ifb

aviesan
alliance nationale
pour les sciences de la vie et de la santé

Structural Variant detection

Gabrièle Adam - INRAE

École de bioinformatique AVIESAN-IFB-INSERM 2025

Définition

- Consensus actuel : Réarrangement génomique >50bp

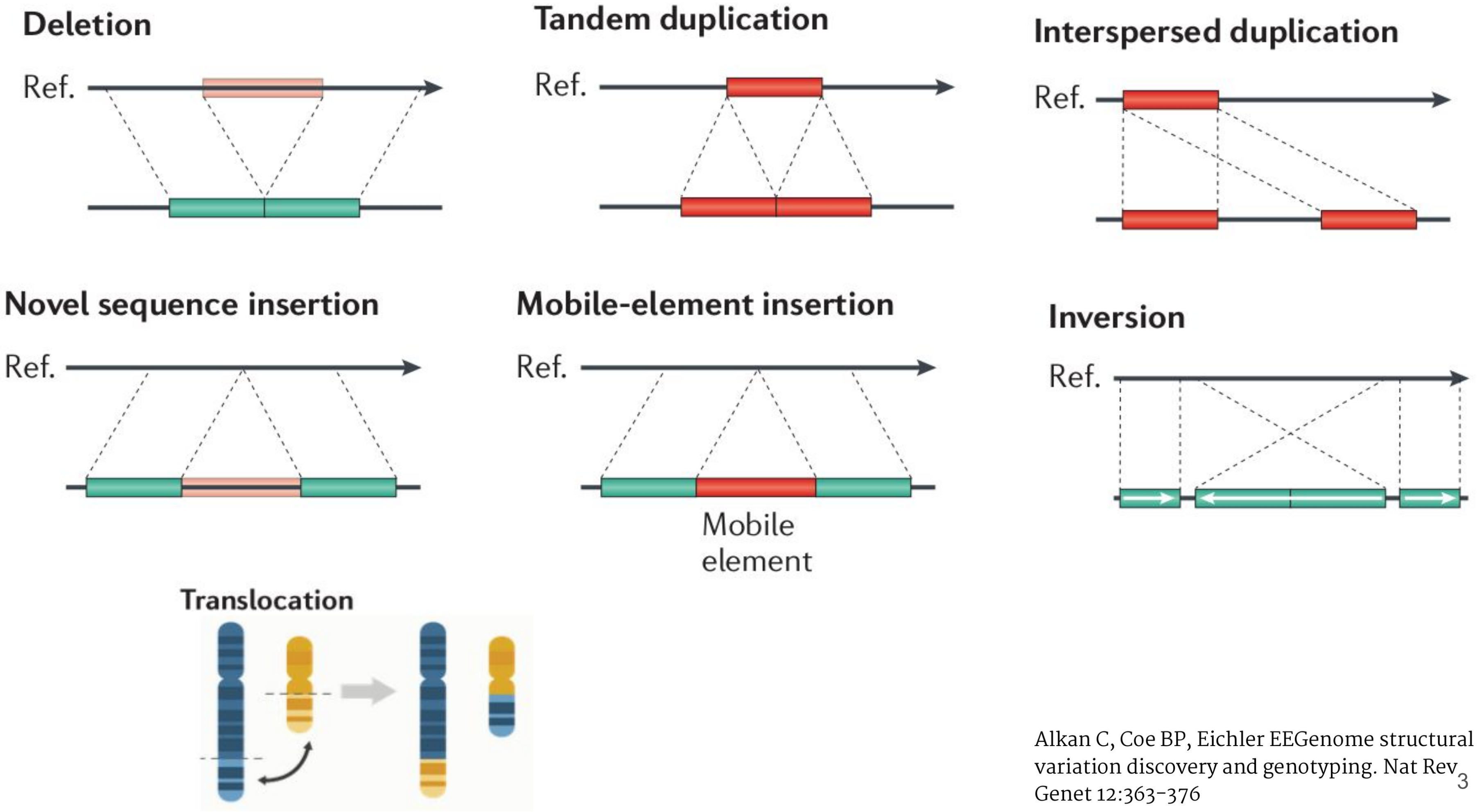
- Différents types de variants structuraux :

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

  - Délétion
  - Duplication

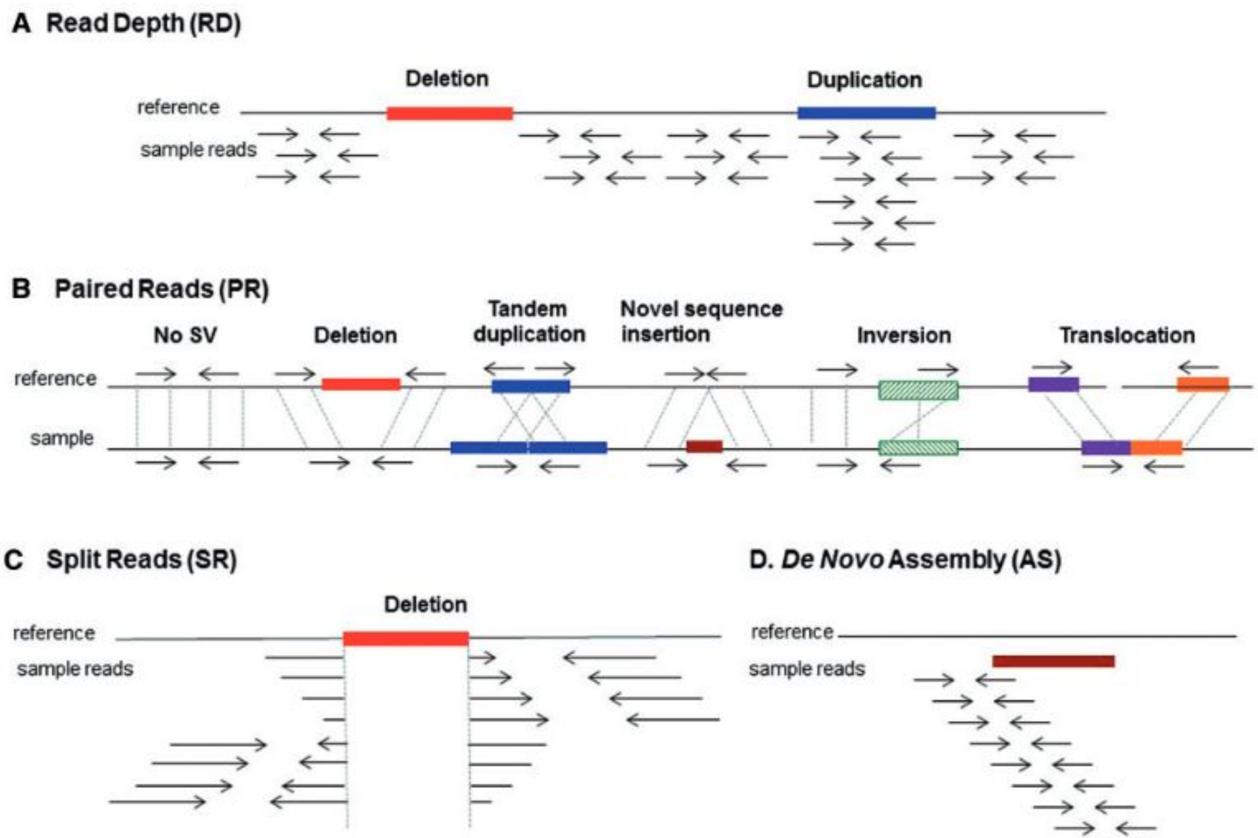
→ Réarrangements équilibrés

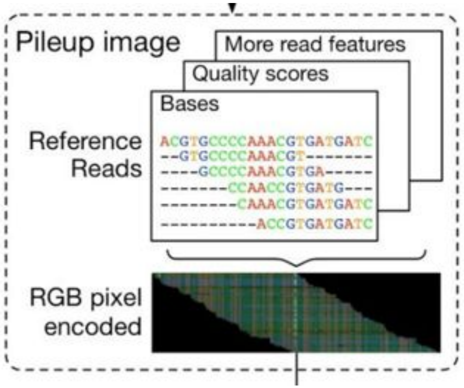
  - Insertion
  - Inversion
  - Translocation



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

Principe de détection des SVs





Et maintenant avec
des réseaux neuronaux !

Review > Brief Funct Genomics. 2015 Sep;14(5):305-14. doi: 10.1093/bfgp/elv014. Epub 2015 Apr 15.

A decade of structural variants: description, history and methods to detect structural variation

Geòrgia Escaramís, Elisa Docampo, Raquel Rabionet

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)

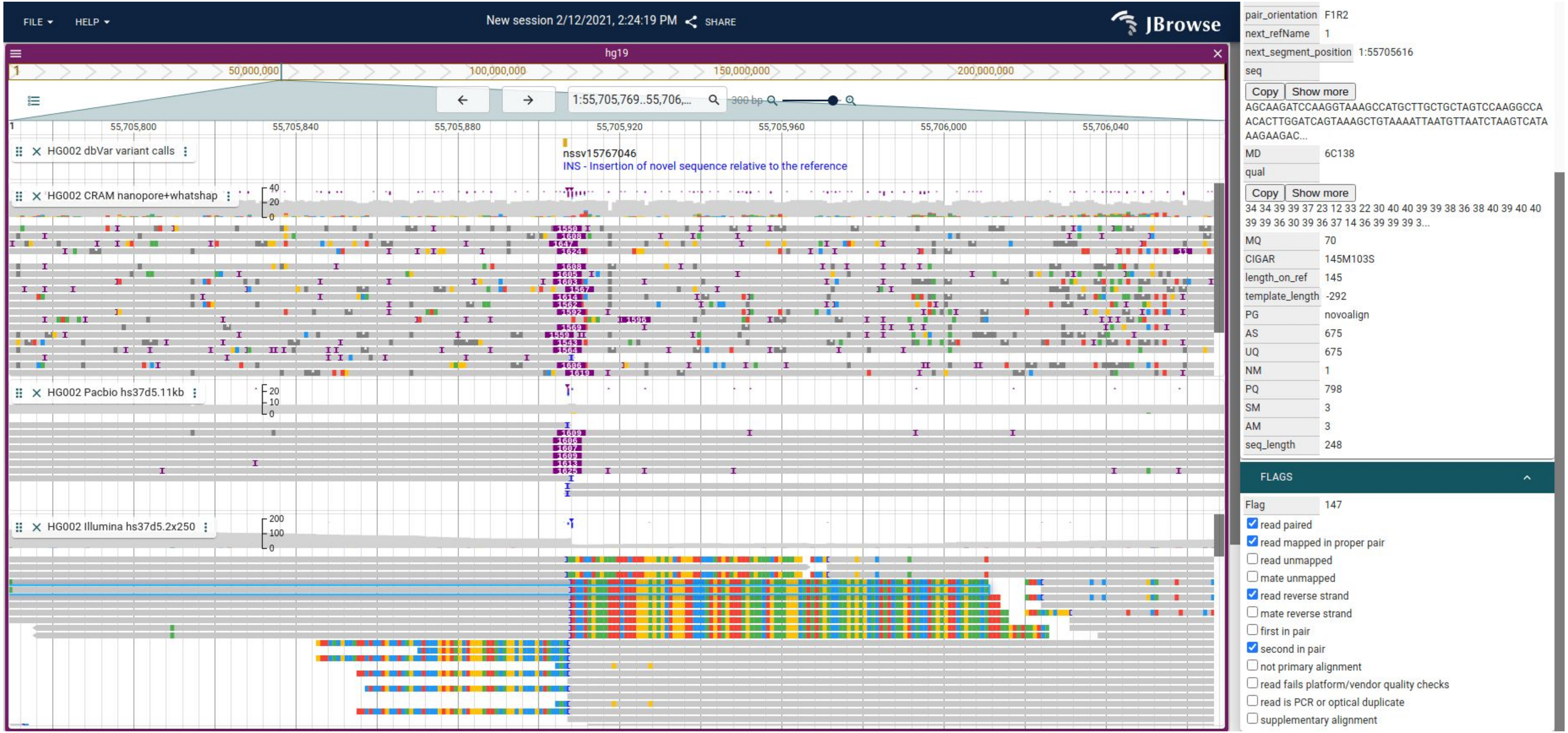
/!\ Un calling consensus avec plusieurs outils de détection peut être utile avec des données short reads /!\

Long reads (PacBio/MinION) :

→ Meilleure caractérisation des altérations des régions répétées

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

Sedlazeck, F. J. et al.Nat. Methods 15, 461–468 (2018).



pair_orientation	F1R2
next_refName	1
next_segment_position	1:55705616
seq	

Copy Show more

AGCAAGATCCAAGGTAAAGCCATGCTTGCTGCTAGTCCAAGGCCA
ACACTTGGATCAGTAAAGCTGTAAAATTAATGTTAATCTAAGTCATA
AAGAAGAC...

MD	6C138
qual	

Copy Show more

34 34 39 39 37 23 12 33 22 30 40 40 39 39 38 36 38 40 39 40 40
39 39 36 30 39 36 37 14 36 39 39 39 3...

MQ	70
CIGAR	145M103S
length_on_ref	145
template_length	-292
PG	novoalign
AS	675
UQ	675
NM	1
PQ	798
SM	3
AM	3
seq_length	248

FLAGS

Flag	147

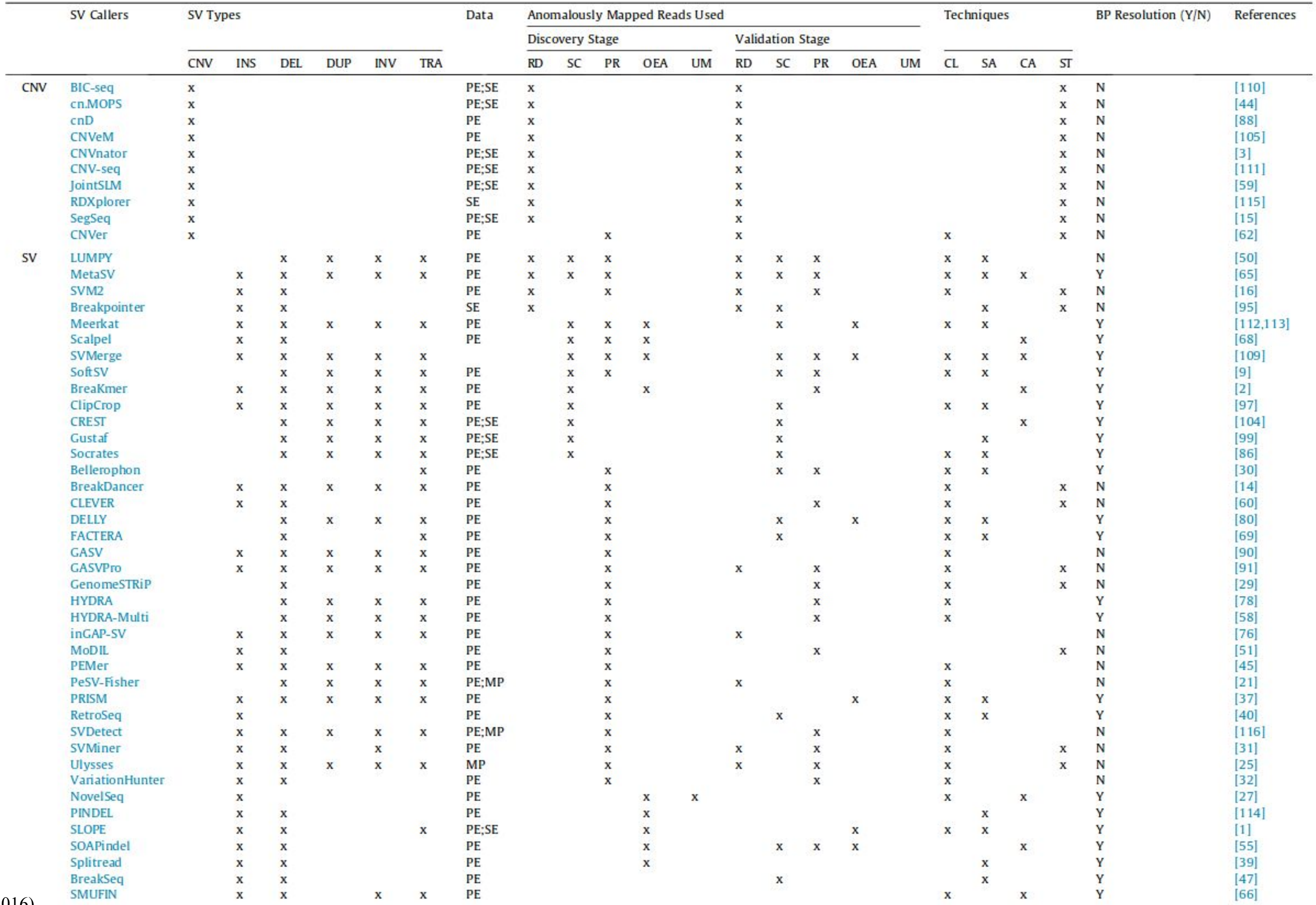
☑ read paired
☑ read mapped in proper pair
☐ read unmapped
☐ mate unmapped
☑ read reverse strand
☐ mate reverse strand
☐ first in pair
☑ second in pair
☐ not primary alignment
☐ read fails platform/vendor quality checks
☐ read is PCR or optical duplicate
☐ supplementary alignment

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	SV Types						Data	Anomalously Mapped Reads Used										Techniques				BP Resolution (Y/N)	References
									Discovery Stage					Validation Stage										
		CNV	INS	DEL	DUP	INV	TRA		RD	SC	PR	OEA	UM	RD	SC	PR	OEA	UM	CL	SA	CA	ST		
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	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	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	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		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]

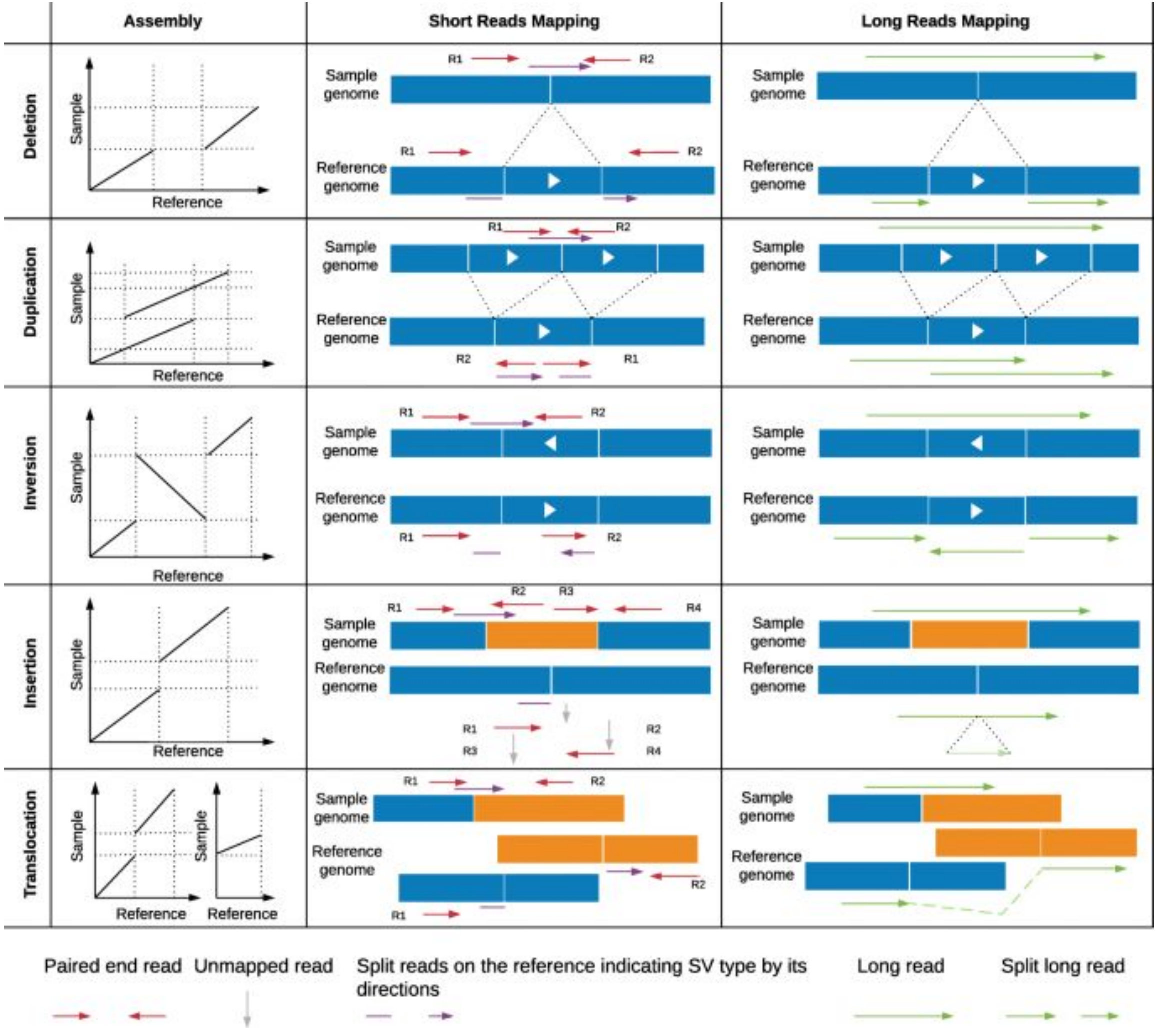
P. Guan, W.-K. Sung / Methods 102 (2016) 36–49

Peiyong Guan, , Methods 102 (2016)

Outils en long reads

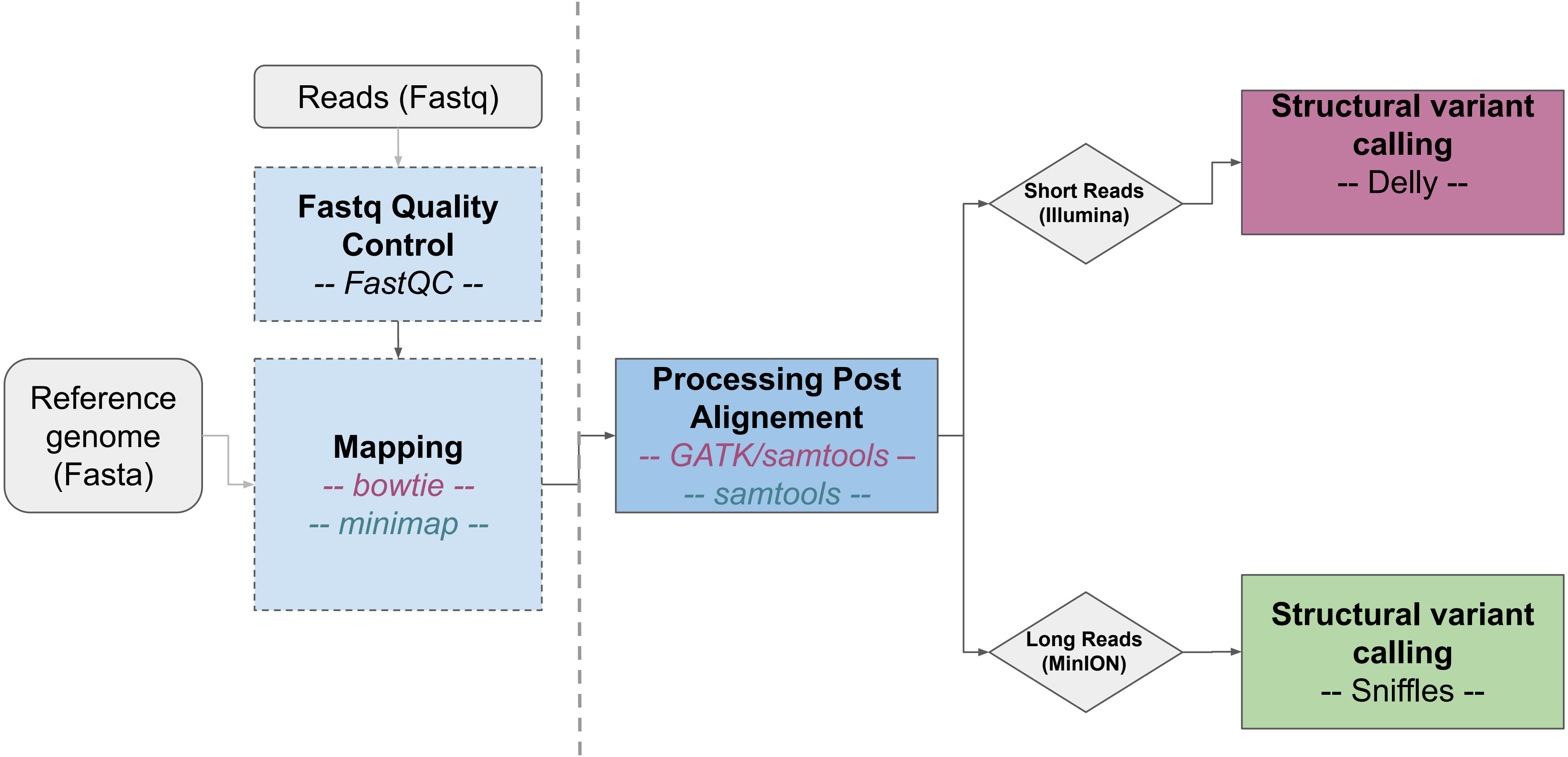
Outils	Read type	Variant type	Auteurs
DeepVariant	short/long	SNV/indel	Poplin et al.
NanoCaller	long	SNV/indel	Ahsan et al
PEPPERa	long	SNV/indel	Shafin et al
cuteSV	long	SV/indel	Jiang et al.
Dysgu	short/long	SV/indel	Cleal et al.
pbsv	long	SV/indel	PacBio SMRT Linkb
Sniffles	long	SV/indel	Sedlazeck et al.
SVDSS	long	SV/indel	Denti et al.
SVIM	long	SV/indel	Heller and Vingron
Deep SV	long	SV/indel	Cai et al.
Hysa	short/long	SV/indel	Fan et al
NanoSV	long	SV/indel	Euskirchen et al.
PBHoney	long	SV/indel	English et al.

A quoi vont ressembler les SVs dans les données short et long reads ?



Mahmoud et al, Genome Biology 2019

Workflow



Reads (Fastq)

Fastq Quality Control
-- FastQC --

Reference genome (Fasta)

Mapping
-- bowtie --
-- minimap --

Processing Post Alignement
-- GATK/samtools –
-- samtools --

Structural variant calling
-- Delly --

Structural variant calling
-- Sniffles --

Rappel Mapping

-> Short Reads

```
bowtie2 --threads 4  --very-sensitive --no-unal -x genomeRef -1 R1.fq.gz -2 R2. fq.gz -S output.sam
```

->Long Reads

```
minimap2 -t 4 -ayYL --MD --eqx -x asm20 Ref.fa subreads > output.sam
```

```
# a output in sam format
# -Y  use soft clipping for supplementary alignments
#  -L  write CIGAR with >65535 ops at the CG tag
#  -MD output the MD tag
#   --eqx   write =/X CIGAR operator
#asm20Use this if the average divergence is around several percent.
```

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

Jeux de données #2 : SVs

Zymoseptoria tritici : Champignon ascomycète, pathogène du blé tendre, responsable d’une maladie foliaire (septoriose).

- Principale maladie du blé (jusqu’à 50% de perte de rendement).
- Haploïde, génome de 40 Mb séquencé en 2011 : 13 chromosomes essentiels + 8 chromosomes accessoires
- Souche séquencée avec **deux technologies : Illumina et MinIon**

Your turn !

Retrouvez les délétions de grande taille

Aller au jupyterNoteBook

Visualisation sous IGV (Bonus)

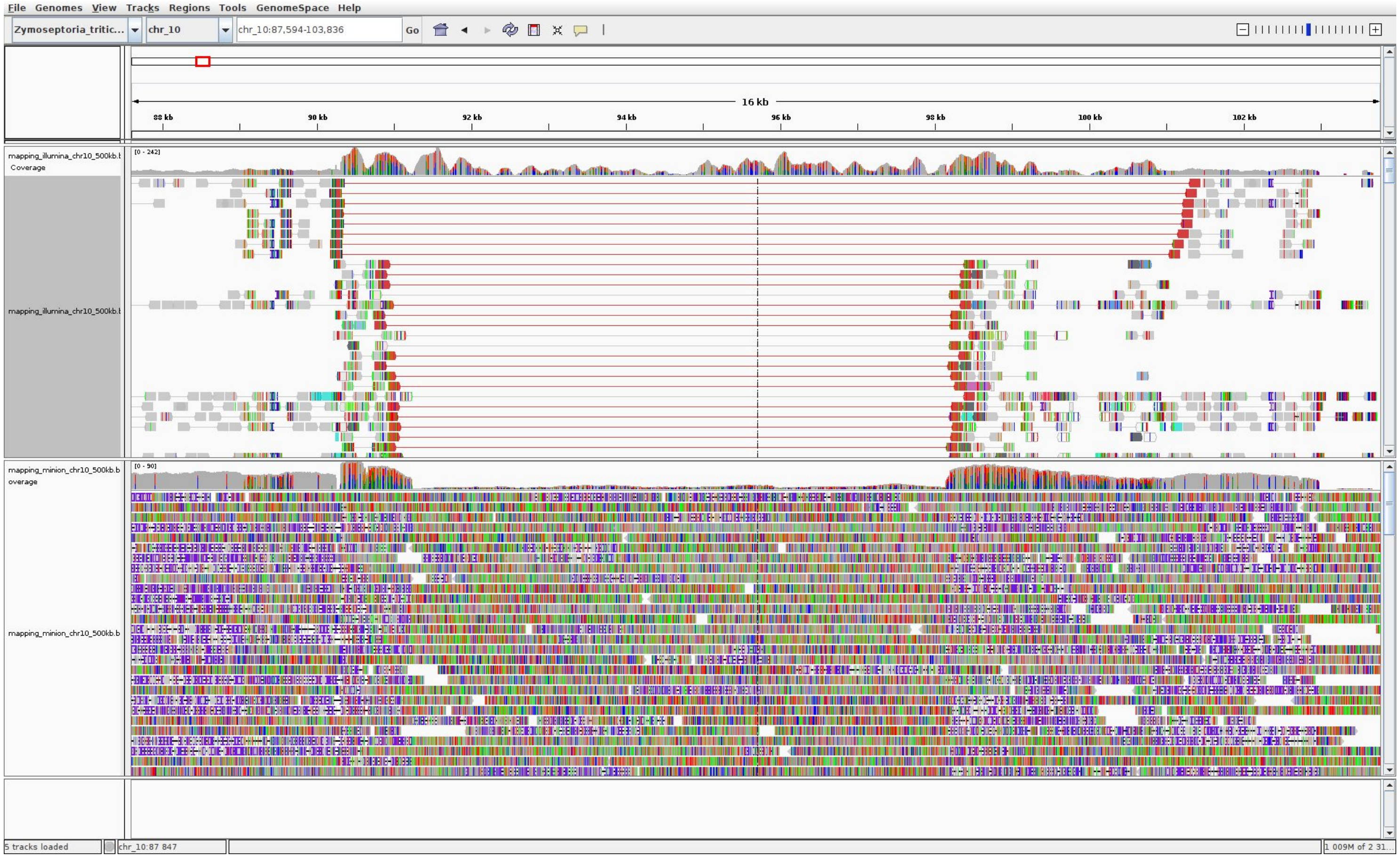
- Télécharger en local les fichiers BAM et leurs index à travers votre session Jupyter

  → `Zymoseptoria_tritici.fa/fai`
  → mapping_illumina_chr10_500kb.bam/bam.bai
  → mapping_minion_chr10_500kb.bam/bam.bai

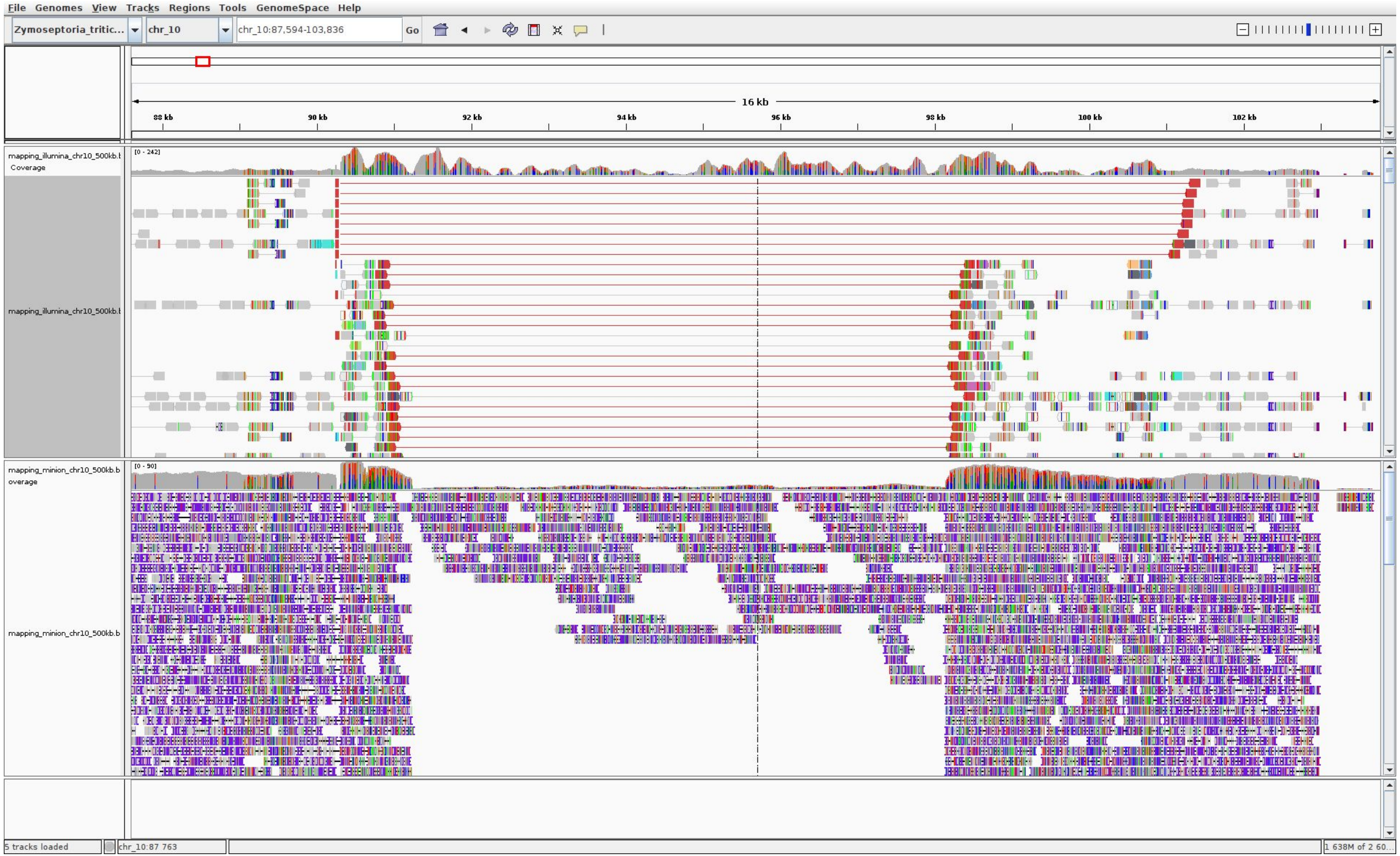
- Charger le génome de référence

- Ouvrir les fichiers BAM correspondant aux deux analyses (short et long reads)

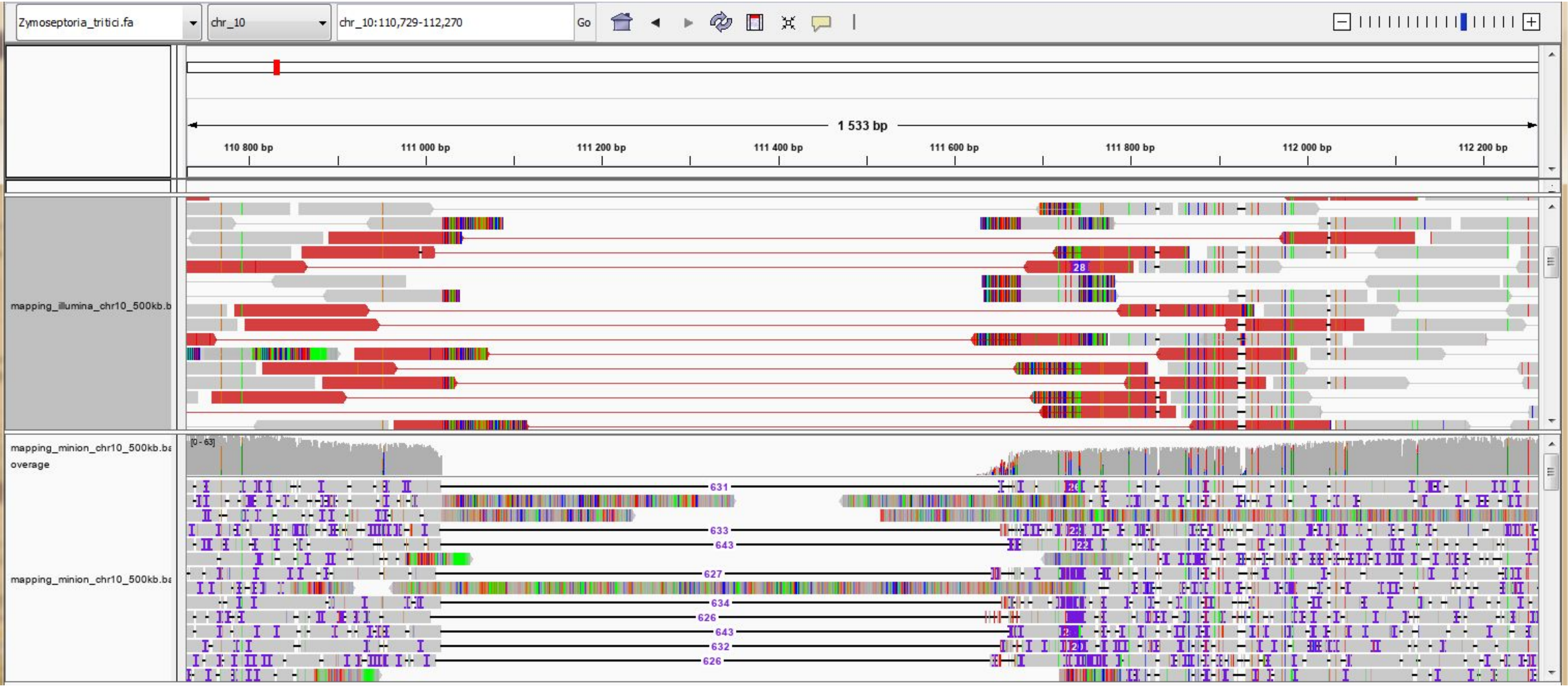
deletion 90309-101040 (illumina), 91233-98159 (Minion)


File Genomes View Tracks Regions Tools GenomeSpace Help
Zymoseptoria_tritic...
chr_10
chr_10:87,594-103,836
Go
16 kb
88 kb
90 kb
92 kb
94 kb
96 kb
98 kb
100 kb
102 kb
mapping_illumina_chr10_500kb.t Coverage
[0 - 242]
mapping_illumina_chr10_500kb.t
mapping_minion_chr10_500kb.b overage
[0 - 90]
mapping_minion_chr10_500kb.b
5 tracks loaded
chr_10:87 847
1 009M of 2 31...

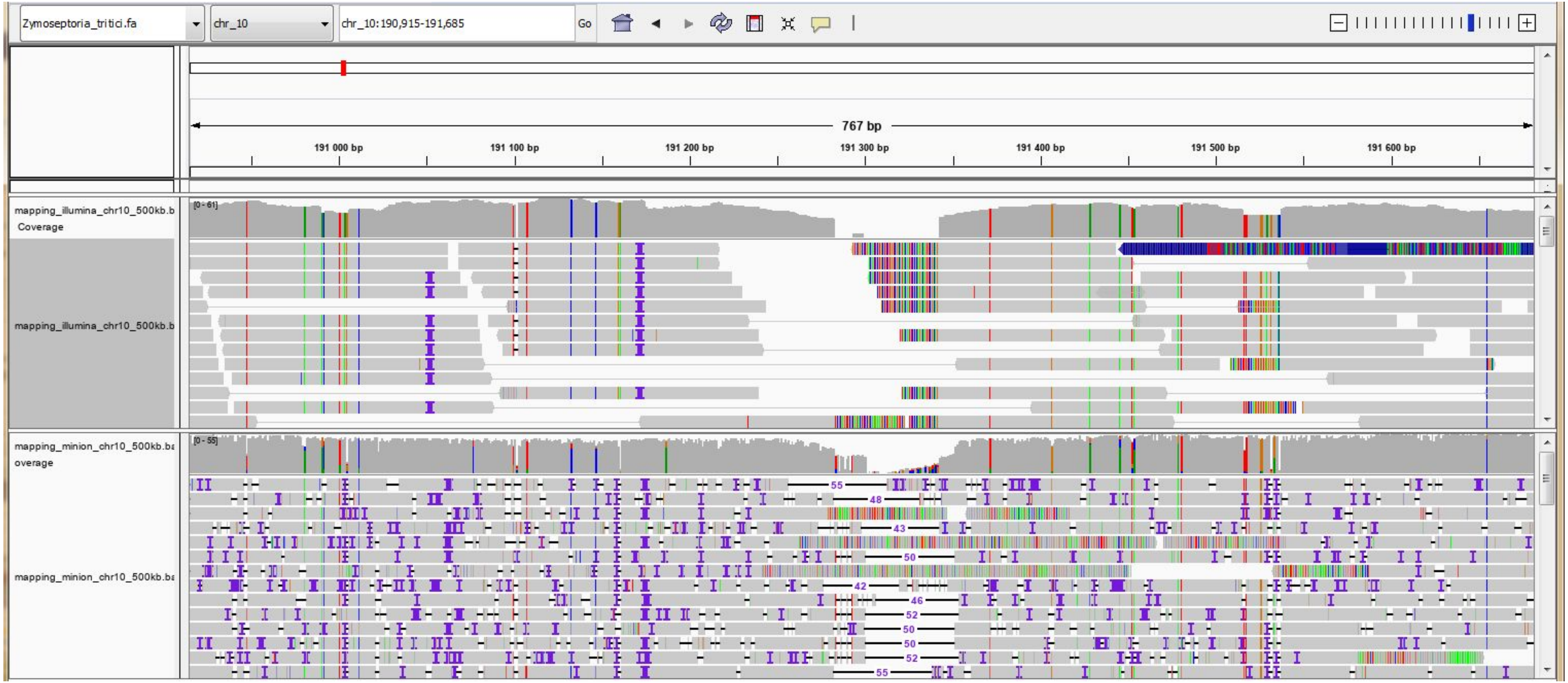
deletion 90309-101040 (illumina), 91233-98159 (Minion)


File Genomes View Tracks Regions Tools GenomeSpace Help
Zymoseptoria_tritic...
chr_10
chr_10:87,594-103,836
Go
16 kb
88 kb
90 kb
92 kb
94 kb
96 kb
98 kb
100 kb
102 kb
mapping_illumina_chr10_500kb.t Coverage
[0 - 242]
mapping_illumina_chr10_500kb.t
mapping_minion_chr10_500kb.b overage
[0 - 90]
mapping_minion_chr10_500kb.b
5 tracks loaded
chr_10:87 763
1 638M of 2 60...

deletion 111021-111676



deletion 191291-191343



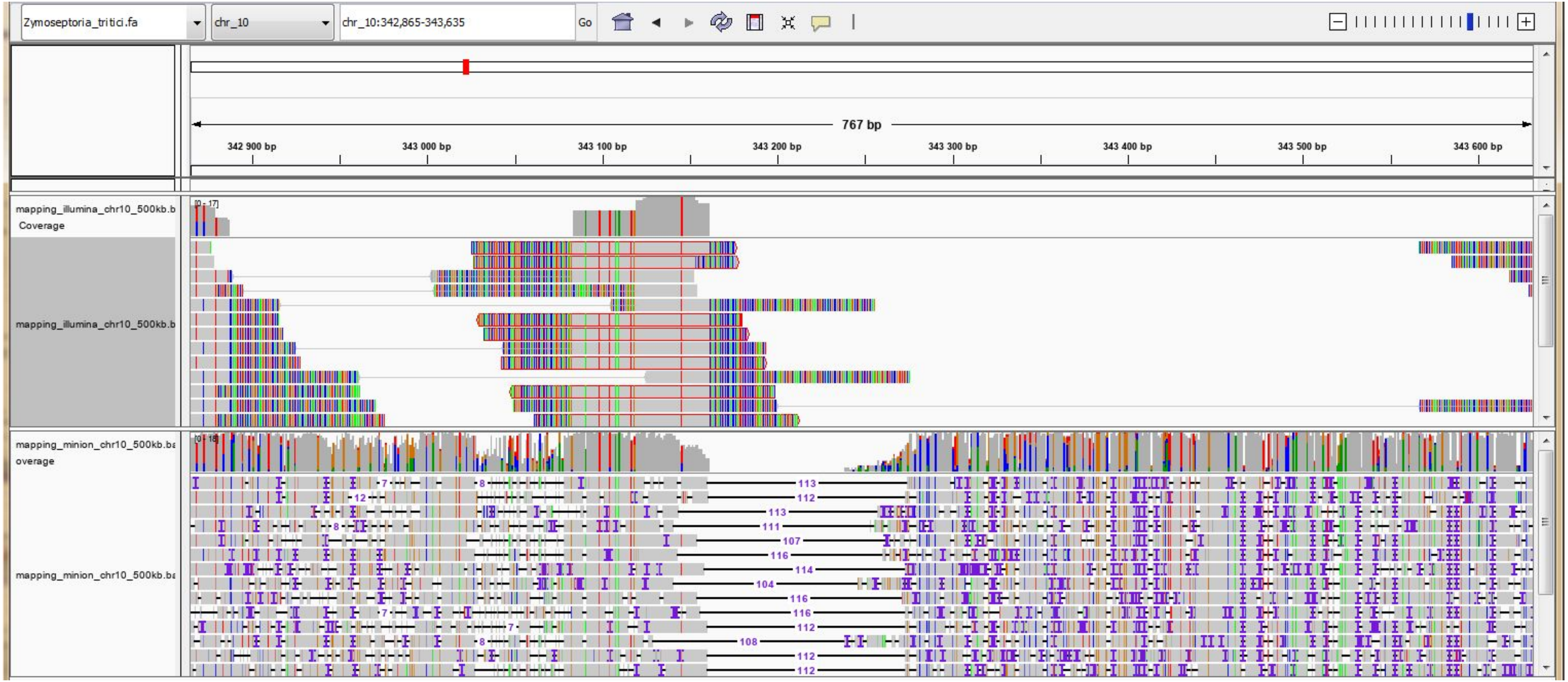
Zymoseptoria_tritici.fa

chr_10

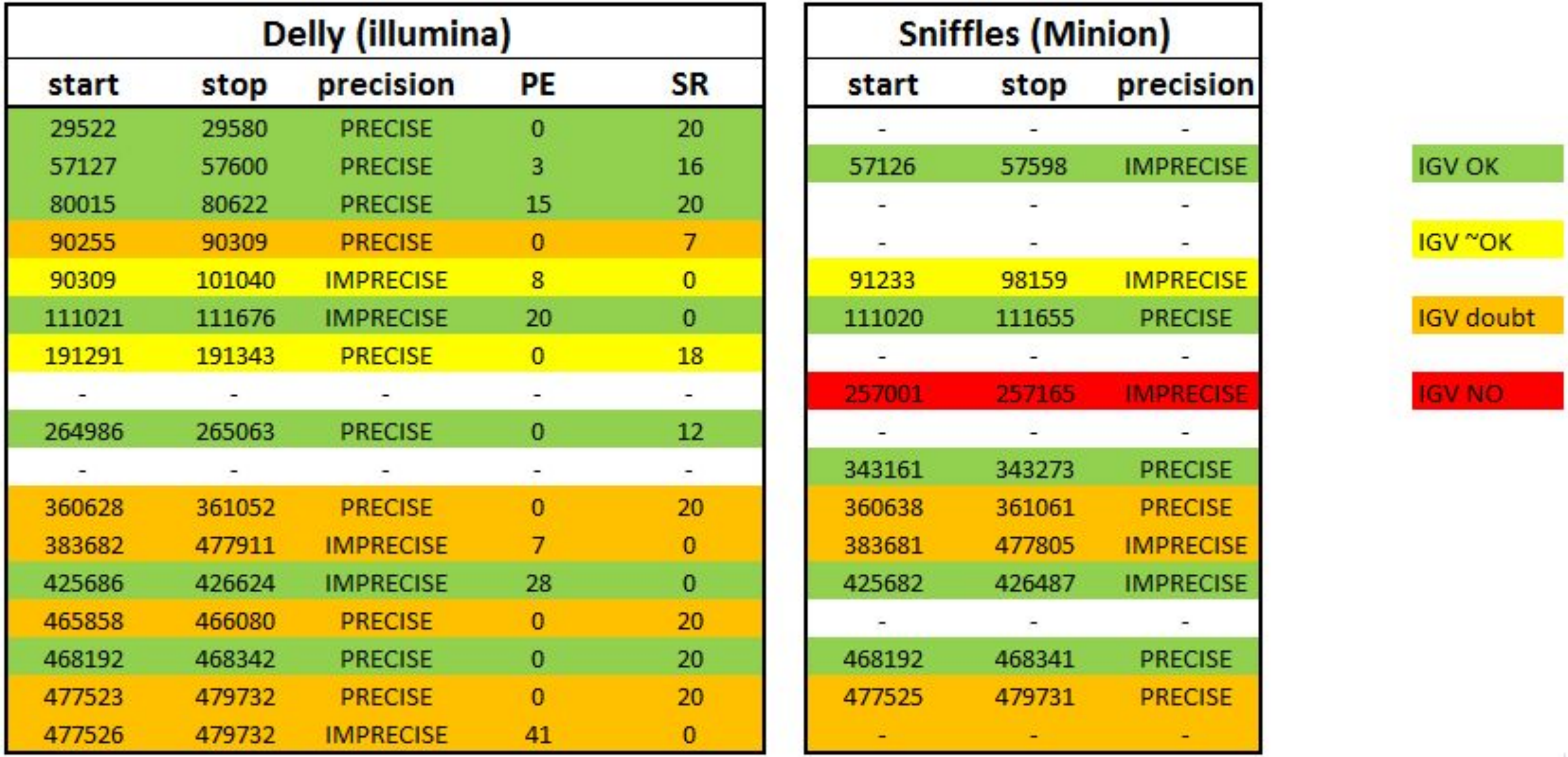
chr_10:190,915-191,685

Go
767 bp
191 000 bp
191 100 bp
191 200 bp
191 300 bp
191 400 bp
191 500 bp
191 600 bp
mapping_illumina_chr10_500kb.b Coverage
[0 - 61]
mapping_illumina_chr10_500kb.b
mapping_minion_chr10_500kb.ba overage
[0 - 55]
mapping_minion_chr10_500kb.ba
55
48
43
50
42
46
52
50
50
52
55

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