	[62]
SV	LUMPY			x	x	x	x	PE	x	x	x			x	x	x			x	x			N	[50]
	MetaSV		x	x	x	x	x	PE	x	x	x			x	x	x			x	x	x		Y	[65]
	SVM2		x	x				PE	x		x				x	x			x			x	N	[16]
	Breakpointer		x	x				SE	x	x				x	x					x		x	N	[95]
	Meerkat		x	x	x	x	x	PE		x	x	x			x		x		x	x			Y	[112,113]
	Scalpel		x	x				PE		x	x	x									x		Y	[68]
	SVMerge		x	x	x	x	x			x	x			x	x	x			x	x	x		Y	[109]
	SoftSV			x	x	x	x	PE	x	x				x	x				x	x			Y	[9]
	BreaKmer		x	x	x	x	x	PE		x		x			x						x		Y	[2]
	ClipCrop		x	x	x	x	x	PE		x					x				x	x			Y	[97]
	CREST			x	x	x	x	PE;SE		x					x						x		Y	[104]
	Gustaf			x	x	x	x	PE;SE		x					x					x			Y	[99]
	Socrates			x	x	x	x	PE;SE		x					x				x	x			Y	[86]
	Bellerophon						x	PE			x			x	x				x	x			Y	[30]
	BreakDancer		x	x	x	x	x	PE			x			x					x			x	N	[14]
	CLEVER		x	x				PE			x				x				x			x	N	[60]
	DELLY			x	x	x	x	PE			x				x		x		x	x			Y	[80]
	FACTERA			x			x	PE			x				x				x	x			Y	[69]
	GASV		x	x	x	x	x	PE			x								x				N	[90]
	GASVPro		x	x	x	x	x	PE			x			x		x			x			x	N	[91]
	GenomeSTRiP			x				PE			x					x			x		x		N	[29]
	HYDRA			x	x	x	x	PE			x					x			x				Y	[78]
	HYDRA-Multi			x	x	x	x	PE			x					x			x				Y	[58]
	inGAP-SV		x	x	x	x	x	PE	x					x									N	[76]
	MoDIL		x	x				PE			x					x						x	N	[51]
	PEMer		x	x	x	x	x	PE			x								x				N	[45]
	PeSV-Fisher			x	x	x	x	PE;MP			x			x					x				N	[21]
	PRISM		x	x	x	x	x	PE			x						x		x	x			Y	[37]
	RetroSeq		x					PE			x				x				x	x			Y	[40]
	SVDetect		x	x	x	x	x	PE;MP			x					x			x				N	[116]
	SVMiner		x	x	x	x		PE			x			x		x			x			x	N	[31]
	Ulysses		x	x	x	x	x	MP			x			x		x			x			x	N	[25]
	VariationHunter		x	x				PE			x					x			x				N	[32]
	NovelSeq		x					PE				x	x						x		x		Y	[27]
	PINDEL		x	x				PE			x									x			Y	[114]
	SLOPE		x	x			x	PE;SE			x						x		x	x			Y	[1]
	SOAPindel		x	x				PE			x			x		x	x				x		Y	[55]
	Splitread		x	x				PE			x									x			Y	[39]
	BreakSeq		x	x				PE							x					x			Y	[47]
	SMUFIN		x	x		x	x	PE											x		x		Y	[66]

Outils en long reads

- PBHoney, 2014

- SMRT-SV, 2015

- Hysa, 2016 (hybrid avec short reads)

- NanoSV, 2017

- Sniffles, 2018

Partie TP

Data : souche de *Zymoseptoria tritici* séquencées à la fois en Illumina et en MinION.

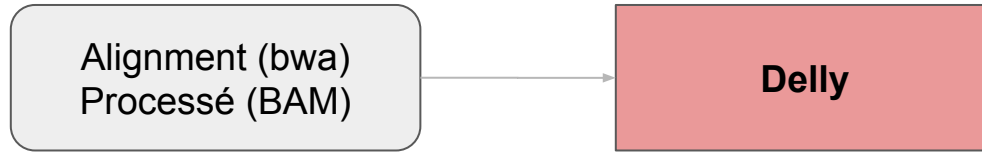→ chaque set de reads a été aligné sur le génome de référence avec les outils dédiés

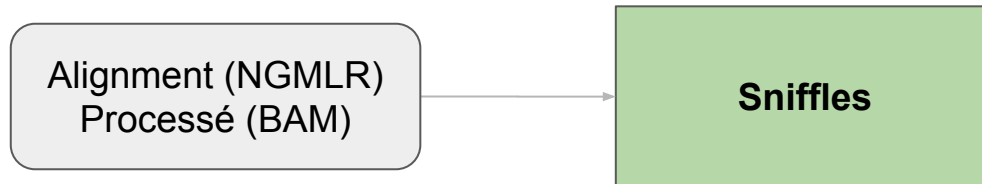→ les données ont été réduites aux premiers 500kb du chr10

Tools :

- **Delly** *(Bioinformatics, Volume 28, Issue 18, 15 September 2012, Pages i333-i339,*
  https://doi.org/10.1093/bioinformatics/bts378)

- **Sniffles** *(Nature Methods volume 15, pages 461-468 (2018) ,*
  https://www.nature.com/articles/s41592-018-0001-7) with **NGMLR** mapping

Workflow - Variants Structuraux

Short Reads (Illumina)

Alignment (bwa) Processé (BAM) → Delly

Long Reads (MinION)

Alignment (NGMLR) Processé (BAM) → Sniffles

Préparation des données

```
# Copie des données SV
$ cp -R /shared/projects/ebai2019/atelier_variant/data/sv ~/tp_sv
$ cd ~/tp_sv

# Indexation des fichiers
$ module load samtools/1.10

$ samtools index mapping_illumina_chr10_500kb.bam
$ samtools index mapping_minion_chr10_500kb.bam
$ samtools faidx Zymoseptoria_tritici.fa

$ ls -l
```

```
13812904 Oct 28 14:41 mapping_illumina_chr10_500kb.bam
    1720 Oct 28 14:51 mapping_illumina_chr10_500kb.bam.bai
43323244 Oct 28 14:41 mapping_minion_chr10_500kb.bam
    9040 Oct 28 14:51 mapping_minion_chr10_500kb.bam.bai
40348870 Oct 28 14:41 Zymoseptoria_tritici.fa
     606 Oct 28 14:44 Zymoseptoria_tritici.fa.fai
```

Delly

```
$ mkdir -p delly/logs
$ cd delly

$ module load delly/0.8.3
$ delly        # (v0.8.3)
$ delly call

Usage: delly call [OPTIONS] -g <ref.fa> <sample1.sort.bam> <sample2.sort.bam> ...

Generic options:
  -? [ --help ]                 show help message
  -t [ --svtype ] arg (=ALL)    SV type to compute [DEL, INS, DUP, INV,
                                BND, ALL]
  -g [ --genome ] arg           genome fasta file
  -x [ --exclude ] arg          file with regions to exclude
  -o [ --outfile ] arg (="sv.bcf")  SV BCF output file
```

Delly

```
$ sbatch -J delly -o logs/delly.out -e logs/delly.err --mem=8G --wrap=" \
    delly call -g ~/tp_sv/Zymoseptoria_tritici.fa \
    -o SV_calling_illumina.bcf ~/tp_sv/mapping_illumina_chr10_500kb.bam"
```

```
$ less SV_calling_illumina.bcf
# "delly/SV_calling_illumina.bcf" may be a binary file.  See it anyway? n
```

```
# Conversion en fichier vcf
$ module load bcftools/1.10.2

$ sbatch -J bcf_to_vcf -o logs/bcf_to_vcf.out -e logs/bcf_to_vcf.err --wrap=" \
    bcftools view SV_calling_illumina.bcf > SV_calling_illumina.vcf"

$ less -S SV_calling_illumina.vcf        # "Q" pour quitter
```

Header du vcf de Delly

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20200804
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=BND,Description="Translocation">
##ALT=<ID=INS,Description="Insertion">
##FILTER=<ID=LowQual,Description="Poor quality and insufficient number of PEs and SRs.">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="PE confidence interval around END">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="PE confidence interval around POS">
##INFO=<ID=CHR2,Number=1,Type=String,Description="Chromosome for POS2 coordinate in case of an inter-chromosomal translocation">
##INFO=<ID=POS2,Number=1,Type=Integer,Description="Genomic position for CHR2 in case of an inter-chromosomal translocation">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the structural variant">
##INFO=<ID=PE,Number=1,Type=Integer,Description="Paired-end support of the structural variant">
##INFO=<ID=MAPQ,Number=1,Type=Integer,Description="Median mapping quality of paired-ends">
##INFO=<ID=SRMAPQ,Number=1,Type=Integer,Description="Median mapping quality of split-reads">
##INFO=<ID=SR,Number=1,Type=Integer,Description="Split-read support">
##INFO=<ID=SRQ,Number=1,Type=Float,Description="Split-read consensus alignment quality">
##INFO=<ID=CONSENSUS,Number=1,Type=String,Description="Split-read consensus sequence">
##INFO=<ID=CE,Number=1,Type=Float,Description="Consensus sequence entropy">
##INFO=<ID=CT,Number=1,Type=String,Description="Paired-end signature induced connection type">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Insertion length for SVTYPE=INS.">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=PRECISE,Number=0,Type=Flag,Description="Precise structural variation">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVMETHOD,Number=1,Type=String,Description="Type of approach used to detect SV">
##INFO=<ID=INSLEN,Number=1,Type=Integer,Description="Predicted length of the insertion">
##INFO=<ID=HOMLEN,Number=1,Type=Integer,Description="Predicted microhomology length using a max. edit distance of 2">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Log10-scaled genotype likelihoods for RR,RA,AA genotypes">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=FT,Number=1,Type=String,Description="Per-sample genotype filter">
##FORMAT=<ID=RC,Number=1,Type=Integer,Description="Raw high-quality read counts or base counts for the SV">
##FORMAT=<ID=RCL,Number=1,Type=Integer,Description="Raw high-quality read counts or base counts for the left control region">
##FORMAT=<ID=RCR,Number=1,Type=Integer,Description="Raw high-quality read counts or base counts for the right control region">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Read-depth based copy-number estimate for autosomal sites">
##FORMAT=<ID=DR,Number=1,Type=Integer,Description="# high-quality reference pairs">
##FORMAT=<ID=DV,Number=1,Type=Integer,Description="# high-quality variant pairs">
##FORMAT=<ID=RR,Number=1,Type=Integer,Description="# high-quality reference junction reads">
##FORMAT=<ID=RV,Number=1,Type=Integer,Description="# high-quality variant junction reads">
##reference=Zymoseptoria_tritici.fa
##contig=<ID=chr_1,length=6088797>
```

15

Delly : comptage du nombre de SVs

```
# Combien de variants ?
$ grep -v -c "^#" SV_calling_illumina.vcf
```

```
# Combien de variants de bonne qualité ?
$ grep -v "^#" SV_calling_illumina.vcf | grep -v -c "LowQual"
```

Delly : comptage du nombre de SVs

```
# Combien de variants de bonne qualité de type Deletion…

$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep -c "<DEL>"

$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep -c "<DUP>"

$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep -c "<INV>"

$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep -c "<BND>"

$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep -c "<INS>"
```

Delly : extraction des informations

```
$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep "<DEL>"
```

Delly : extraction des informations des délétions

```
#Récupération du start des variants
$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep "<DEL>" | \
    cut -f1,2 > delly_del_start.txt
```

```
#Récupération des autres informations
$ grep -v "^#" SV_calling_illumina.vcf | grep -v "LowQual" | grep "<DEL>" | \
    cut -f8 | cut -d ";" -f1,4,5,13 | sed "s/;/\t/g" > delly_del_info.txt
```

```
#Fusion des deux fichiers
$ paste -d '\t' delly_del_start.txt delly_del_info.txt > delly_del.txt
```

```
#Formattage et ménage
$ awk '{print $1"\t"$2"\t"$4"\t"$3"\t"$5"\t"$6}' delly_del.txt | sed "s/END=//g" \
    > delly_del.csv
$ rm delly_del_info.txt delly_del_start.txt delly_del.txt
```

Workflow - Variants Structuraux

Short Reads (Illumina)

Alignment Processé (BAM) → Delly

- -

Long Reads (MinION)

Alignment Processé (BAM) → Sniffles

Détection de données long reads avec Sniffles

```
$ module load sniffles/1.0.11
$ sniffles --help
Usage: sniffles [options] -m <sorted.bam> -v <output.vcf>
Version: 1.0.11
Contact: fritz.sedlazeck@gmail.com

Input/Output:
    -m <string>,  --mapped_reads <string>
    (required)  Sorted bam File
    -v <string>,  --vcf <string>
    VCF output file name []
    -b <string>,  --bedpe <string>
        bedpe output file name []
    --Ivcf <string>
    Input VCF file name. Enable force calling []
    --tmp_file <string>
    path to temporary file otherwise Sniffles will use the current directory. []
( -l <int>,  --min_length <int>
    Minimum length of SV to be reported. [30] )
```

Sniffles

```
$ mkdir -p ~/tp_sv/sniffles/logs
$ cd ~/tp_sv/sniffles

$ sbatch -J sniffles -o logs/sniffles.out -e logs/sniffles.err --mem=8G --wrap=" \
    sniffles -l 100 -m ~/tp_sv/mapping_minion_chr10_500kb.bam \
    -v SV_calling_minion.vcf"

$ less -S SV_calling_minion.vcf
```

Header du vcf de Sniffles

```
##fileformat=VCFv4.3
##source=Sniffles
##fileDate=20191028
##contig=<ID=chr_1,length=6088797>
##contig=<ID=chr_2,length=3860111>
##contig=<ID=chr_3,length=3505381>
##contig=<ID=chr_4,length=2880011>
##contig=<ID=chr_5,length=2861803>
##contig=<ID=chr_6,length=2674951>
##contig=<ID=chr_7,length=2665280>
##contig=<ID=chr_8,length=2443572>
##contig=<ID=chr_9,length=2142475>
##contig=<ID=chr_10,length=1682575>
##contig=<ID=chr_11,length=1624292>
##contig=<ID=chr_12,length=1462624>
##contig=<ID=chr_13,length=1185774>
##contig=<ID=chr_14,length=773098>
##contig=<ID=chr_15,length=639501>
##contig=<ID=chr_16,length=607044>
##contig=<ID=chr_17,length=584099>
##contig=<ID=chr_18,length=573698>
##contig=<ID=chr_19,length=549847>
##contig=<ID=chr_20,length=472105>
##contig=<ID=chr_21,length=409213>
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=INVDUP,Description="InvertedDUP with unknown boundaries">
##ALT=<ID=TRA,Description="Translocation">
##ALT=<ID=INS,Description="Insertion">
##INFO=<ID=CHR2,Number=1,Type=String,Description="Chromosome for END coordinate in case of a translocation">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the structural variant">
##INFO=<ID=MAPQ,Number=1,Type=Integer,Description="Median mapping quality of paired-ends">
##INFO=<ID=RE,Number=1,Type=Integer,Description="read support">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=PRECISE,Number=0,Type=Flag,Description="Precise structural variation">
##INFO=<ID=UNRESOLVED,Number=0,Type=Flag,Description="An insertion that is longer than the read and thus we cannot predict the full size.">
```

Sniffles : comptage du nombre de SVs

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | wc -l
```

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DEL" | wc -l

$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DUP" | wc -l

$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "INV" | wc -l

$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "TRA" | wc -l

$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "INS" | wc -l
```

Sniffles : extraction des positions des délétions

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DEL"
```

Sniffles : extraction des positions des délétions

revoir les commandes, DEL peuvent apparaître différemment

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DEL" | cut -f -2 \
    > sniffles_del_start.txt
```

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DEL" | cut -d ";" -f 4 | \
    cut -d "=" -f 2 > sniffles_del_stop.txt
```

```
$ cat SV_calling_minion.vcf | grep ^chr_10 | grep "DEL" | cut -f 8 | \
    cut -d ";" -f 1 > sniffles_del_infos.txt
```

```
$ paste sniffles_del_start.txt sniffles_del_stop.txt sniffles_del_infos.txt \
    > sniffles_del.csv
```

```
$ rm sniffles_del_start.txt sniffles_del_stop.txt sniffles_del_infos.txt
```

Comparaison des résultats de Delly et Sniffles

```
$ cd ~/tp_sv
$ cat delly/delly_del.csv
$ cat sniffles/sniffles_del.csv
```
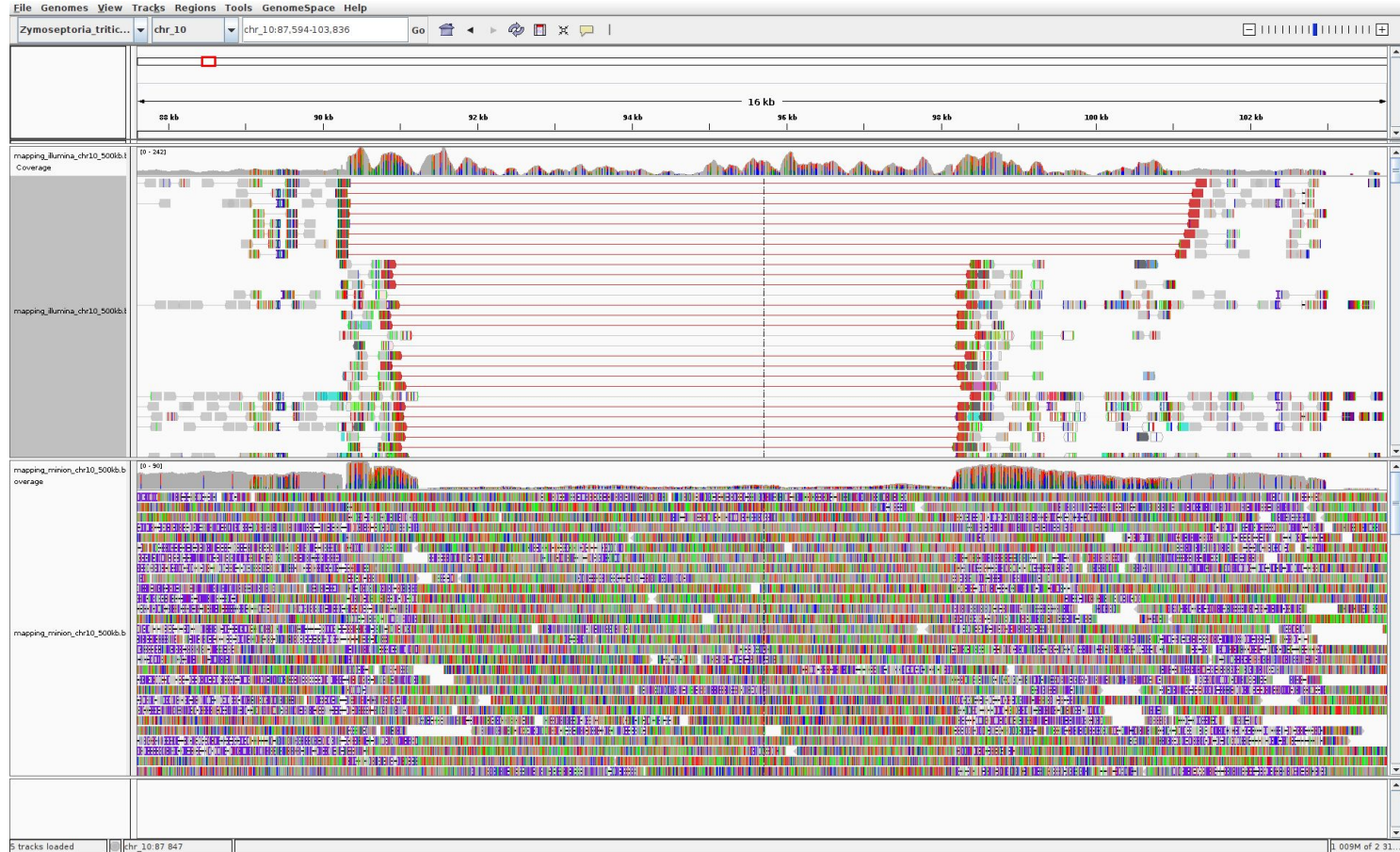
Delly (Illumina)				
Start	End	precision	PairEnd	Split Reads
29522	29580	PRECISE	0	20
32733	32783	PRECISE	0	20
57127	57600	PRECISE	3	16
80015	80622	PRECISE	15	20
90255	90309	PRECISE	0	7
90309	101040	IMPRECISE	8	0
111021	111676	IMPRECISE	20	0
191291	191343	PRECISE	0	20
-	-	-	-	-
264986	265063	PRECISE	0	12
267829	267857	PRECISE	0	19
-	-	-	-	-
360628	361052	PRECISE	0	20
383682	477911	IMPRECISE	7	0
425686	426624	IMPRECISE	28	0
459094	459124	PRECISE	0	12
465858	466080	PRECISE	0	20
468192	468342	PRECISE	0	20
477523	479732	PRECISE	41	20
496882	496919	PRECISE	0	20

| Sniffles (Minion)		
Start	End	precision
-	-	-
-	-	-
57126	57598	IMPRECISE
-	-	-
-	-	-
91233	98159	IMPRECISE
111020	111655	PRECISE
-	-	-
257001	257165	IMPRECISE
-	-	-
-	-	-
343161	343273	PRECISE
360638	361061	PRECISE
383681	477805	IMPRECISE
425682	426487	IMPRECISE
-	-	-
-	-	-
468192	468341	PRECISE
477525	479731	PRECISE
-	-	-

27

Visualisation sous IGV

- Chargement du génome de référence

  → ~/tp_sv/Zymoseptoria_tritici.fa

- Ouvrez à partir d'un fichier les fichiers BAM correspondant aux deux analyses (short et long reads) :

  → ~/tp_sv/mapping_illumina_chr10_500kb.bam
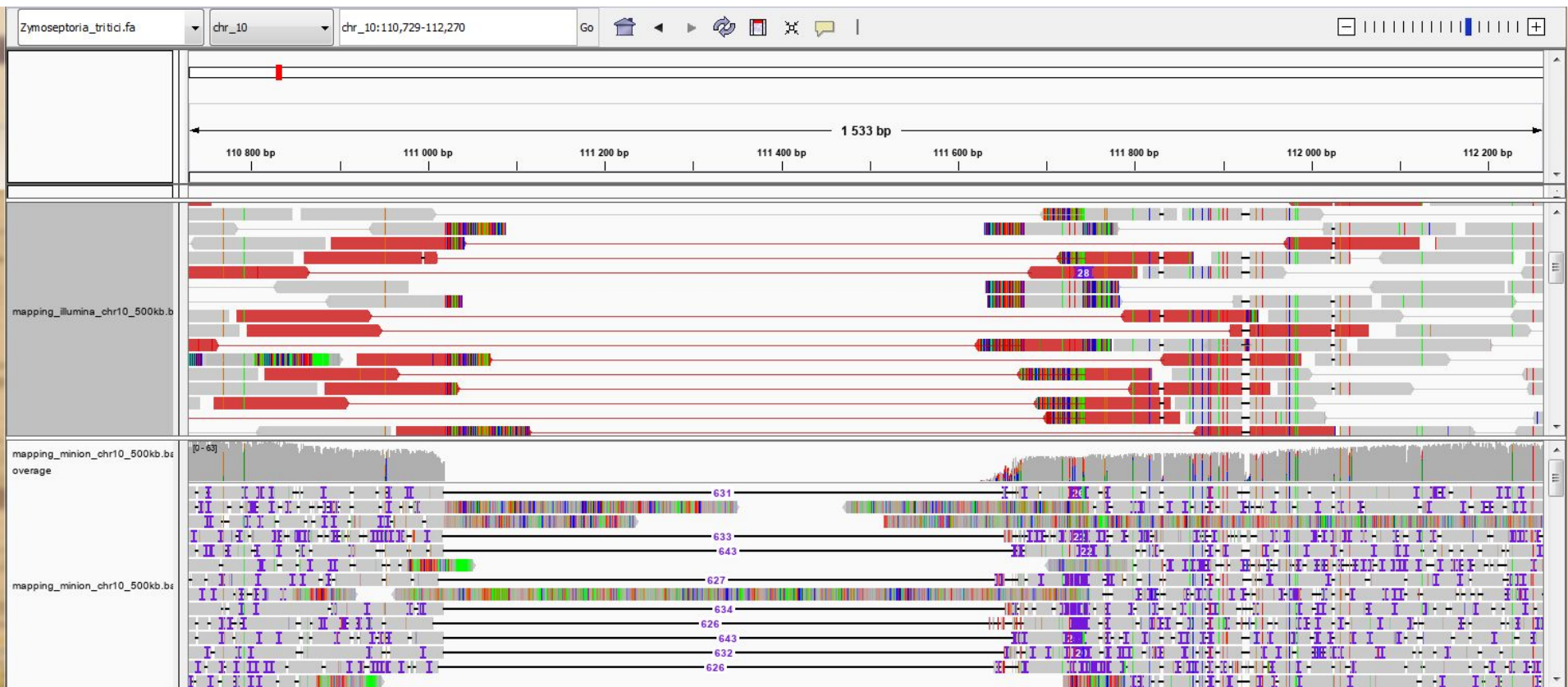 → ~/tp_sv/mapping_minion_chr10_500kb.bam

deletion 90309-101040 (illumina), 91233-98159 (Minion)

deletion 90309-101040 (illumina), 91233-98159 (Minion)
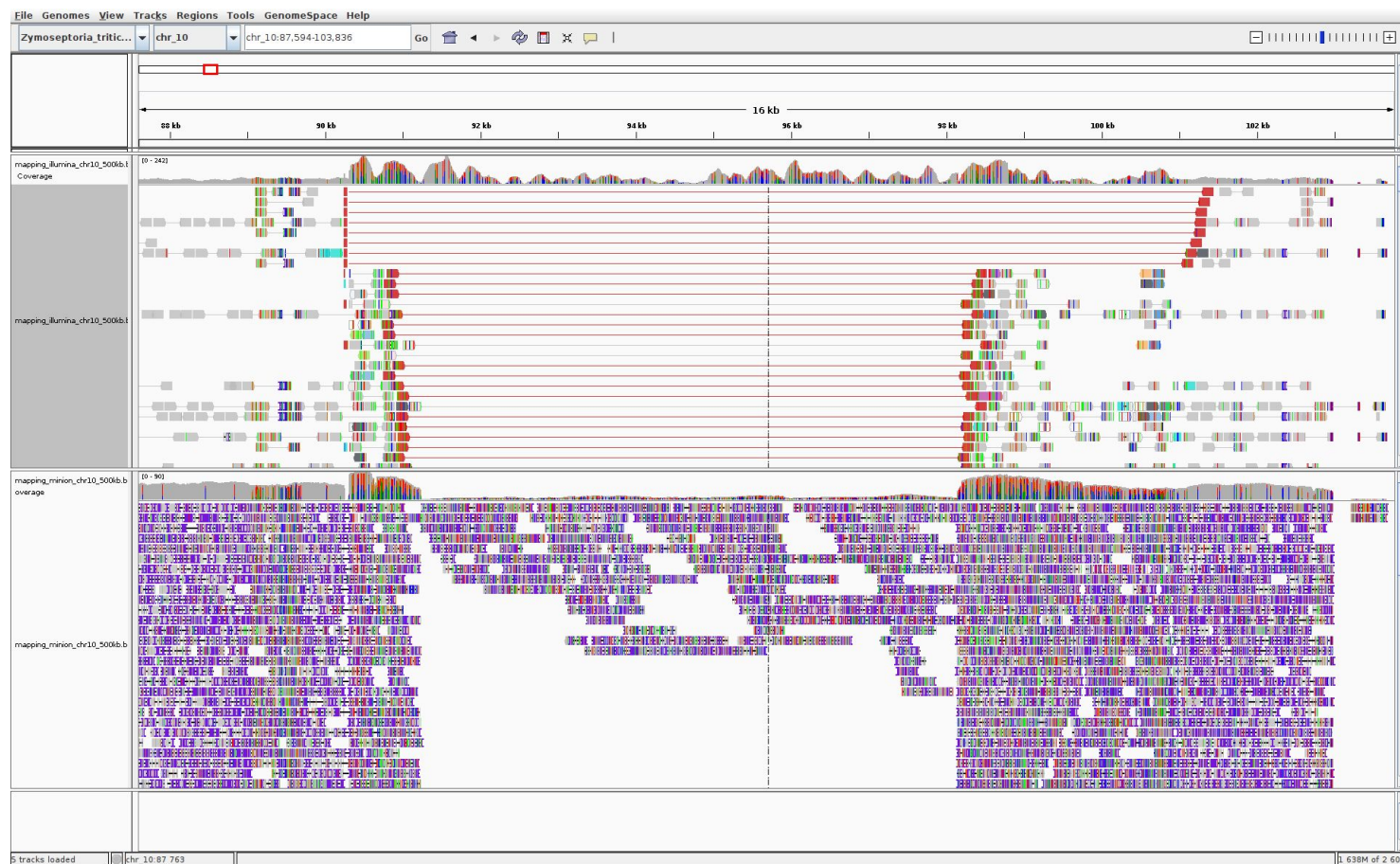
deletion 111021-111676

deletion 191291-191343

deletion 343161-343273

Comparaison des résultats de Delly et Sniffles

Delly (illumina)				
start	stop	precision	PE	SR
29522	29580	PRECISE	0	20
57127	57600	PRECISE	3	16
80015	80622	PRECISE	15	20
90255	90309	PRECISE	0	7
90309	101040	IMPRECISE	8	0
111021	111676	IMPRECISE	20	0
191291	191343	PRECISE	0	18
-	-	-	-	-
264986	265063	PRECISE	0	12
-	-	-	-	-
360628	361052	PRECISE	0	20
383682	477911	IMPRECISE	7	0
425686	426624	IMPRECISE	28	0
465858	466080	PRECISE	0	20
468192	468342	PRECISE	0	20
477523	479732	PRECISE	0	20
477526	479732	IMPRECISE	41	0

Sniffles (Minion)		
start	stop	precision
-	-	-
57126	57598	IMPRECISE
-	-	-
-	-	-
91233	98159	IMPRECISE
111020	111655	PRECISE
-	-	-
257001	257165	IMPRECISE
-	-	-
343161	343273	PRECISE
360638	361061	PRECISE
383681	477805	IMPRECISE
425682	426487	IMPRECISE
-	-	-
468192	468341	PRECISE
477525	479731	PRECISE
-	-	-

IGV OK

IGV ~OK

IGV doubt

IGV NO

Conclusion

- La détection des SVs manque de précision et engendre des faux positifs et faux négatifs

    → **Nécessité de croiser différents outils/technologies**

    → **Nécessité de bien utiliser les métriques des outils**

    → **Nécessité d'une bonne profondeur (variant hétérozygote)**

- Vérifier visuellement les résultats sur IGV permet d'augmenter la confiance dans les SVs détectés