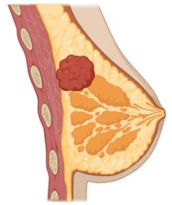
De l'ADN à l'ARN : Optimiser l'analyse des séquences longues avec des pipelines bioinformatiques performants

Camille AUCOUTURIER / Antoine CHOUTEAU

Laboratoire de biologie et de génétique du cancer · Centre François Baclesse (Caen, France)
INSERM U1245 - FHU G4 Genomic

Séminaire BioinfoDiag – 02/04/25

Contexte :



- Gènes de prédispositions aux cancers du **sein** et de **l'ovaire**

- Intérêt du séquençage *long read* :

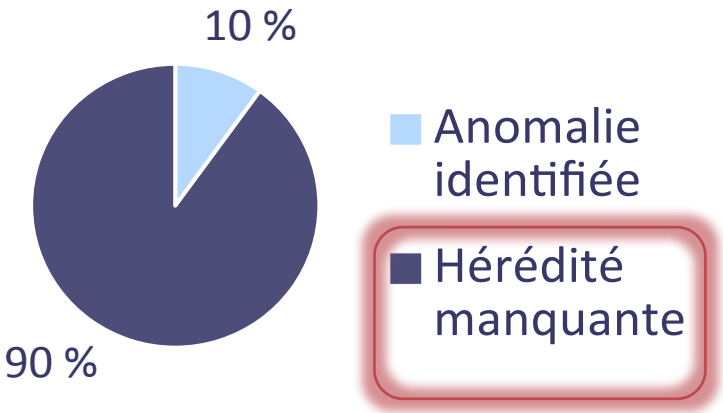
Syndrome HBOC :
Hereditary Breast and Ovarian Cancer

Exploration

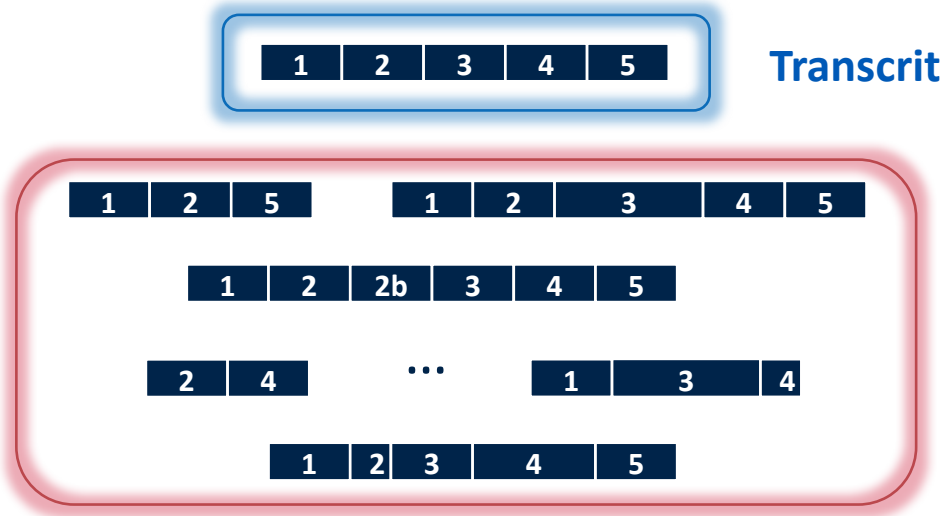
***DNAseq "short read"* :**
13 gènes diagnostiques
(*Moretta J et al, 2018*):

BRCA1 *BRCA2* *PALB2* *RAD51C*
RAD51D *MLH1* *MSH2* *MSH6*
PMS2 *TP53* *STK11* *PTEN*
EPCAM

Résultats



Diversité des transcrits ARNm :



Transcrits alternatifs :

- **Saut(s) d'exon(s)**
- **Utilisation de nvx sites d'épissage**
- **Rétention d'intron(s)**

Exploration des structures
Identification transcrits aberrants

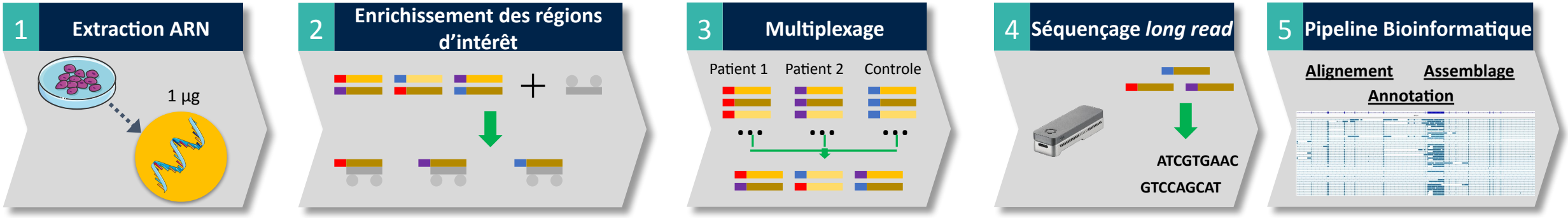
Troisième génération de séquençage: *long read*

- **Oxford Nanopore Technologies® (ONT)**

Obtenir un niveau suffisant de profondeur de couverture (gènes faiblement exprimés) :

Approche ciblée par capture : enrichissement

RNAseq *long read* :



Lignées lymphoblastoïdes

Panel 28 gènes :

ATM	*BAP1*	*BRCA1*
BRCA2	*BRIP1*	*CDH1*
CHEK2	*MLH1*	*MLH3*
MRE11	*MSH2*	*MSH6*
MUTYH	*NBN*	*PALB2*
PMS1	*PMS2*	*PTEN*
RAD50	*RAD51*	*RAD51B*
RAD51C	*RAD51D*	*RINT1*
STK11	*TP53*	*XRCC2*
XRCC3		

Groupes :

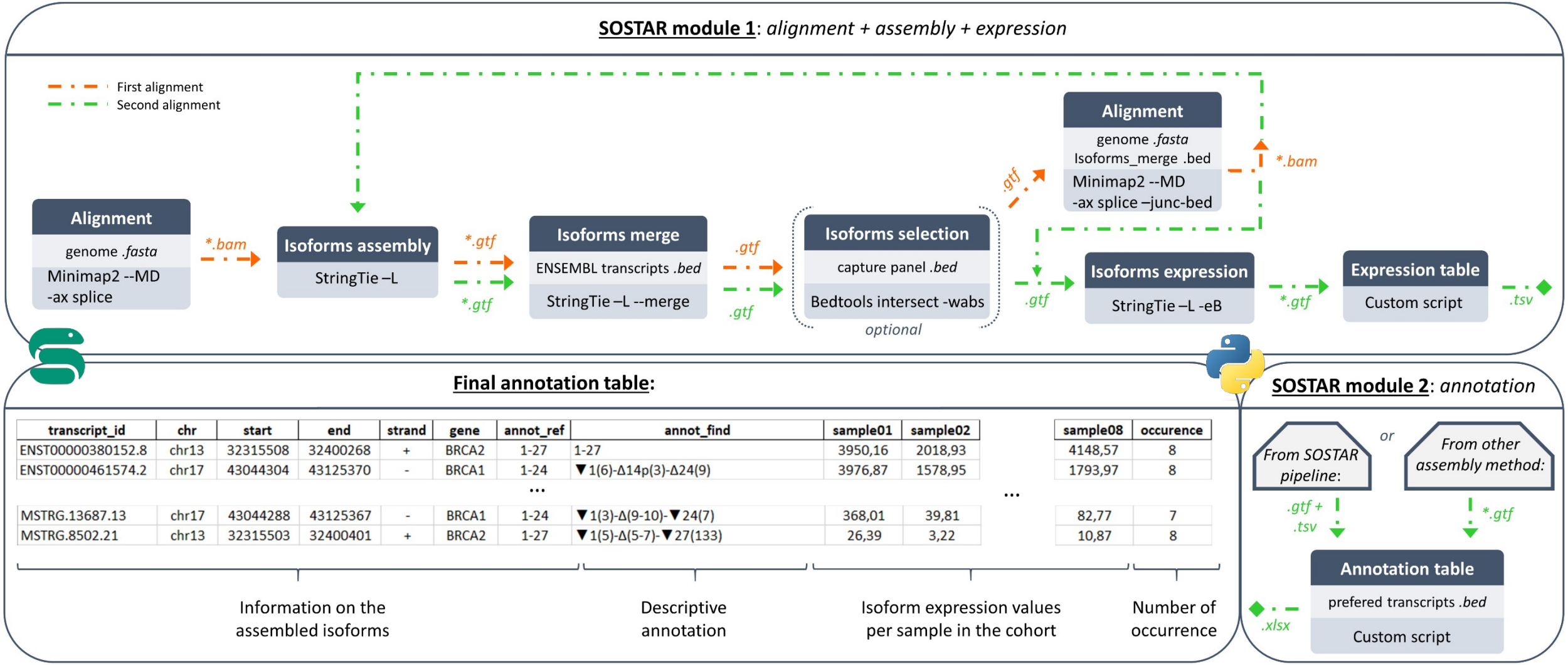
- **4 donneurs sains :** CASOHAR (CAncer du Sein et/ou de l'Ovaire Héréditaire – ARN, N°ID-RCB 2015-A00598-41)
- **2 patients porteurs d'anomalies génétiques impactant l'épissage du gène *BRCA1***

MinION (*Oxford Nanopore Technologies ONT*)

Pipeline SOSTAR : *iSofOrmS annoTAtoR* pipeline

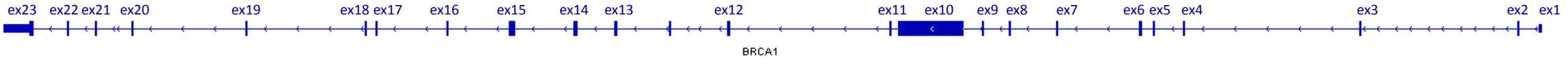
Disponible sur GitHub : https://github.com/LBGC-CFB/SOSTAR

Pipeline SOSTAR :



Module SOSTAR :

Transcrit « préféré » de référence :

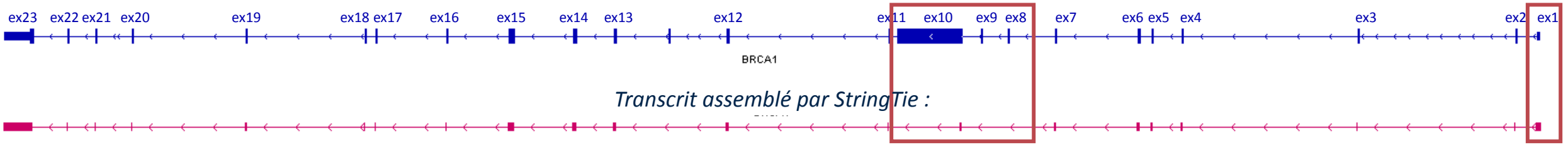


Transcrit assemblé par StringTie :

Comment annoter les différents transcrits assemblés par StringTie et non connus dans les bases de données actuelles ?

Module SOSTAR :

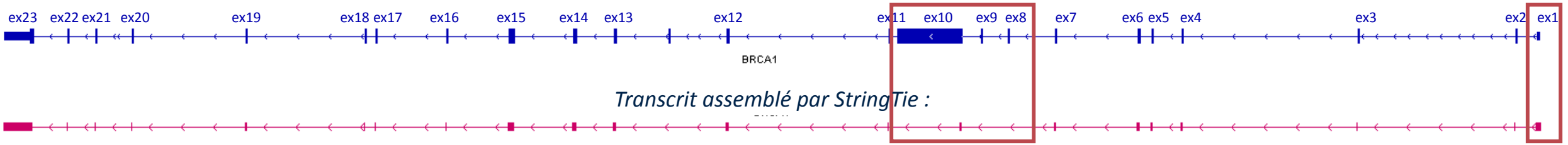
Transcrit « préféré » de référence :



Transcrit assemblé par StringTie :

Module SOSTAR :

Transcrit « préféré » de référence :



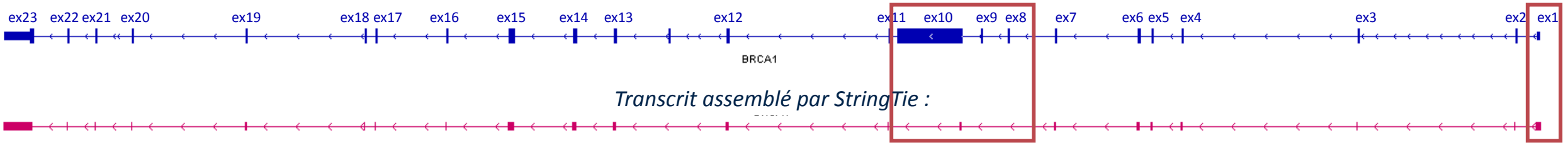
Transcrit assemblé par StringTie :

SOSTAR

▼1(119),▼1q(89)-Δ(8-9)-Δ10q(3309)

Module SOSTAR :

Transcrit « préféré » de référence :



Transcrit assemblé par StringTie :

SOSTAR

▼1(119),▼1q(89)-Δ(8-9)-Δ10q(3309)

- Nomenclature descriptive par rapport aux coordonnées d'un "transcrit préféré" défini (ex : *BRCA1 NM_007294)* :

  → inclut uniquement les **évènements d'épissage alternatif**

Δ = saut (*relatif à **l'exon***)	**int** = intronisation d'exon
▼ = insertion (*relatif à **l'intron***)	**exo** = exonisation d'intron
p = proximal (accepteur)	**(67)** = nb nucl
q = distal (donneur)	**[ ]** = intervalle d'évènement
, = événements discontinus	**327** = position
- = événements continus	

Algorithme SOSTAR :

Input

Output

SOSTAR : *iSofOrmS annoTAtoR*

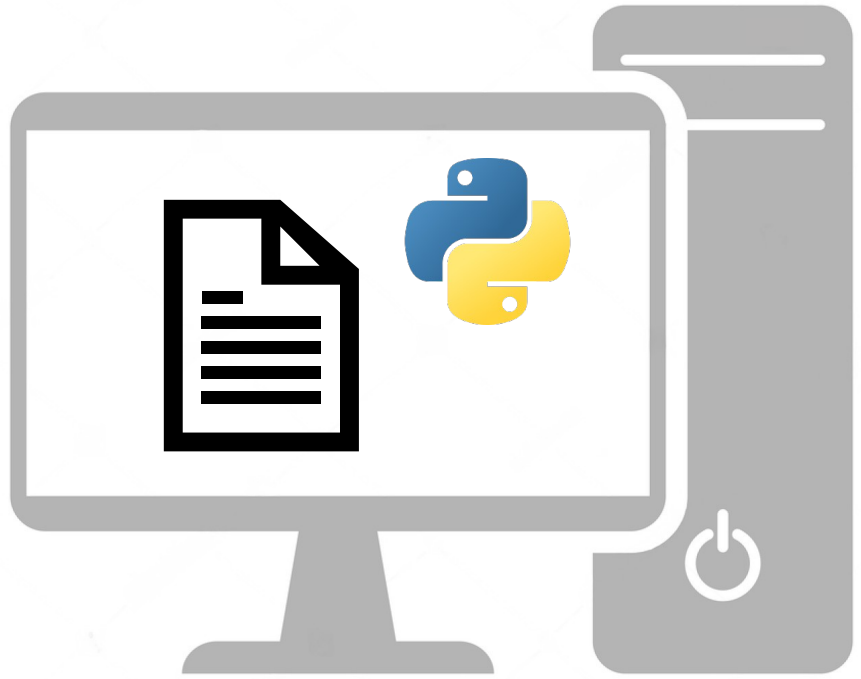
Fichier GTF

Isoformes de référence (MANE Select) pour chaque gène du panel

+

Fichiers GTF

Isoformes assemblées par StringTie pour chaque patient

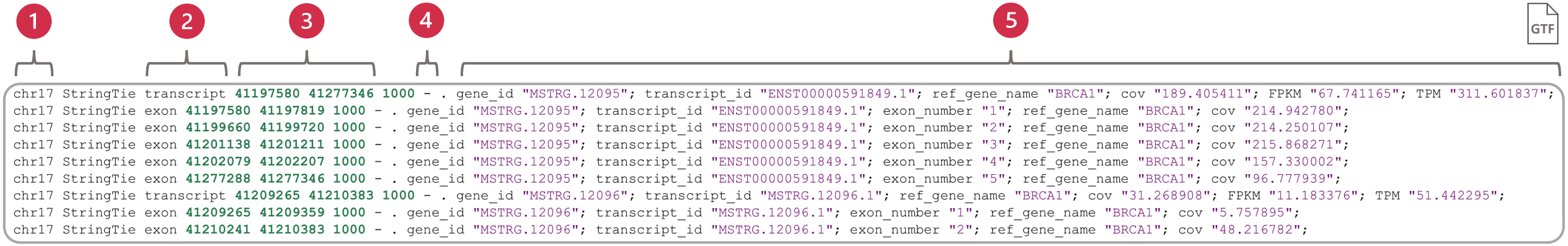


Script codé en langage Python

Tableau .xlsx

Fichier GTF : *General Transfer Format*

Fichier texte tabulé : description des structures des transcrits




```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
chr17 StringTie exon 41209265 41209359 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "5.757895";
chr17 StringTie exon 41210241 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "48.216782";
```

1. Chromosome
2. *Feature* : *transcript, exon, CDS, …*
3. Coordonnées start/end *feature*
4. Brin
5. Attributs
   - *Gene_id* : identifiant du locus génomique du transcrit
   - *Transcript_id* : identifiant du transcrit
   - *Ref_gene_name* : gène
   - *Cov, FPKM, TPM :* données d'expression du transcrit

Principales étapes script :

Lecture ligne à ligne des différents fichiers gtf :

Stockage coordonnées exons transcrits de références

- Initialisation du transcrit à annoter
- Annotation de chaque exon assemblé selon les règles définies
- Vérification si intronisation d'exon
- Vérification des délétions complètes d'exons
- Regroupement des évènements selon règles définies
- Génération du tableau final

Initialisation annotation transcrit :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268906"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
    list_exon_found = []
```

- Initialisation annotation : stockage dans liste = « list_exon_found »

```
dic_attr = {attr.split()[0]: attr.split()[1].replace('"',"") for attr in line.split("\t")[-1].split(";") if attr.split()}
if dic_attr["gene_name"] in dic_tr_ref.keys():
    gene = dic_attr["gene_name"]
    transcript_ID = dic_attr["transcript_id"]
    cov = dic_attr["cov"] if ("cov" in dic_attr.keys()) else 1
```

- Stockage des attributs : **gène / identifiant du transcrit / couverture**

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```python
if lines[2] == "exon":
    if lines[6] == "+":
        start_exon_stg = int("+"+lines[3]); end_exon_stg = int("+"+lines[4])
    elif lines[6] == "-":
        start_exon_stg = int("-"+lines[4]); end_exon_stg = int("-"+lines[3])
```

- **Stockage des coordonnées exon assemblés** → inversion des coordonnées (start/end) si brin antisens « - »

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```python
count = 0
for exon_ref in dic_tr_ref[gene]:
    count += 1
    if lines[6] == "+":
        start_exon_ref = int("+"+exon_ref[0]); end_exon_ref = int("+"+exon_ref[1]); exon_nb = int(exon_ref[2]); sign = "+"
    elif lines[6] == "-":
        start_exon_ref = int("-"+exon_ref[1]); end_exon_ref = int("-"+exon_ref[0]); exon_nb = int(exon_ref[2]); sign = "-"
```

- Recherche localisation évènement : **comparaison des coordonnées des exons du transcrit de référence et celles du transcrit assemblé par StringTie** → itérations sur la **liste des coordonnées des exons** du transcrit de référence

Tests logiques : if, else if, else

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

Annotation selon la nomenclature définie :

```
if start_exon_stg in range(start_exon_ref, end_exon_ref+1):
    #case exon == exon
    if (start_exon_stg == start_exon_ref) and (end_exon_stg == end_exon_ref):
        list_exon_found.append(f"{exon_nb}"); break
```

- Cas : coordonnées **exon transcrit StringTie = coordonnées exon transcrit référence**

```
    #case del start exon
    if (start_exon_stg > start_exon_ref) and (end_exon_stg == end_exon_ref):
        if exon_nb != 1:
            list_exon_found.append(f"Δ{exon_nb}p({start_exon_stg-start_exon_ref})"); break
        elif exon_nb == 1:
            list_exon_found.append(f"Δ{exon_nb}({start_exon_stg-start_exon_ref})"); break
```

- Cas **délétion début exon**

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```python
elif end_exon_stg in range(start_exon_ref, end_exon_ref+1):
    #case ins start exon
    if (start_exon_stg < start_exon_ref) and (end_exon_stg == end_exon_ref):
        if exon_nb != 1:
            list_exon_found.append(f"▼{exon_nb-1}p({start_exon_ref-start_exon_stg})"); break
        elif exon_nb == 1:
            list_exon_found.append(f"▼{exon_nb}({start_exon_ref-start_exon_stg})"); break
```

- Cas **insertion début exon**

```
...
```

11 tests logiques

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_exon_found = [Δ1(43),▼1q(89); 2; 3; 6; Δ7q(67)]

```
list_exon_found = check_intro(list_exon_found, dic_tr_ref, gene)
```

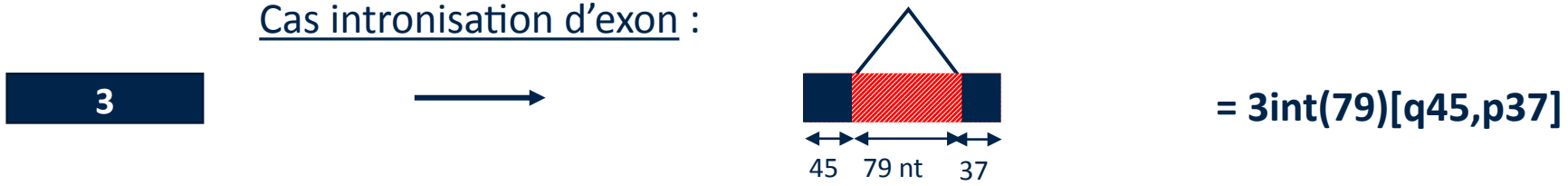
Vérification si présence d'une intronisation d'exon et modification en conséquence

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

Cas intronisation d'exon :



= 3int(79)[q45,p37]

- StringTie : 2 lignes « exon »
- list_exon_found = [Δ3q(116); Δ3p(124)]

```
list_exon_found = check_intro(list_exon_found, dic_tr_ref, gene)
```

Vérification si présence d'une intronisation d'exon et modification en conséquence

list_exon_found = [**3int(79)[q45,p37]**]

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_exon_found = [Δ1(43),▼1q(89); 2; 3; 6; Δ7q(67)]

```
list_annot_keep = [ex for ex in list_exon_found if not ex.isdigit() is True]
```

Nouvelle liste qui garde uniquement les anomalies

list_annot_keep = [Δ1(43),▼1q(89); Δ7q(67)]

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_exon_found = [Δ1(43),▼1q(89); 2; 3; 6; Δ7q(67)]

```
list_exon_nb_found = list(set([get_ex_nb(exon) for exon in list_exon_found if not ("exo" in exon)]))
```

Nouvelle liste regroupant les numéros d'exons (complets ou non) retrouvés

list_exon_nb_found = [1; 2; 3; 6; 7]

Présence uniquement des exons (*normaux ou non*) assemblés → délétions complètes d'exons ?

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_exon_nb_found = [1; 2; 3; 6; 7]

```
list_exon_nb_ref = [int(ex[-1]) for ex in dic_tr_ref[gene]]
```

Nouvelle liste correspondant aux numéros d'exons du transcrit de référence

list_exon_nb_ref = [1; 2; 3; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24]

```
list_del_exon = sorted([i for i in list_exon_nb_ref if i not in list_exon_nb_found])
```

Nouvelle liste correspondant aux exons délétés

list_del_exon = [5; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24]

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_del_exon = [5; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24]

```
list_annot_keep.extend("Δ" + str(ex) for ex in list_del_exon)
```

Ajout des exons délétés dans la liste des annotations à garder

list_annot_keep = [Δ1(43),▼1q(89); Δ6q(67); Δ5; Δ8; Δ9; Δ10; Δ11; Δ12; Δ13; Δ14; Δ15; Δ16; Δ17; Δ18; Δ19; Δ20; Δ21; Δ22; Δ23; Δ24]

```
list_annot_final = check_next(list_annot_sorted)
list_annot_final = "-".join(list_annot_final)
```

Regroupement des évènements selon la nomenclature

Annotation :

GTF

```
chr17 StringTie transcript 41197580 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; ref_gene_name "BRCA1"; cov "189.405411"; FPKM "67.741165"; TPM "311.601837";
chr17 StringTie exon 41197580 41197819 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "1"; ref_gene_name "BRCA1"; cov "214.942780";
chr17 StringTie exon 41199660 41199720 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "2"; ref_gene_name "BRCA1"; cov "214.250107";
chr17 StringTie exon 41201138 41201211 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "3"; ref_gene_name "BRCA1"; cov "215.868271";
chr17 StringTie exon 41202079 41202207 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "4"; ref_gene_name "BRCA1"; cov "157.330002";
chr17 StringTie exon 41277288 41277346 1000 - . gene_id "MSTRG.12095"; transcript_id "ENST00000591849.1"; exon_number "5"; ref_gene_name "BRCA1"; cov "96.777939";
chr17 StringTie transcript 41209265 41210383 1000 - . gene_id "MSTRG.12096"; transcript_id "MSTRG.12096.1"; ref_gene_name "BRCA1"; cov "31.268908"; FPKM "11.183376"; TPM "51.442295";
```

```
if (lines[2] == "transcript"):
```

- Nouveau transcrit → « **clôture » de l'annotation du transcrit précédent** :

list_annot_final = [Δ1(43),▼1q(89)-Δ5-Δ7q(67)-Δ(8-24)]

```
df_tmp = pd.DataFrame.from_dict({"transcript_id": [info[4]], "chr": [info[0]], "start": [info[1]], "end":
[info[2]], "strand": [info[3]], "gene": [gene], "annot_ref": [str(list_exon_nb_ref[0])
+"-"+str(list_exon_nb_ref[-1])], "annot_find": [list_annot_final], filename: [round(float(cov), 2)]})
```

Stockage dans le tableau final avec ajout des données d'expression

...

Ecriture dans fichier un excel .xlsx

Description des isoformes assemblées :

- 6 patients → 792 isoformes assemblées
- Majorité **d'isoformes partagées** par tous les patients (~ 63 %)
- **Isoformes uniques** exprimées par un seul patient (~ 10%)

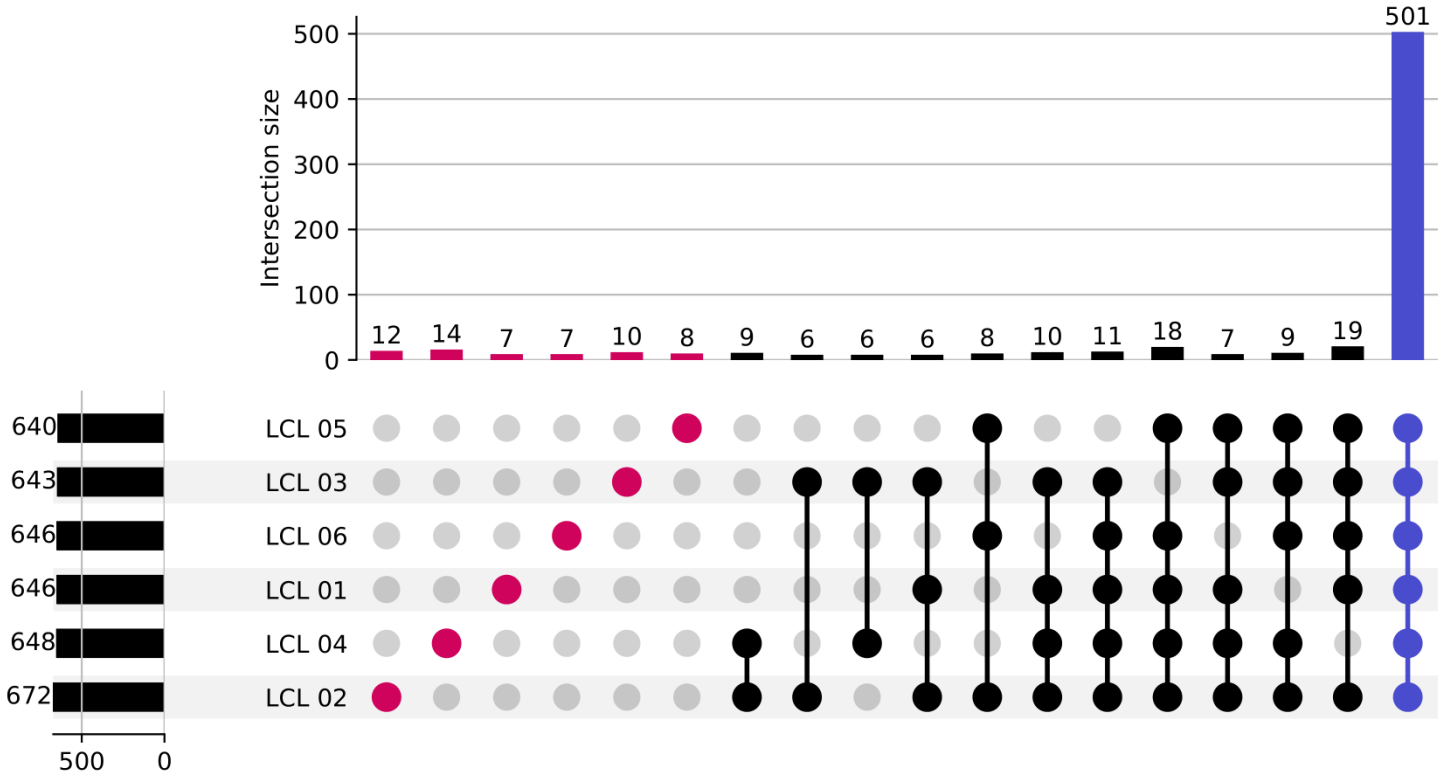


Tableau final généré :

```
python3 SOSTAR.py -I /mnt/data/auccam/test/expression -O /mnt/data/auccam/test -R
/mnt/data/auccam/References/ref_gencode_hg38_MANE_transcripts.gtf
```

< 1 min

Isoformes assemblées pour par gène

transcript_id	chr	start	end	strand	gene	annot_ref	annot_find	LCL 01	LCL 02	LCL 03	LCL 04	LCL 05	LCL 06	occurence
ENST00000357654.3	chr17	41196312	41277387	-	BRCA1	1-24	1-24	3226,55	2076,07	2018,37	1560,78	482,43	912,17	6
MSTRG.6981.10	chr13	32889631	32974418	+	BRCA2	1-27	Δ1(14)-10exo(224)[p2515,q138]-Δ11-▼11p(1782)-▼27(1071)	678,69	189,89	401,99	135,34	613,59	718,32	6
MSTRG.11818.19	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼7p(58)-▼24(7)	38,59	20,95	16,97	8,98	6,76	472,96	6
MSTRG.6981.26	chr13	32889640	32984016	+	BRCA2	1-27	Δ1(23)-Δ5-▼27(10669)	63,35	47,06	76,55	70,34	279,85	477,45	6
MSTRG.11818.14	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-▼24(7)	6,94	5,48	5,32	0	2,26	381,96	5
MSTRG.11818.16	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-▼24(7)	0	0	0	1,81	0	358,3	2
MSTRG.11818.6	chr17	41196305	41277550	-	BRCA1	1-24	▼1(163)-Δ(9-10)-Δ11q(3309)-▼24(7)	1366,17	443,66	719,55	430,45	91,94	339,24	6
ENST00000346315.3	chr17	41196313	41277468	-	BRCA1	1-24	▼1(81),Δ1q(6)-Δ(14-17)-Δ24(1)	1143,56	271,35	295,38	539,43	381,93	282,13	6
MSTRG.11818.5	chr17	41196305	41277384	-	BRCA1	1-24	Δ1(3)-Δ(9-10)-▼24(7)	724,34	116,1	828,07	20,54	64,52	182,52	6
MSTRG.11818.17	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-Δ11q(3309)-▼24(7)	0	0	0	2,67	0	145,91	2
MSTRG.11818.3	chr17	41196276	41277370	-	BRCA1	1-24	Δ1(17)-Δ8p(3)-▼10-▼24(36)	755,22	197,79	417,68	127,73	172,14	117,88	6
MSTRG.6981.1	chr13	32889602	32984016	+	BRCA2	1-27	▼1(15)-Δ12-▼27(10669)	22,56	16,69	18,82	2,19	69,94	110,68	6
ENST00000471181.2	chr17	41197646	41277500	-	BRCA1	1-24	▼1(113)-13exo(66)[p3004,q2719]-Δ14p(3)-Δ24(1334)	281,55	51,9	208,95	124,77	59,44	99,02	6
MSTRG.11818.18	chr17	41196305	41277500	-	BRCA1	1-24	▼1(113)-Δ5q(22)-▼24(7)	324,32	132,39	141,55	122,89	99,25	92,81	6
MSTRG.6981.4	chr13	32889631	32977455	+	BRCA2	1-27	Δ1(14)-11int(4254)[q489,p189]-▼27(4108)	108,83	52,59	33,07	23,34	62,28	89,89	6

Informations sur les isoformes assemblées

Annotation descriptive

Valeur d'expression des isoformes par échantillon dans la cohorte

Nombre d' occurrence

Tableau final généré :

```
python3 SOSTAR.py -I /mnt/data/auccam/test/expression -O /mnt/data/auccam/test -R
/mnt/data/auccam/References/ref_gencode_hg38_MANE_transcripts.gtf
```

< 1 min

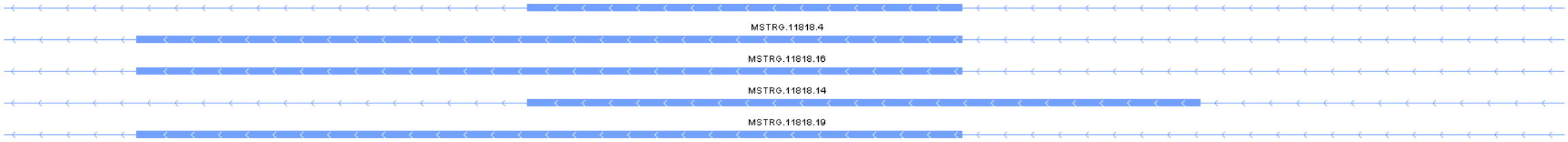
Isoformes assemblées pour par gène

transcript_id	chr	start	end	strand	gene	annot_ref	annot_find	LCL 01	LCL 02	LCL 03	LCL 04	LCL 05	LCL 06	occurence
ENST00000357654.3	chr17	41196312	41277387	-	BRCA1	1-24	1-24	3226,55	2076,07	2018,37	1560,78	482,43	912,17	6
MSTRG.6981.10	chr13	32889631	32974418	+	BRCA2	1-27	Δ1(14)-10exo(224)[p2515,q138]-Δ11-▼11p(1782)-▼27(1071)	678,69	189,89	401,99	135,34	613,59	718,32	6
MSTRG.11818.19	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼7p(58)-▼24(7)	38,59	20,95	16,97	8,98	6,76	472,96	6
MSTRG.6981.26	chr13	32889640	32984016	+	BRCA2	1-27	Δ1(23)-Δ5-▼27(10669)	63,35	47,06	76,55	70,34	279,85	477,45	6
MSTRG.11818.14	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-▼24(7)	6,94	5,48	5,32	0	2,26	381,96	5
MSTRG.11818.16	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-▼24(7)	0	0	0	1,81	0	358,3	2
MSTRG.11818.6	chr17	41196305	41277550	-	BRCA1	1-24	▼1(163)-Δ(9-10)-Δ11q(3309)-▼24(7)	1366,17	443,66	719,55	430,45	91,94	339,24	6
ENST00000346315.3	chr17	41196313	41277468	-	BRCA1	1-24	▼1(81),Δ1q(6)-Δ(14-17)-Δ24(1)	1143,56	271,35	295,38	539,43	381,93	282,13	6
MSTRG.11818.5	chr17	41196305	41277384	-	BRCA1	1-24	Δ1(3)-Δ(9-10)-▼24(7)	724,34	116,1	828,07	20,54	64,52	182,52	6
MSTRG.11818.17	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-Δ11q(3309)-▼24(7)	0	0	0	2,67	0	145,91	2
MSTRG.11818.3	chr17	41196276	41277370	-	BRCA1	1-24	Δ1(17)-Δ8p(3)-▼10-▼24(36)	755,22	197,79	417,68	127,73	172,14	117,88	6
MSTRG.6981.1	chr13	32889602	32984016	+	BRCA2	1-27	▼1(15)-Δ12-▼27(10669)	22,56	16,69	18,82	2,19	69,94	110,68	6
ENST00000471181.2	chr17	41197646	41277500	-	BRCA1	1-24	▼1(113)-13exo(66)[p3004,q2719]-Δ14p(3)-Δ24(1334)	281,55	51,9	208,95	124,77	59,44	99,02	6
MSTRG.11818.18	chr17	41196305	41277500	-	BRCA1	1-24	▼1(113)-Δ5q(22)-▼24(7)	324,32	132,39	141,55	122,89	99,25	92,81	6
MSTRG.6981.4	chr13	32889631	32977455	+	BRCA2	1-27	Δ1(14)-11int(4254)[q489,p189]-▼27(4108)	108,83	52,59	33,07	23,34	62,28	89,89	6

Comparaison des <u>valeurs d'expression</u> entre les patients pour mettre en évidence les isoformes aberrantes

Valeur d'expression des isoformes par échantillon dans la cohorte

Nombre d' occurrence



Description des annotations :

```
python3 SOSTAR.py -I /mnt/data/auccam/test/expression -O /mnt/data/auccam/test -R
/mnt/data/auccam/References/ref_gencode_hg38_MANE_transcripts.gtf
```

< 1 min

transcript_id	chr	start	end	strand	gene	annot_ref	annot_find	LCL 01	LCL 02	LCL 03	LCL 04	LCL 05	LCL 06	occurence
ENST00000357654.3	chr17	41196312	41277387	-	BRCA1	1-24	1-24	3226,55	2076,07	2018,37	1560,78	482,43	912,17	6
MSTRG.6981.10	chr13	32889631	32974418	+	BRCA2	1-27	Δ1(14)-10exo(224)[p2515,q138]-Δ11-▼11p(1782)-▼27(1071)	678,69	189,89	401,99	135,34	613,59	718,32	6
MSTRG.11818.19	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼7p(58)-▼24(7)	38,59	20,95	16,97	8,98	6,76	472,96	6
MSTRG.6981.26	chr13	32889640	32984016	+	BRCA2	1-27	Δ1(23)-Δ5-▼27(10669)	63,35	47,06	76,55	70,34	279,85	477,45	6
MSTRG.11818.14	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-▼24(7)	6,94	5,48	5,32	0	2,26	381,96	5
MSTRG.11818.16	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-▼24(7)	0	0	0	1,81	0	358,3	2
MSTRG.11818.6	chr17	41196305	41277550	-	BRCA1	1-24	▼1(163)-Δ(9-10)-Δ11q(3309)-▼24(7)	1366,17	443,66	719,55	430,45	91,94	339,24	6
ENST00000346315.3	chr17	41196313	41277468	-	BRCA1	1-24	▼1(81),Δ1q(6)-Δ(14-17)-Δ24(1)	1143,56	271,35	295,38	539,43	381,93	282,13	6
MSTRG.11818.5	chr17	41196305	41277384	-	BRCA1	1-24	Δ1(3)-Δ(9-10)-▼24(7)	724,34	116,1	828,07	20,54	64,52	182,52	6
MSTRG.11818.17	chr17	41196305	41277540	-	BRCA1	1-24	▼1(153)-▼8q(95)-Δ(9-10)-Δ11q(3309)-▼24(7)	0	0	0	2,67	0	145,91	2
MSTRG.11818.3	chr17	41196276	41277370	-	BRCA1	1-24	Δ1(17)-Δ8p(3)-▼10-▼24(36)	755,22	197,79	417,68	127,73	172,14	117,88	6
MSTRG.6981.1	chr13	32889602	32984016	+	BRCA2	1-27	▼1(15)-Δ12-▼27(10669)	22,56	16,69	18,82	2,19	69,94	110,68	6
ENST00000471181.2	chr17	41197646	41277500	-	BRCA1	1-24	▼1(113)-13exo(66)[p3004,q2719]-Δ14p(3)-Δ24(1334)	281,55	51,9	208,95	124,77	59,44	99,02	6
MSTRG.11818.18	chr17	41196305	41277500	-	BRCA1	1-24	▼1(113)-Δ5q(22)-▼24(7)	324,32	132,39	141,55	122,89	99,25	92,81	6
MSTRG.6981.4	chr13	32889631	32977455	+	BRCA2	1-27	Δ1(14)-11int(4254)[q489,p189]-▼27(4108)	108,83	52,59	33,07	23,34	62,28	89,89	6

- **Transcrits connus** dans les bases de données ~ 57%
- **Transcrits *de novo*** (tag MSTRG)

Evénements aberrants localisés en **début** et **fin** de transcrit :

- Position 5’ : n = 489 (165 uniques)
- Position 3’ : n = 168 (103 uniques)

Variabilité des régions UTRs

Erreurs d’alignement et/ou d’assemblage

Amélioration du script :

- **Nettoyage des UTRs** : regroupement des transcrits / somme des couvertures

transcript_id	chr	start	end	strand	gene	annot_t	annot_find	barcode01
ENST00000261584.9/MSTRG.23769.15/MSTRG.23769.2	chr16	23603165	23641310	-	PALB2	1-13	1-13	3273,19
MSTRG.23769.18	chr16	23603028	23641402	-	PALB2	1-13	Δ11	559,68
MSTRG.23769.32	chr16	23603161	23641348	-	PALB2	1-13	▼1q(715)-▼1p(1447)	385,94
MSTRG.23769.23	chr16	23603028	23641400	-	PALB2	1-13	▼2p(36)	225,06
MSTRG.23769.20	chr16	23603028	23641396	-	PALB2	1-13	Δ7p(10)	194,58
MSTRG.23769.17	chr16	23603028	23641402	-	PALB2	1-13	1exo(134)[p2517,q329]	183,75
MSTRG.23769.6/ENST00000566069.6	chr16	23603587	23641402	-	PALB2	1-13	Δ12	174,83

- Calcul du **TQL** = *Transcript Quantification Level*

$$TQL = \frac{E_{i,j}}{\sum_k E_{k,j}}$$

- $E_{i,j}$ = expression de l'isoforme (i) du gène (j)
- $\sum_k E_{k,j}$ = somme des expressions des isoformes (k) du gène (j)

- Ajout de calculs de **Zscore** :

$$Zscore = \frac{X - \mu}{\sigma} \qquad Zscore_{median} = \frac{X - median}{\sigma}$$

exp_global | exp_global_smoothed | zscore_expglob | tql | zscore_tql

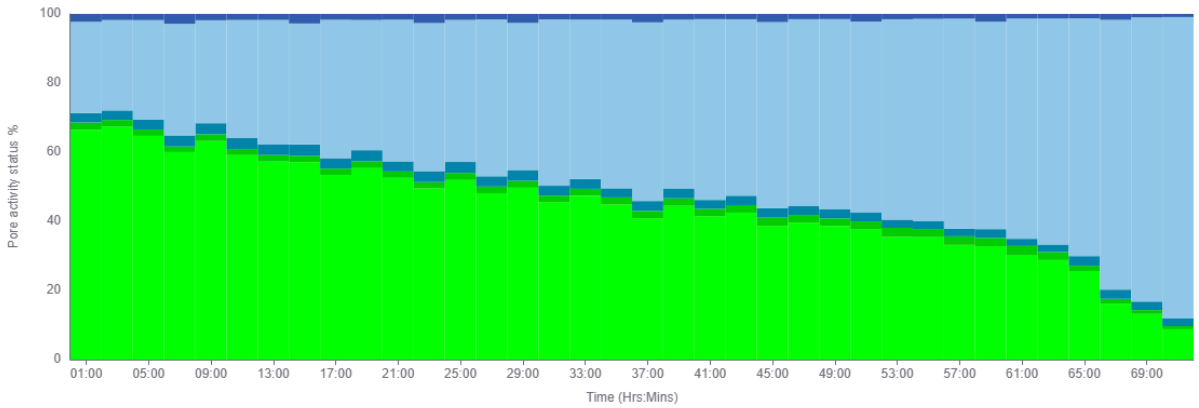
Poursuite des analyses par RNAseq *long read* :

- Séquençage de nouveaux patients :

Type	Nombre échantillons
Contrôles positifs	n = 4
Patientes sélectionnées pour exploration hérédité manquante	n = 40
Contrôles négatifs	n = 20

- Selon le même protocole que le premier run + séquençage sur PromethION P2 Solo

Séquençage en **2 x 32** échantillons



P2Solo

PromethION Flow Cell

Perspectives :

- Améliorations **alignement / assemblage**

- Améliorations **SOSTAR** :

01 **Calcul statistique** : basé sur SpliceLauncher ? *(Leman et al, 2020)*

02 Définir des **signatures** d'expression d'isoformes ?

03 Définir des **critères qualités**

Séquençage ADN *Long Read*

- Objectifs

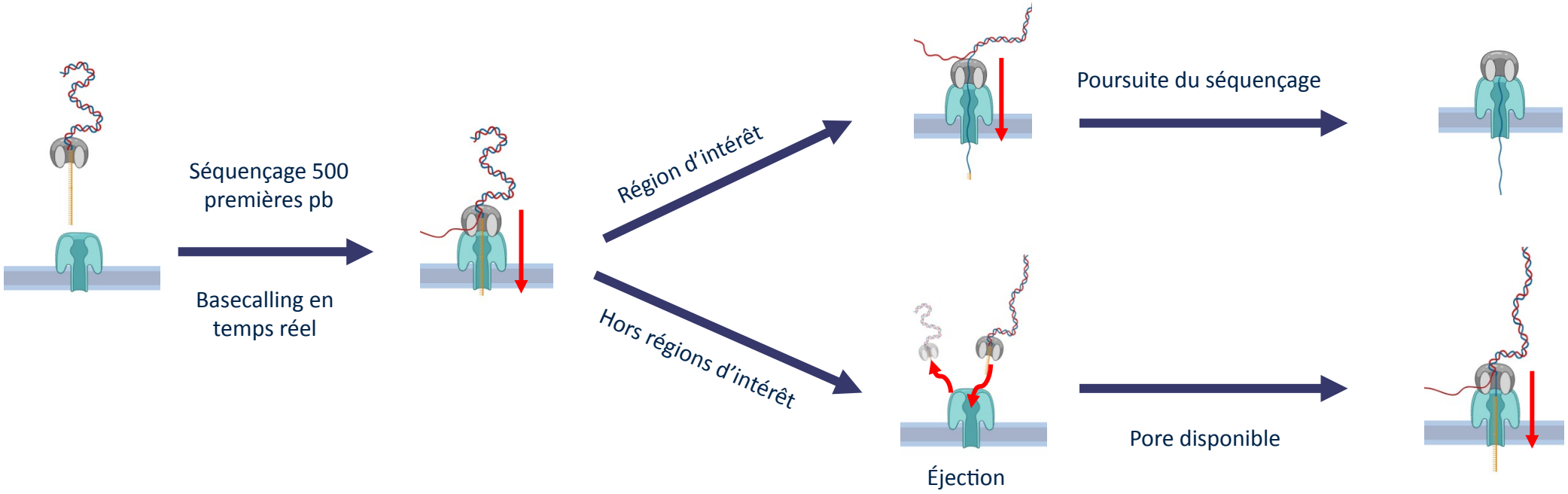
-> Détecter des variants dans les régions régulatrices

-> Détecter de nouveaux variants dans les régions introniques profondes

-> Caractériser des événements de grandes tailles

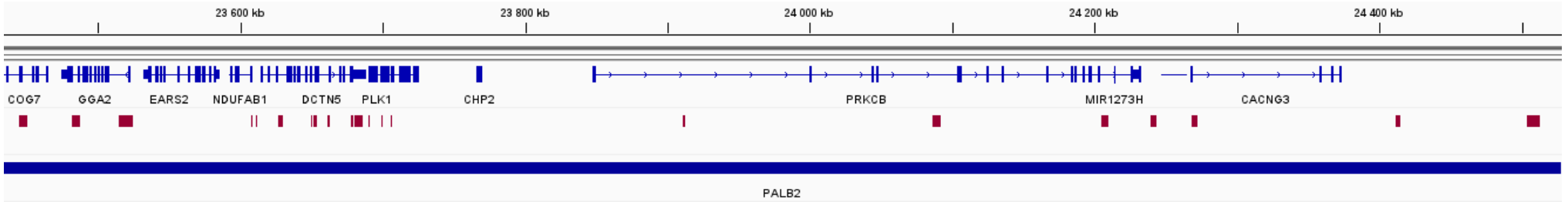
200 patientes déjà séquencées

Adaptive Sampling



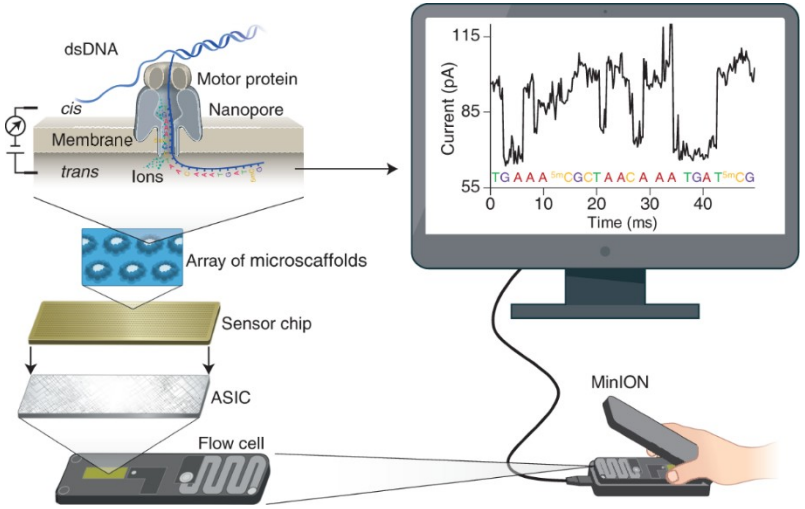
BED avec 232 gènes d'intérêts ± 15kb et élargi aux régions régulatrices pour 5 gènes (*BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, *RAD51D*) -> ~1% du génome

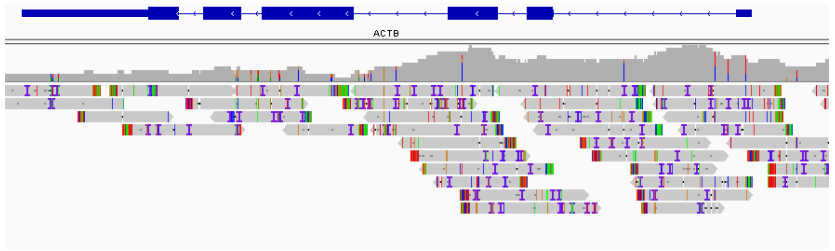
Enhancers issus de GREEN DB[2]

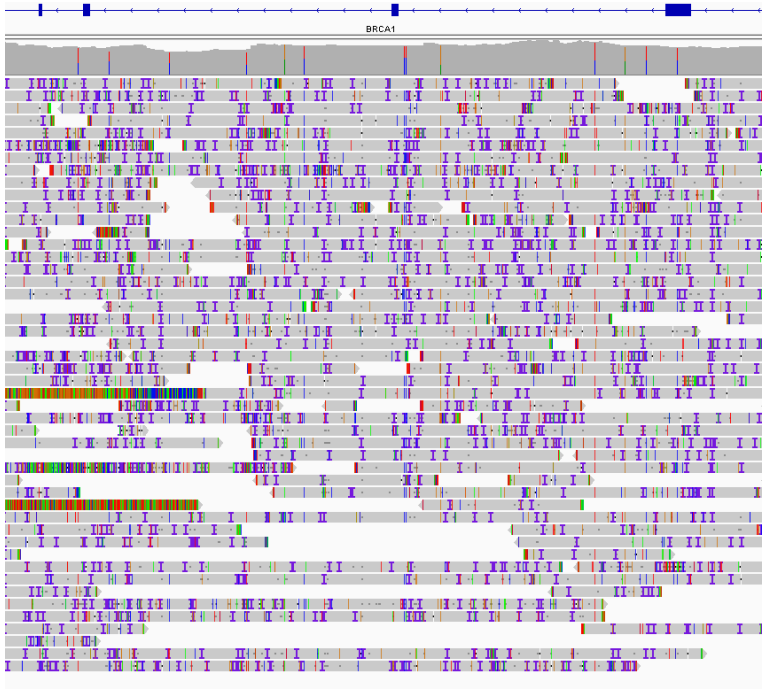


Input

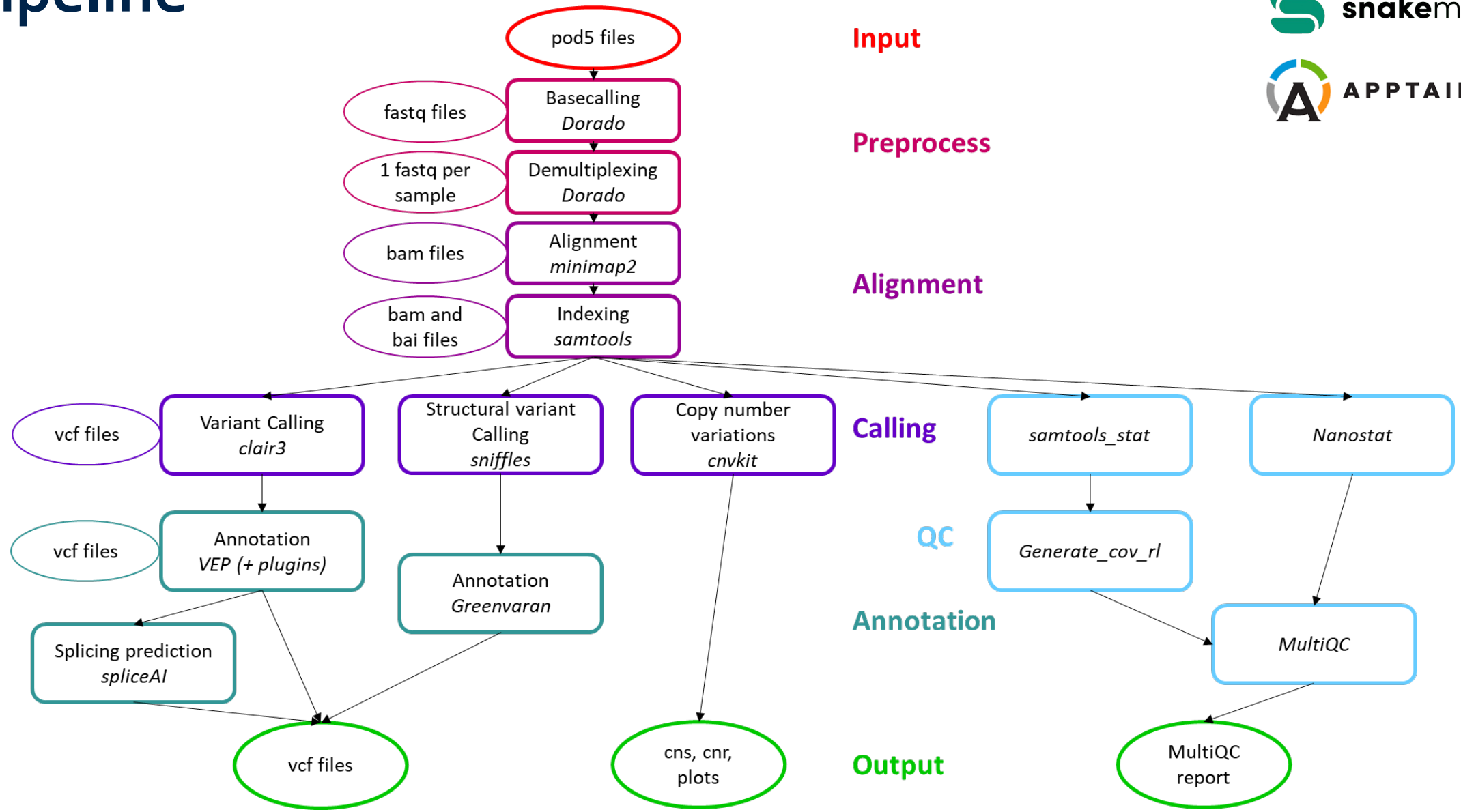
- 1 flowcell -> 4 patients
- 1 flowcell -> ~1To de data (fichiers **.pod5**)
- 1 flowcell -> 1 run de pipeline
- Couverture : ~5x *off target* – ~50x *on target*







Pipeline

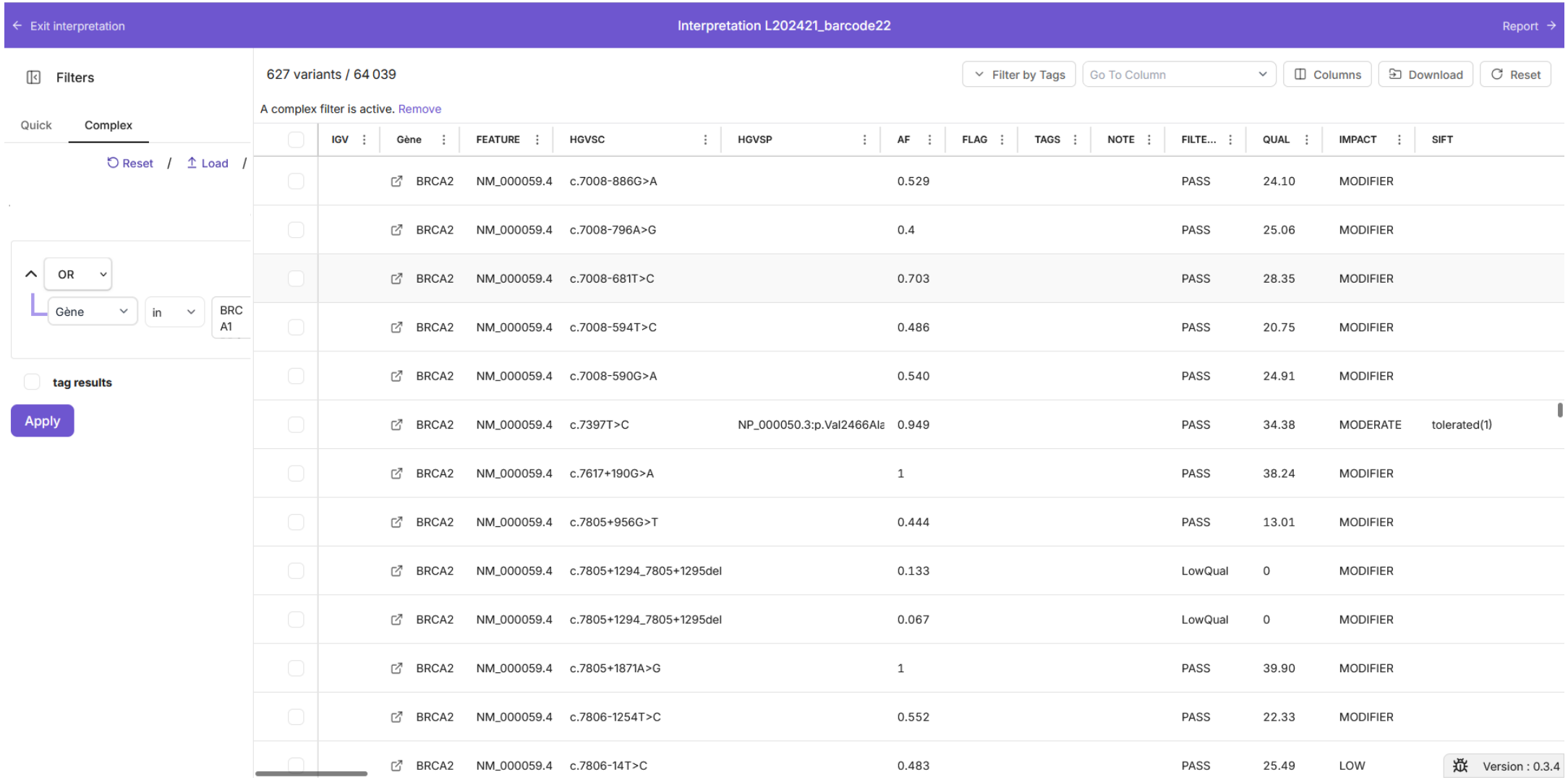

pod5 files
Input
fastq files
Basecalling
Dorado
Preprocess
1 fastq per sample
Demultiplexing
Dorado
bam files
Alignment
minimap2
Alignment
bam and bai files
Indexing
samtools
vcf files
Variant Calling
clair3
Structural variant Calling
sniffles
Copy number variations
cnvkit
Calling
samtools_stat
Nanostat
vcf files
Annotation
VEP (+ plugins)
QC
Generate_cov_rl
Annotation
Greenvaran
Annotation
Splicing prediction
spliceAI
MultiQC
vcf files
cns, cnr, plots
Output
MultiQC report

Hardware

- Workstation dédiée avec 3 GPUs optimisés pour le calcul (NVIDIA A6000 ADA)
- GPUs utilisés par 3 process :

- De loin les process les plus longs :
  - Basecall + Demultiplexing = ~17h
  - spliceAI = ~1h
  - pipeline complet = ~20h
- Optimiser les autres outils pour utiliser du calcul GPU ?

Visualisation : Diagho



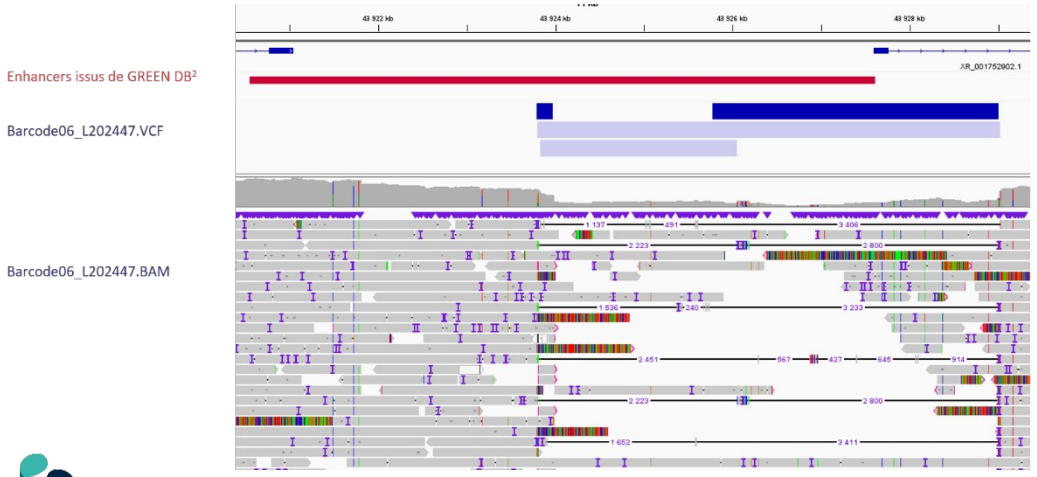
Résultats préliminaires

Sur les 40 premiers patients analysés par notre pipeline :

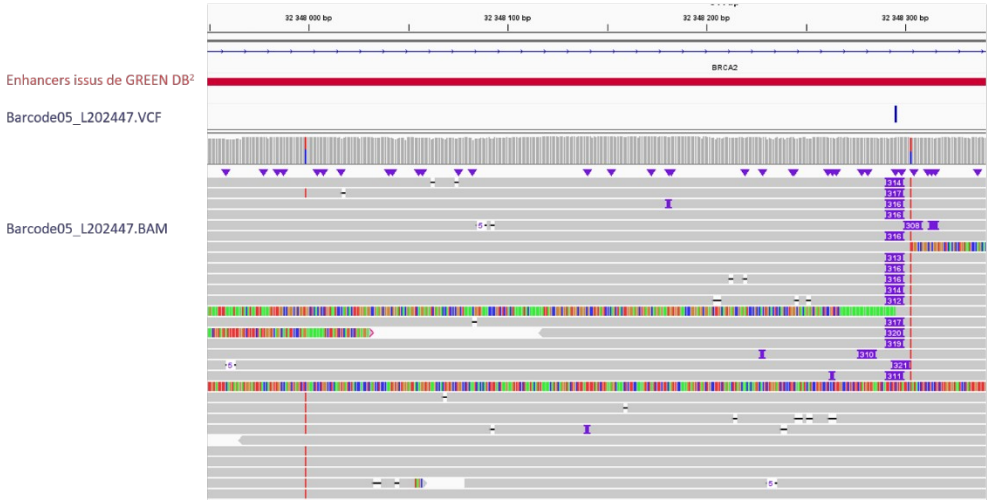
- 13 variants impactant de possibles éléments régulateurs
  - Dont 1 ayant un score élevé FINSURF

- 1 variant d'épissage
- 2 variants structuraux

- **Délétion de ≈5000pb dans une région régulatrice de *BRCA1* retrouvée chez un patient**



- **Insertion de 316pb « <u>en phase</u> » avec une substitution C>T dans l'intron 13 de *BRCA2* retrouvée chez 3 patients**



Perspectives

- Méthylation
  - Refaire le basecall (uniquement zones d'intérêt ?) avec bases modifiées
  - Implémenter Modkit

- Optimisation GPU

- Pepper-Deepvariant ?

- Validation en comparant avec le *short read*

Remerciements

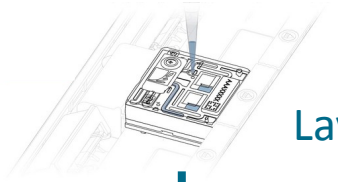
Dr Sophie KRIEGER
Dr Raphael LEMAN
Dr Laurent CASTERA
Dr Agathe RICOU
Dr Flavie BOULOUARD

Dr Nicolas GOARDON
Thibaut LAVOLE
Alexandre ATKINSON
Julien LEVILLY
Céline QUESNELLE

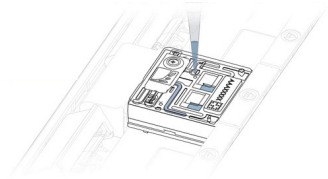
Appel à projets 2018

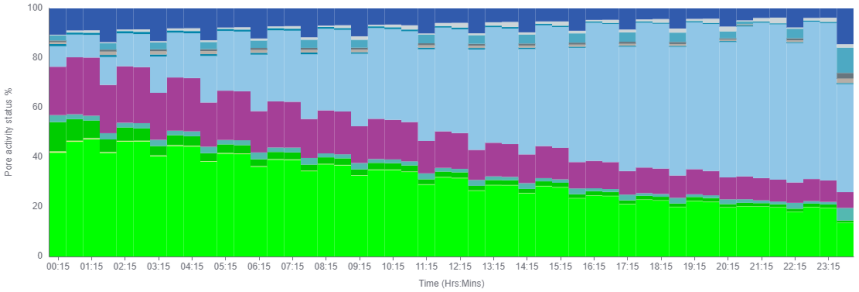
Rotary Club
Lions Club

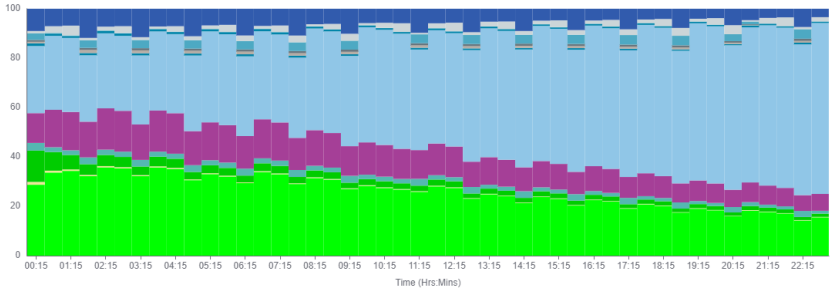
Suivi du *run DNA LR* :

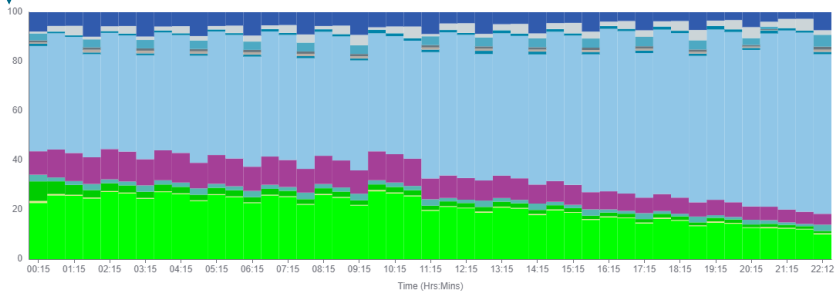


Lavage *flowcell* + rechargement









0h à 24h

24h à 48h

48h à 72h

→ Génération de **1T** de fichiers *.pod5 par *flow cell*

Run health

PORE ACTIVITY

The Pore activity graph shows the performance of your sample as it is being sequenced during a run.

Legend

Sequencing Pore currently sequencing	Adapter Pore currently sequencing adapter	Pore available Pore available for sequencing	Unavailable Pore unavailable for sequencing	Active feedback Channel ejecting analyte	No pore No pore detected in channel
Out of range-high Current is positive but unavailable for sequencing	Out of range-low Current is negative but unavailable for sequencing	Multiple Multiple pores detected. Unavailable for sequencing.	Saturated The channel has switched off as current levels exceed hardware limitations	Zero Pore currently unavailable for sequencing	Channel disabled Channel is disabled and awaiting another pore scan
Unclassified Pore status unknown					

Rapport multiQC :

MultiQC

Centre Baclesse unicancer NORMANDIE - CAEN

DNASeq_LongRead_Adaptive_Sampling

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2023-09-13, 11:36 based on data in: `/mnt/data/auccam/DNASeq_LR_P2solo/L202336`

NanoStat

NanoStat various statistics from a long read sequencing dataset in fastq, bam or sequencing summary format. *DOI: 10.1093/bioinformatics/bty149.*

Aligned stats

NanoStat statistics from BAM files.

Copy table | Configure Columns | Plot | Showing 4/4 rows and 7/10 columns.

Sample Name	Median length	Read N50	Median Qual	Median Identity	# Reads (K)	Total Bases (Mb)	Aligned Bases (Mb)
barcode05	568 bp	603 bp	13.9	95.9%	18 584.1	12 717.0	12 499.7
barcode06	568 bp	600 bp	13.9	95.9%	29 524.0	19 777.6	19 529.2
barcode07	565 bp	596 bp	13.9	95.9%	23 994.9	15 504.0	15 253.5
barcode08	566 bp	597 bp	13.9	95.9%	28 015.5	18 301.8	18 041.8

Toolbox

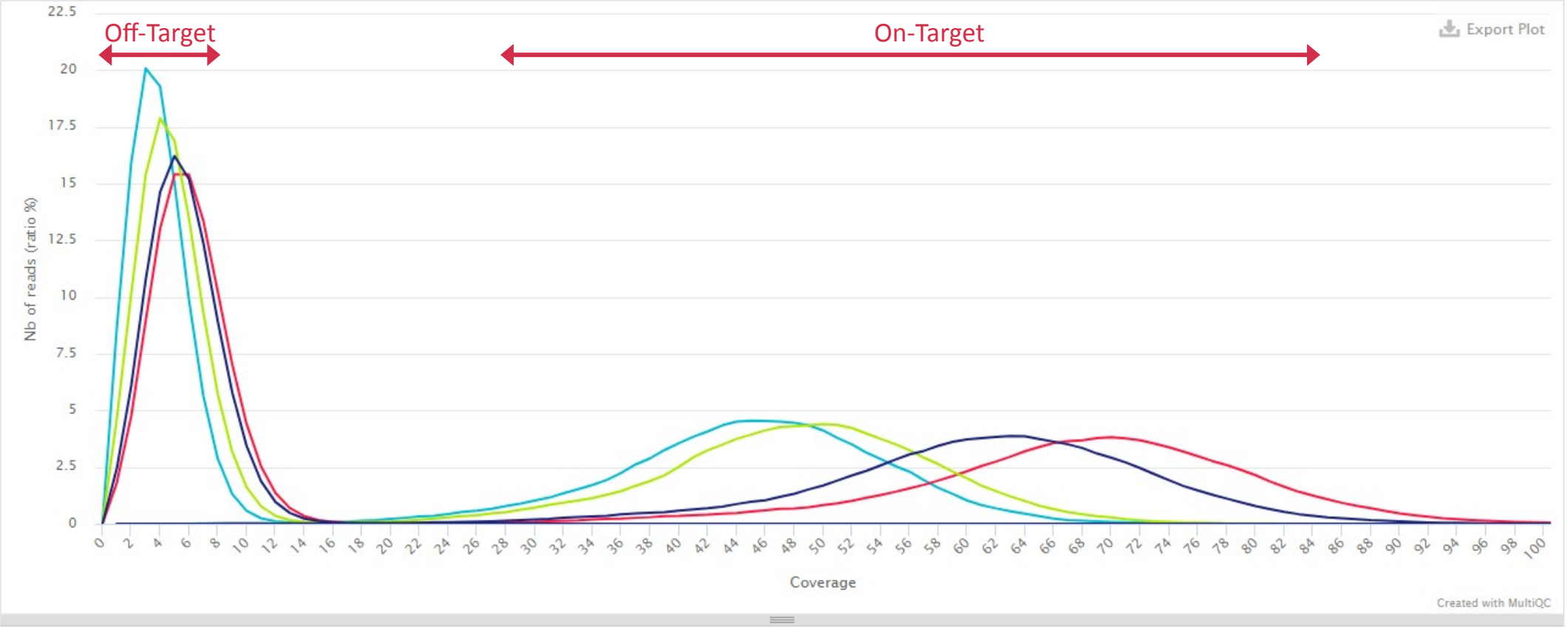
Rapport multiQC :

Adaptive sampling statistics

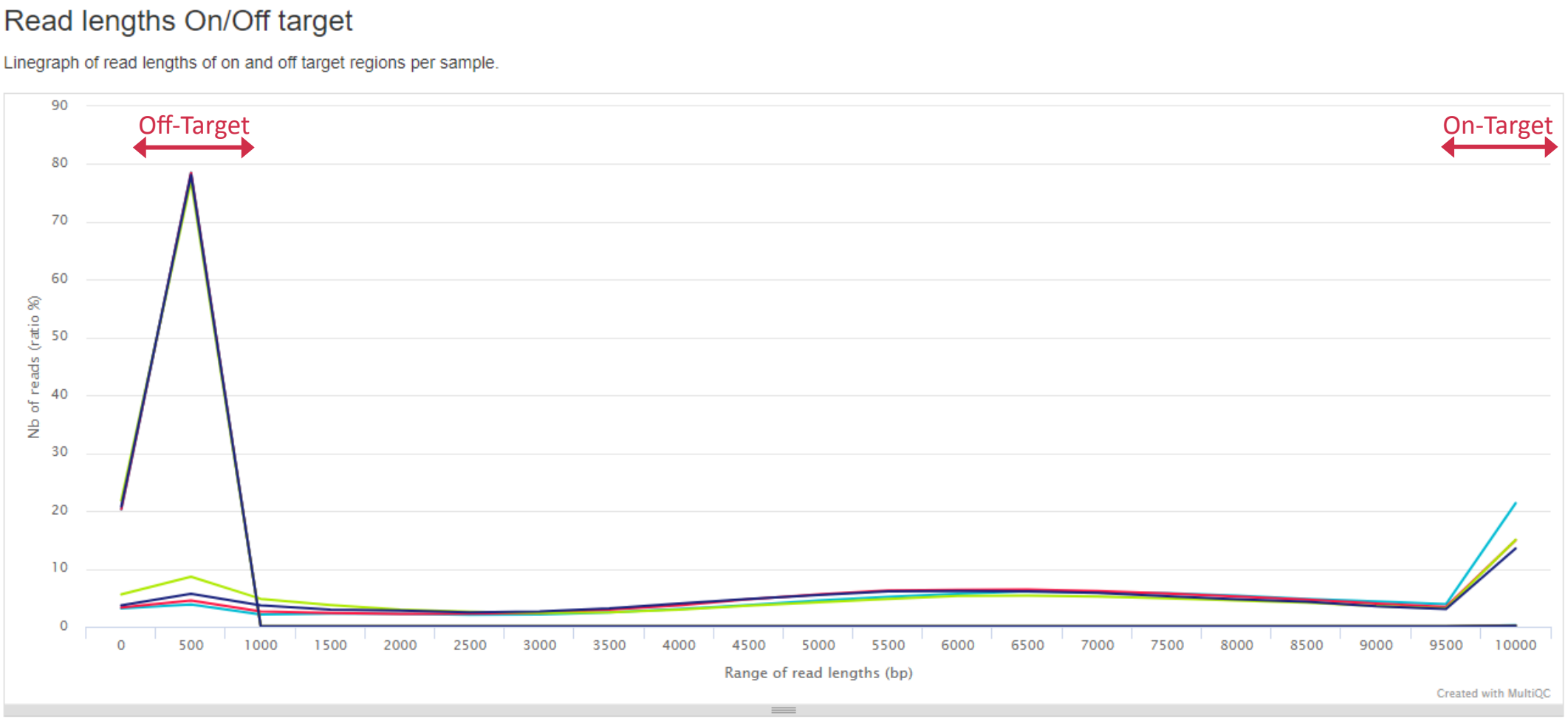
Various statistics of adaptive sampling results. Using `samtools stats` on bedfile used during the adaptive sampling (i.e. on target region) and on the bedtool complement of this bedfile (off target regions).

Coverage On/Off target

Linegraph of coverage of on and off target regions per sample.

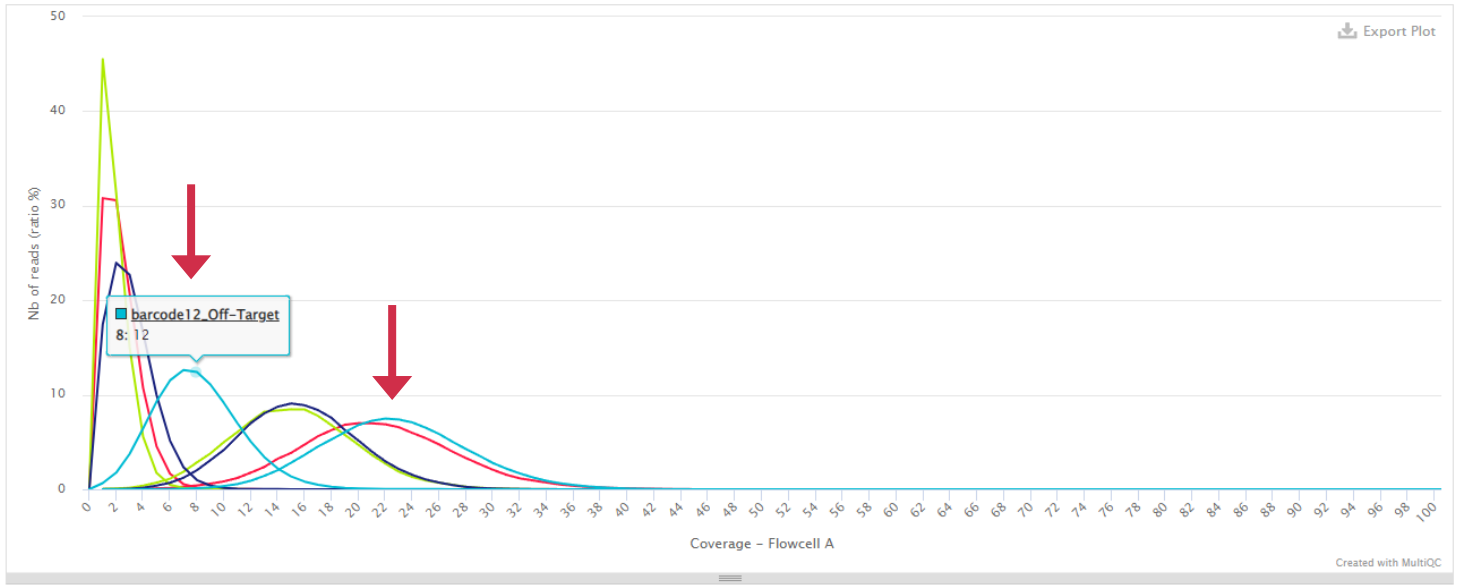


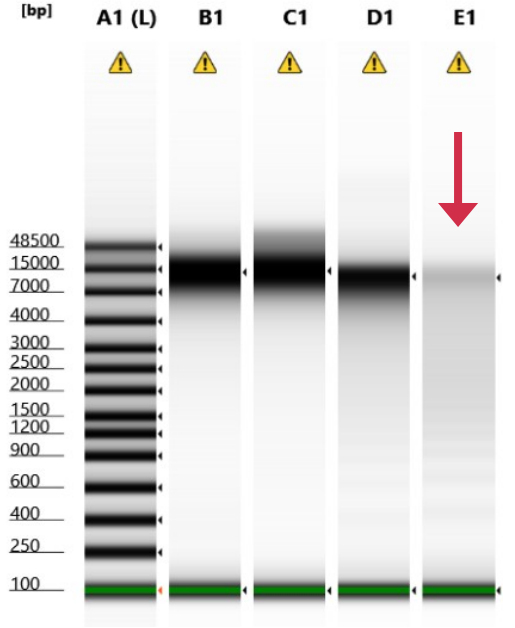
Rapport multiQC :



Rapport multiQC :

Sample Name	Median length	Read N50	Median Qual	Median Identity	# Reads (K)	Total Bases (Mb)	Aligned Bases (Mb)
barcode09	516 bp	547 bp	15.0	96.8%	10 708.9	6 561.3	6 474.2
barcode10	512 bp	548 bp	15.0	96.8%	7 250.3	4 582.0	4 522.1
barcode11	493 bp	524 bp	14.8	96.7%	18 382.3	9 298.2	9 106.4
barcode12	484 bp	514 bp	14.8	96.7%	53 930.4	25 626.9	24 939.1





→ **Séquençage privilégié des petits fragments du bc12 bloquant les pores disponibles pour les autres bc**