Club bioinformatique
Analyse primaire de génomes

Bénédicte Condamine et Marie Petitjean

IAME

Résumé

Mapping, assemblage et contrôle qualité. Comment lancer soi-même son analyse primaire ?

Table des matières

1 Mapping/Assemblage et Contrôle qualité **2**
- 1.1 Mapping ou Assemblage ? . . . 2
- 1.2 Mapping . . . 2
  - 1.2.1 BWA . . . 2
  - 1.2.2 Minimap2 . . . 2
  - 1.2.3 Samtools . . . 3
  - 1.2.4 Bamstat . . . 4
- 1.3 Assemblage . . . 5
  - 1.3.1 SPAdes . . . 5
  - 1.3.2 Flye . . . 5
  - 1.3.3 Unicycler . . . 5
  - 1.3.4 QUAST . . . 6

2 Supplément **8**
- 2.1 Format SAM . . . 8
  - 2.1.1 Fichier SAM . . . 8
  - 2.1.2 Flag . . . 9
  - 2.1.3 Code CIGAR . . . 10
- 2.2 Samtools view . . . 10
- 2.3 Visualisation de l'alignement . . . 11

1 Mapping/Assemblage et Contrôle qualité

1.1 Mapping ou Assemblage ?

A partir des données de séquençage, il existe deux types d'analyses principales. D'une part le mapping qui nécessite un génome de référence et d'autre part, l'assemblage qui ne dépend que de lui même. Chacune des techniques à ses avantages et ses inconvenients et permet des analyses différentes.

Alignement	Assemblage
Référence obligatoire	Pas besoin de référence
Régions manquantes visibles	Visualisation de régions spécifiques
Analyse de SNPs	Notions de synténie *
Peut être réalisé avec une couverture faible	Nécessite une certaine couverture

** enchainement des gènes sur un chromosome, notion de positionnement.*

1.2 Mapping

Le mapping permet à partir de reads de générer un alignement sur un génome de référence.

1.2.1 BWA

BWA est un algorithme d'alignement de reads (mapping) sur un "large" génome de référence. Il permet à partir de reads (fichier au format fastq) de générer un fichier d'alignement au format sam.

Pour générer un alignement, 2 étapes sont nécessaires :

- l'indexation du génome de référence avec bwa index, cela sert principalement à gagner de la mémoire et du temps de calcul.
- l'alignement sur génome de référence avec bwa mem

bwa
$ bwa index reference.fasta
-t <nombre> : nombre de thread
-o <fichier> : le nom du fichier de sortie de bwa mem
$ bwa mem reference.fasta -t 2 -o output.sam sample_R1.fastq.gz sample_R2.fastq.gz

(?) Indexer le génome de référence (/home/progs/docker/Club_bioinfo/Mapping_assembly/data)
(?) Lancer un alignement de l'échantillon nettoyé précédement sur le génome de référence avec 5 threads et donner le nom du fichier résultat "sample_illumina.sam".

1.2.2 Minimap2

Le mapping avec Minimap2 se fait directement depuis la référence et avec les reads nanopore.

minimap2
-ax <technologie> : map-ont pour indiquer mapping avec les reads nanopore, map-pb pour le pacbio
$ minimap2 -ax map-ont ref.fa ont.fq.gz > aln.sam

(?) Lancer l'alignement avec minimap de reads_nanopore.fastq sur reference.fasta

1.2.3 Samtools

Samtools (http://www.htslib.org/doc/samtools.html) est une suite de logiciels qui permettent de manipuler des alignements au format BAM, de les importer ou les exporter au format SAM, de les trier, de les fuisonner, de les indexer ou encore de filtrer une région particulière.

Tri → "samtools sort" permet d'ordonner les reads en fonction de leurs positions sur la séquence de référence. De nombreux logiciels, utilisant un format SAM ou BAM, nécessitent un fichier d'alignement où les reads sont triés.

samtools sort
-o <fichier> : nom du fichier de sortie
-O <format> : format de sortie (bam, sam)
-@ <nombre> : nombre de threads
$ samtools sort -o out.bam -O bam -@ nbthreads [in.sam|in.bam]

(?) Trier l'alignement précédent.

Index → "samtools index" permet d'indexer un fichier BAM. De nombreux logiciels, utilisant un BAM, nécessitent un fichier d'alignement où les reads sont triés et indexés.

samtools index
-b : création d'un index bai à partir d'un fichier bam
$ samtools index -b <in.bam>

(?) Indexer l'alignement trié.

Statistique → "samtools idxstats" permet d'obtenir de simple statistiques descriptives. Il génère un fichier tabulé avec une ligne par référence et une ligne pour les reads non alignés. Chaque ligne contient quatre éléments, le nom de la référence, la longueur de la référence, le nombre de reads alignés sur la référence et le nombre de reads non-alignés (lorsque l'autre read de la paire est aligné).

NC_000913.3	4 641 652	2 327 266	21 766
*	0	0	680 814

Extrait de résultats

samtools idxstats
$ samtools idxstats in.bam > fichier_idxstats.txt

(?) Utiliser idxstats avec l'alignement.

Profondeur → "samtools depth" calcule la profondeur de la couverture à chaque position.

```
NC_000913.3   1   46
NC_000913.3   2   47
NC_000913.3   3   48
NC_000913.3   4   49
NC_000913.3   5   49
...
```

Extrait de résultats

samtools depth
-a : pour toutes les positions de la référence (même celle avec une couverture de 0)
$ samtools depth -a [in.sam|in.bam] > fichier_depths.txt

(?) Calculer la profondeur de chaque base de l'alignement.

1.2.4 Bamstat

Bamstat (http://bamstats.sourceforge.net/) calcule et génère des graphiques des différentes métriques des fichiers SAM/BAM.
Le fichier généré précise par défaut le nombre de base de la référence, la couverture moyenne, médiane, l'écart-type, 1er quartile (25%), 3ème quartile (75%), 2.5% percentile, 97.5% percentile, le minimum et le maximum. Selon les options, le fichier affichera les mêmes métriques pour la couverture des régions couvertes, la qualité de mapping et la longueur des reads.

```
Coverage
ref         N           mean      median    sd        q1        q3        2.5%      97.5%     min       max
1           4,984,245   256.34    302.00    139.83    259.00    334.00    0.00      407.00    0         6927
2           138,218     55.11     0.00      365.78    0.00      0.00      0.00      302.00    0         4636

Coverage (mapped regions only)
ref         N           mean      median    sd        q1        q3        2.5%      97.5%     min       max
1           4,051,241   315.37    313.00    73.75     288.00    342.00    228.00    412.00    1         6927
2           22,012      346.07    158.00    859.92    31.00     261.00    1.00      4229.68   1         4636

Mapping qualities
ref         N           mean      median    sd        q1        q3        2.5%      97.5%     min       max
1           8,733,241   58.79     60.00     7.75      60.00     60.00     46.00     60.00     0         60
2           58,877      50.76     60.00     19.67     60.00     60.00     0.00      60.00     0         60

Read lengths
ref         N           mean      median    sd        q1        q3        2.5%      97.5%     min       max
1           8,733,241   150.46    151.00    7.52      151.00    151.00    151.00    151.00    30        151
2           58,877      149.67    151.00    11.79     151.00    151.00    151.00    151.00    30        151
```

Extrait de fichier de bamstat

bamstat
-i <fichier> : fichier d'alignement
-o <fichier> : fichier de sortie
-v <format> : format de la sortie 'simple' (défaut) ou 'html'
-m : qualités de mapping
-q : couverture pour les régions couvertes
-l : longueurs des reads
$ java -Xmx4g -jar /opt/progs/BAMStats/BAMStats.jar [option] -o out.tab -i <in.bam>

(?) Vérifier l'alignement avec Bamstat.

1.3 Assemblage

L'assemblage est un procédé qui permet à partir des reads de reconstituer *de novo* un génome c'est à dire sans référence préalable.

1.3.1 SPAdes

SPAdes (http://cab.spbu.ru/software/spades/) est un outil d'assemblage de génomes. A partir des reads générés par séquençage haut-débit, il permet d'obtenir un ensemble de contigs représentant le génome séquencé. Le nombre et la taille des contigs vont dépendre de la qualité de l'assemblage et du type de séquençeur utilisé.

L'ensemble des contigs assemblés se trouve dans le fichier contigs.fasta. En cas de besoin, les informations fournies par le logiciel se trouvent dans le fichier spades.log.

spades
-k <liste de nombre> : 21,33,55,77 (liste de tailles de k-mer utilisés)
- -careful : minimise le nombre de mismatch dans les contigs
-t <nombre> : nombre de threads
-1 <fichier> : fichier fastq R1
-2 <fichier> : fichier fastq R2
-o <dossier> : dossier contenant les résultats
$ spades -k 21,33,55,77 - -careful -t 10 -1 sample_R1.fastq.gz -2 sample_R1.fastq.gz -o spades_results

(?) Lancer l'assemblage de sample.

1.3.2 Flye

Flye est un outil pour l'assemblage de génome avec des longs reads.

flye
- -nano-raw <fichier> : fichier contenant les reads nanopore
-g <nombre> : taille du génome attendu
-t <nombre> : nombre de threads
-o <dossier> : dossier contenant les résultats
$ flye –nano-raw fastq/sample.fastq -g 5500000 -o flye_assembly -t 10

(?) Lancer l'assemblage de reads_nanopore.fastq.

1.3.3 Unicycler

Unicycler est un logicel d'assemblage hubride. Il prend en entrée des reads Illumina et des reads Nanopore. Au début, un assemblage des reads Illumina est réalisé avec Spades, puis les reads nanopore sont utilisés pour compléter l'assemblage.

unicycler
-1 <fichier> : fichier contenant les reads Illumina R1
-2 <fichier> : fichier contenant les reads Illumina R2
-l <fichier> : fichier contenant les reads Nanopore
-o <dossier> : dossier contenant les résultats (sera créé)
$ unicycler -1 reads_R1.fastq -2 reads_R2.fastq -l reads_nanopore.fastq -o unicycler_assembly

(?) Lancer l'assemblage des reads nanopore et illumina.

1.3.4 QUAST

Quast (http://bioinf.spbau.ru/quast) est un logiciel qui permet de donner des métriques sur l'assemblage. Cela permet donc de vérifier la qualité de l'assemblage réalisé.

Assembly	E_coli_test
# contigs (>= 0 bp)	145
# contigs (>= 1000 bp)	110
# contigs (>= 5000 bp)	61
# contigs (>= 10000 bp)	52
# contigs (>= 25000 bp)	41
# contigs (>= 50000 bp)	28
Total length (>= 0 bp)	5641492
Total length (>= 1000 bp)	5617537
Total length (>= 5000 bp)	5512385
Total length (>= 10000 bp)	5440236
Total length (>= 25000 bp)	5272304
Total length (>= 50000 bp)	4777055
# contigs	145
Largest contig	462328
Total length	5641492
GC (%)	57.14
N50	179099
N75	92723
L50	11
L75	21
# N's per 100 kbp	0.00

Exemple de résultats

Les premières métriques concernent le nombre de contigs en fonction de leurs tailles ainsi que la longueur en paires de bases des groupes de contigs considérés.

La deuxième partie indique d'autres métriques. Tout d'abord, le nombre de contigs, puis la taille du plus gros contig et la longueur totale des contigs en bp. Ensuite, on trouve le pourcentage de GC, la valeur du N50 et du N75 ainsi que du L50 et du L75.

Le N50 correspond à la longueur du contig pour laquelle la somme de tous les contigs de tailles supérieures ou égale forment la moitié de la taille du génome. Le L50 est le nombre de contigs correspond à la valeur du N50. Il en est de même pour les valeurs de N75 et L75 avec 75% au lieu de 50%.

Pour finir, on a le nombre de N présent dans l'assemblage pour 100 000 bp.

quast
-t <nombre> : nombre de threads
-o <dossier> : nom du dossier qui contient les résultats
$ quast -t 10 -o quast_results assemblage.fasta

(?) Lancer l'analyse de l'assemblage précédemment réalisé.

2 Supplément

2.1 Format SAM

2.1.1 Fichier SAM

Le fichier sam est un alignement, il indique la position de chaque reads sur une référence. Il est séparé en deux parties : le "header" (lignes commençant par "@") et les reads. Chaque ligne (hors header) représente un alignement d'un read sur la référence.

```
@SQ     SN:NC_000913.3  LN:4641652
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem -t 5 reference.fna -o sample.sam sample_R1
.fastq.gz sample_R2.fastq.gz
ERR434779.1     83      NC_000913.3     81      60      100M    =       51      -130    GTTAC
CTGCCGTGATCAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGT
AC    CCC<B??CCEFFFFHHHFHHGGGE=AGIHGIIIHIIHEGCG@DIGHGGD?IHCIGFGIFFBBFFEGIGGGGHA>FH>ECHGIIIIEG
FACDCFDFFD??@    NM:i:2  MD:Z:14G0T84    MC:Z:100M       AS:i:90 XS:i:0
ERR434779.1     163     NC_000913.3     51      60      100M    =       81      130     AAAGA
GTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGATCAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCA
TA    @@;DDDA?FFFDBHGIBEEEHEGIG>EHDFHHIIGFG>FHICGEGIIIIG9BFEHIIIIEEIIFHGIIEH4@CGHCEEEEHFHEEEE
EECCCC;ACCCCC    NM:i:2  MD:Z:44G0T54    MC:Z:100M       AS:i:90 XS:i:0
ERR434779.2     99      NC_000913.3     19      60      100M    =       97      178     CAACG
GGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGATCAAATTAAAATTTTATTGACT
TA    CCCFFFFFGHHHHJJJJJJJIIHHIJJJJJJJJIJJJIFHGHJJJGIIIJJJJJJJIIJHHHHHHFFFFFEEDDDBCEEDDDDDDDD
EDDDFDDDDDDDC    NM:i:2  MD:Z:76G0T22    MC:Z:100M       AS:i:90 XS:i:0
ERR434779.2     147     NC_000913.3     97      60      100M    =       19      -178    AAATT
AAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAA
AC    DEDEEEEEEEFFFFFFHHHHHHJJJJHJJJJJIGJJJJIIIJIIIJIIGDIIJJJJIHDJIHIJJGJJJJJJJIJHJJJJIJIIIJJ
HGHFHFFFFFCBB    NM:i:0  MD:Z:100        MC:Z:100M       AS:i:100        XS:i:0
```

Extrait de fichier sam

Dans cet exemple, il y a deux lignes dans le "header" :

- **@SQ SN:NC_000913.3 LN:4641652** : "@SQ" indique que la ligne contient des informations sur la référence.
  - "SN" indique le nom de la référence *ie* "NC_000913.3".
  - "LN" indique la taille de la référence *ie* 4 641 652 pb.

- **@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -t 5 reference.fna -o sample.sam sample_R1.fastq.gz sample_R2.fastq.gz** : "@PG" indique que la ligne contient des informations sur l'execution du programme de mapping.
  - "ID" indique l'identifiant du logiciel *ie* "bwa".
  - "PN" indique le nom du logiciel *ie* "bwa".
  - "VN" indique le version du logiciel *ie* "0.7.17-r1188".
  - "CL" indique la ligne de commande utilisée *ie* "bwa mem -t 5 reference.fna -o sample.sam sample_R1.fastq.gz sample_R2.fastq.gz".

Si on prend la première ligne des reads :

```
ERR434779.1     83      NC_000913.3     81      60      100M    =       51      -130    GTTACCTGCCGTG
ATCAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTAC    CCC<B??CCE
FFFFHHHFHHGGGE=AGIHGIIIHIIHEGCG@DIGHGGD?IHCIGFGIFFBBFFEGIGGGGHA>FH>ECHGIIIIEGFACDCFDFFD??@    NM:i:2
  MD:Z:14G0T84    MC:Z:100M       AS:i:90 XS:i:0
```

Extrait de fichier sam

n°	Exemple	Description
1	ERR434779.1	nom du read
2	83	flag du read
3	NC_000913.3	la référence sur laquelle est positionné le read
4	81	position start du read
5	60	qualité du mapping
6	100M	code CIGAR de cet alignement
7	=	référence sur laquelle est alignée l'autre read*
8	51	position de l'autre read de la paire
9	-130	taille de l'insert
10	GTTACCTGCCGTGATCAAA...	séquence nucléique du read
11	CCC<B??CCEFFFFHHHFHHG...	séquence qualité du read
12	NM:i:2	nombre de nucléotides différents
13	MD:Z:14G0T84	information sur les mismatchs
14	MC:Z:100M	code CIGAR de l'autre read
15	AS:i:90	score de logiciel d'alignement
16	XS:i:0	brin de la référence

** lorsque la référence sur laquelle est alignée l'autre read de la paire est "=" cela signifie que les deux reads de la paire sont alignés sur la même référence.*

2.1.2 Flag

Le "FLAG" est une valeur numérique correspondant à une ou plusieurs significations.

Valeur	Description
1	template having multiple segments in sequencing
2	each segment properly aligned according to the aligner
4	segment unmapped
8	next segment in the template unmapped
16	SEQ being reverse complemented
32	SEQ of the next segment in the template being reverse complemented
64	the first segment in the template
128	the last segment in the template
256	secondary alignment
512	not passing filters, such as platform/vendor quality controls
1024	PCR or optical duplicate
2048	supplementary alignment

Dans l'exemple, l'alignement du read ERR434779.1 présente un flag de 83, ce qui correpond à 1+2+16+64. Le read provient donc d'une paire ("1"), cette paire est corectement positionnée

(sur la même référence) ("2"), il s'agit du read du brin reverse ("16") et c'est le premier de la paire ("64").

Le site https://broadinstitute.github.io/picard/explain-flags.html permet de faire des vérifications entre une valeur numéique et sa signification, et inversement.

2.1.3 Code CIGAR

Le code CIGAR est un code de caractère permettant de visualiser comment le read est aligné sur la référence.

Valeur	Description
D	Deletion ; the nucleotide is present in the reference but not in the read
H	Hard Clipping ; the clipped nucleotides are not present in the read.
I	Insertion ; the nucleotide is present in the read but not in the reference
M	Match ; can be an alignment match or mismatch.
N	Skipped region ; a region of nucleotides is not present in the read
P	Padding ; padded area in the read and not in the reference
S	Soft Clipping ; the clipped nucleotides are present in the read
X	Read Mismatch ; the nucleotide is present in the reference
=	Read Match ; the nucleotide is present in the reference

Dans l'exemple, l'alignement est codé en 100M ce qui signifie, que les 100 pb de reads sont alignés.

Le read "ERR434779.14" est aligné avec un code CIGAR "3S97M", ce qui signifie que 3 pb sont "soft clipping", suivi par 97 pb alignées.

Le "clipping" signifie qu'une portion du read à ses extrémités ne s'aligne pas sur la référence. Il existe deux types de "clipping" :

- soft : ce sont des bases en 5' ou 3' qui ne sont pas alignées.
- hard : ce sont des bases en 5' ou 3' qui ne sont pas alignées ET ont été retirées du fichier d'alignement.

Exemple de clipping 3S8M1D6M4S :

```
REF: CTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAA
READ:        gggTCGTCCGT-GAGCATgggg
```

Exemple de skipped region 9M23N8M :

```
REF: CTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAA
READ:        GTGTAACCC.....................TCAGAATA
```

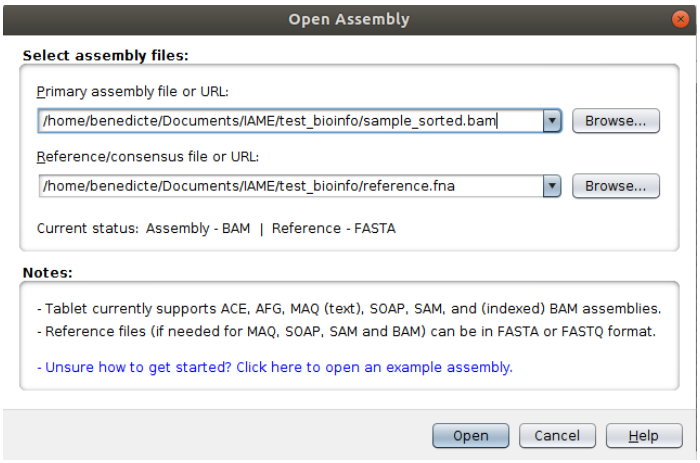
2.2 Samtools view

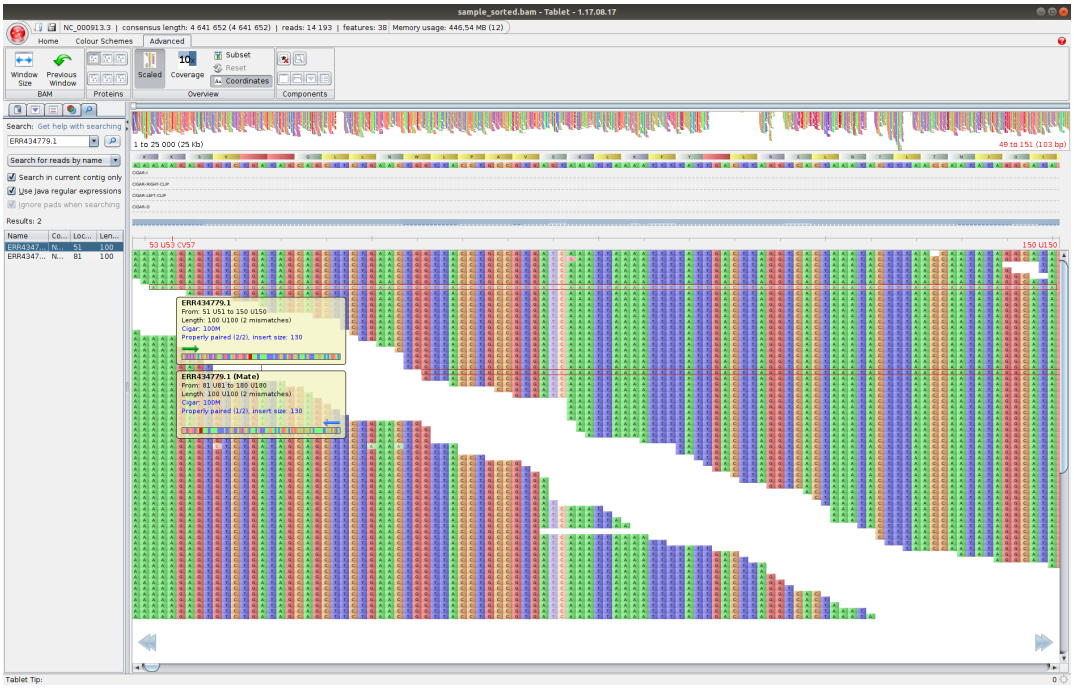
Samtools permet aussi de filtrer les reads d'un alignement en fonction des flags.

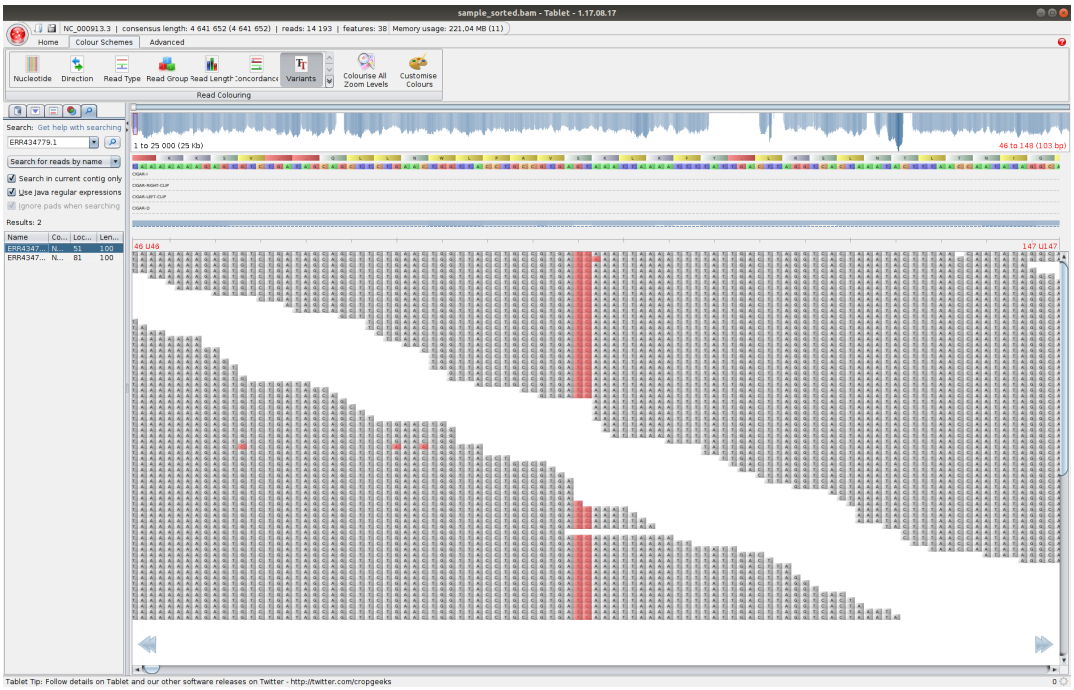
samtoolsview
-f <nombre> : Filtre avec ces flags.
-F <nombre> : Filtre sans ces flags
-o <fichier> : Fichier d'alignement filtré.
$ samtools view -f 83 -o alignement_filtre.bam alignement.bam

2.3 Visualisation de l'alignement

Le logiciel Tablet `https://ics.hutton.ac.uk/tablet/` permet de visualiser un alignement.






sample_sorted.bam - Tablet - 1.17.08.17
Home
Colour Schemes
Advanced
Nucleotide
Direction
Read Type
Read Group
Read Length
Concordance
Variants
Colourise All Zoom Levels
Customise Colours
Read Colouring
Search:
Search for reads by name
Search in current contig only
Use Java regular expressions
Ignore pads when searching
Results: 2
Name
1 to 25 000 (25 kb)
46 to 148 (103 bp)
Tablet Tip: Follow details on Tablet and our other software releases on Twitter - http://twitter.com/cropgeeks