
**Introduction
aux systèmes
UNIX -
Preprocessing
and mapping
of NGS data**

École de bioinformatique
AVIESAN-IFB 2018

Denis Puthier, TAGC/Inserm,
U1090, denis.puthier@univ-amu.fr

Claire Toffano-Nioche, CNRS,
claire.toffano-nioche@u-psud.fr

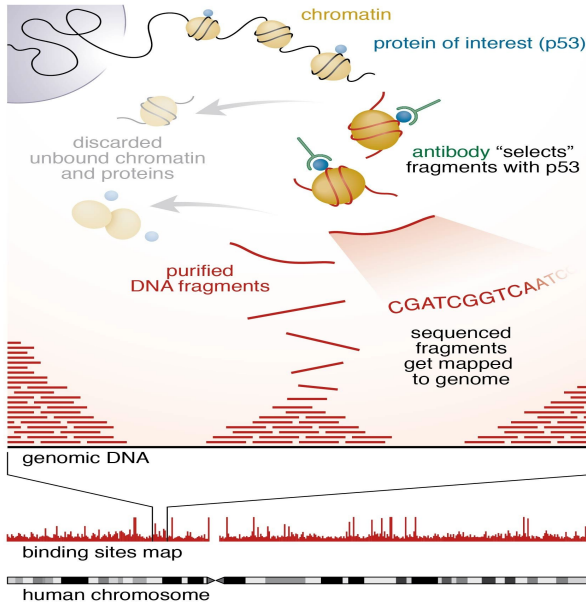
Julien Seiler, IGBMC,
seilerj@igbmc.fr

Gildas le Corguillé,
lecorguille@sb-roscoff.fr

Et tout le staff !!

Présentation du jeu de données

- Immuno-précipitation de chromatine (ChIP-Seq).
 - Un traitement (ADN fragmenté + immunoprécipitation par Ac. anti-ESR1)
 - Un control (~ ADN fragmenté)



Research

GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility

Vasiliki Theodorou,¹ Rory Stark,² Suraj Menon,² and Jason S. Carroll^{1,3,4}

¹Nuclear Receptor Transcription Lab, ²Bioinformatics Core, Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom; ³Department of Oncology, University of Cambridge, Cambridge CB2 0XZ, United Kingdom

Télécharger des fichiers

- On peut utiliser un **navigateur (e.g Cyberduck) pour téléverser sur le serveur**
- **Mieux**, on peut effectuer directement le téléchargement **depuis le terminal** si on dispose de l'**URL**.
 - On utilise alors la commande `wget`.

```
$ cd /shared/projects/<project> # adaptez <project>
$ cd chip-seq/fastq
$ pwd # print working directory
$ srun wget http://denis.puthier.perso.luminy.univ-amu.fr/siNT\_ER\_E2\_r3\_chr21.fastq.gz
$ ls
```

Decompression

- La commande `gunzip`.
 - La commande `gunzip` permet de décompresser un fichier au format `*.gz`. Sa syntaxe générale est la suivante:
 - `gunzip [-cfhkLNqrtVv] [-S suffix] file [file [...]]`

```
$ # on décompresse le fichier *.gz.  
$ srun gunzip siNT_ER_E2_r3_chr21.fastq.gz  
$ # Regardez l'extension du fichier siNT_ER_E2_r3_chr21.fastq.  
$ # Que remarquez vous ?  
$ ls
```

Les lectures brutes (raw reads) sont au format fastq

```
Header      @QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
Sequence    GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+ (optional header) +
Quality     =.+5:<<<<>AA?0A>;A*A#####
```

- La qualité est généralement au format Sanger (cf prochaine diapo).
- Exercice
 - Utilisez une des commandes vues précédemment pour visualiser le contenu du fichier fastq

Les lectures brutes (raw reads) sont au format fastq

Header	@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
Sequence	GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+ (optional header)	+
Quality	=.+5:<<<<>AA?0A>;A*A#####

```
$ # Vous pouvez utiliser la commande less pour visualiser le contenu du fichier.
```

```
$ # q pour quitter
```

```
$ less siNT_ER_E2_r3_chr21.fastq
```

Le score de qualité Sanger

- Une valeur de score Sanger est attribuée à chaque base séquencée
 - Basée sur p , la probabilité d'erreur (i.e. que la base soit fausse)
 - $Q_{Sanger} = -10 \cdot \log_{10}(p)$
 - $p = 0.1 \Leftrightarrow Q_{Sanger} 10$
 - $p = 0.01 \Leftrightarrow Q_{Sanger} 20$
 - $p = 0.001 \Leftrightarrow Q_{Sanger} 30$
 - ...
- Les scores sont encodés en ASCII 33
 - Objectif : compresser les données en diminuant le nombre de caractères utilisés pour encoder la qualité.
- Le score de qualité Sanger varie entre 0 et 40

Le score de qualité Sanger

- ! correspond à 0
- “ correspond à 1
- # correspond à 2
- \$ correspond à 3
- ...
- I correspond à 40

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	\$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(72	48	H	104	68	h
9	09	Horizontal tab	41	29)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	
29	1D	Group separator	61	3D	=	93	5D]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

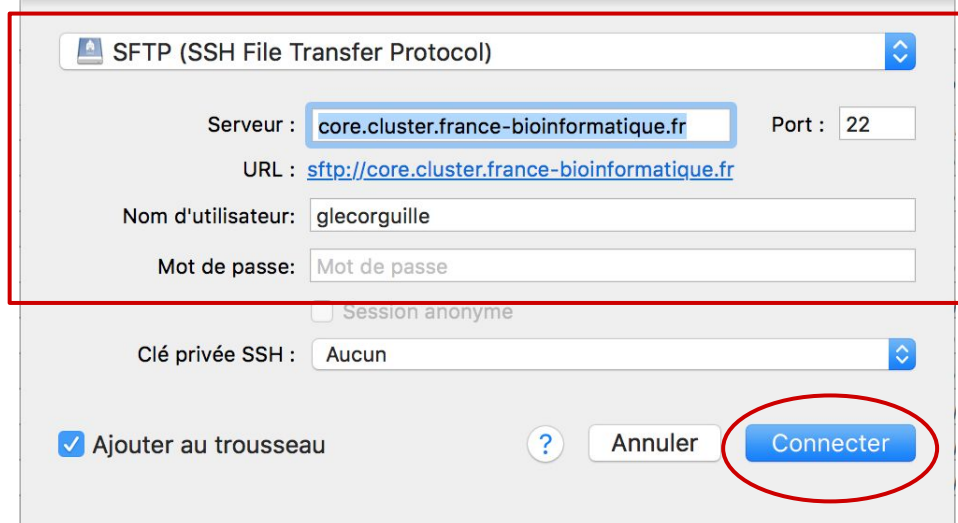
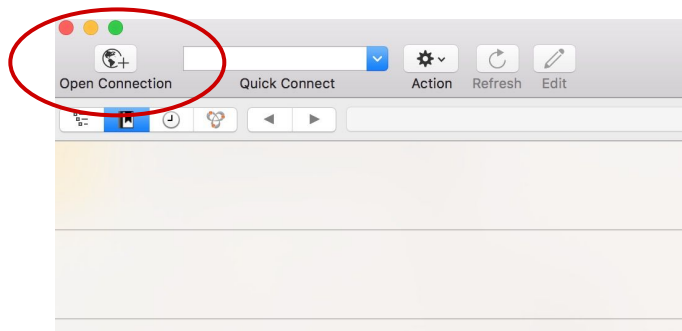
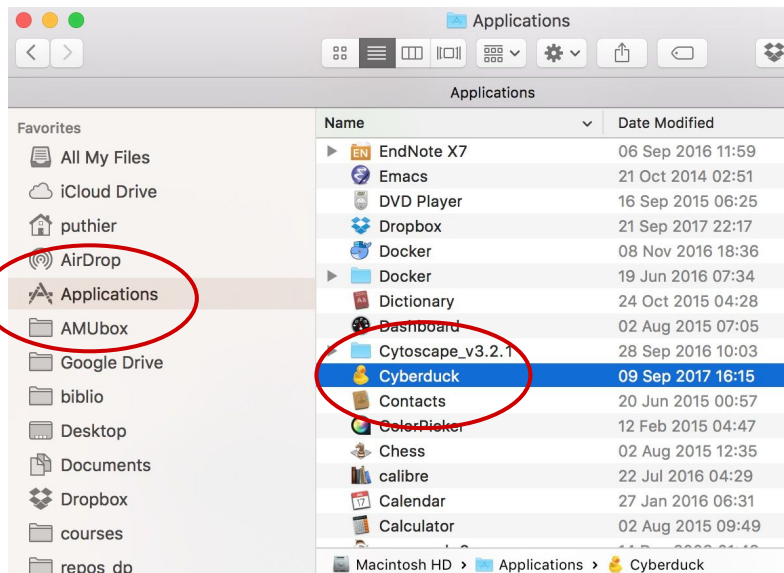
Analyser la qualité avec fastQC

- **Fast Quality Control (FastQC)**

- Propose un certain nombre de diagrammes qualité pour évaluer la qualité du séquençage.

```
$ cd .. # On remonte d'un niveau dans l'arborescence
$ mkdir qc # On créé un répertoire
$ ls -l ; cd qc # Deux instructions sur la même ligne séparées par ';'
$ module load fastqc/0.11.8 # Charge le chemin de fastqc dans l'environnement
$ fastqc -h # Obtenir de l'aide
$ # Lancer fastqc
$ # Ici le \ indique un retour à la ligne mais vous n'êtes pas censé le
$ # taper et aller à la ligne
$ srun fastqc -f fastq -o ./ ../fastq/siNT_ER_E2_r3_chr21.fastq \
    2> siNT_ER_E2_r3_chr21_fastqc.log
$ less siNT_ER_E2_r3_chr21_fastqc.log # la sortie d'erreur de fastqc
$ ls # Que voyez vous ?
```

Télécharger les résultats avec Cyberduck (OSX)



Résultats de FastQC

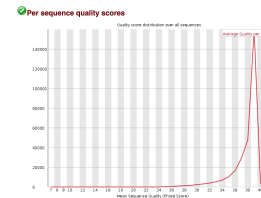
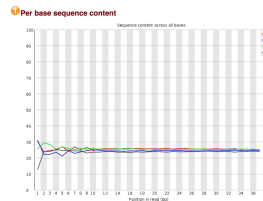
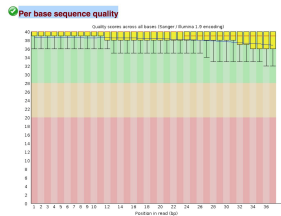
The screenshot shows a web browser window titled "IFB.core.cluster" with the address bar containing the path `/shared/projects/facts/chip-seq/qc`. The browser interface includes navigation buttons like "Ouvrir une connexion", "Connecter rapidement", "Action", "Actualiser", "Éditer", and "Déconnecter". Below the address bar, a table lists files:

Nom du fichier	Taille	Modifié
siNT_ER_E2_r3_chr21_fastqc_log.txt	995 B	31/10/2018 17:04
siNT_ER_E2_r3_chr21_fastqc.html	239.9 KB	31/10/2018 17:04
siNT_ER_E2_r3_chr21_fastqc.zip	297.6 KB	31/10/2018 17:04

The screenshot shows a "Transfers" window with a list of download tasks. The first task is for `siNT_ER_E2_r3_chr21_fastqc.html`, which is marked as "Download complete" and shows a progress bar at 283.2 KB of 283.2 KB, completed on 28 September 2017 at 13:48:10 GMT+2. A second, identical task is shown below it, completed on 28 September 2017 at 13:49:06 GMT+2. At the bottom, the URL `sftp://bioinfo.sb-roscoff.fr/proj...T_ER_E2_r3_chr21_fastqc.html` and the local file path `~/Downloads/siNT_ER_E2_r3_chr21_fastqc.html` are displayed.

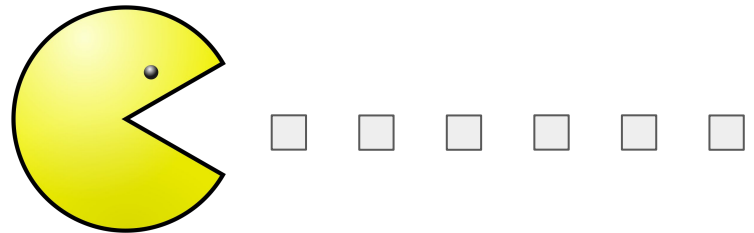
Résultats de FastQC

- Exploration des résultats de fastqc en interactif.
 - A quoi correspond le diagramme “**Per base sequence quality**”.
 - A quoi correspond le diagramme “**Per sequence quality score**” ?
 - A quoi correspond le diagramme “**Per base sequence content**” ?
 - A quoi correspond le diagramme “**Per sequence GC content**” ?
 - A quoi correspond le diagramme “**Per sequence N content**” ?
 - A quoi correspond le diagramme “**Sequence length distribution**” ?
 - A quoi correspond le diagramme “**Sequence duplication level**” ?
 - A quoi correspond le diagramme “**Kmer content**” ?



Rogner les reads

- Une étape de pré-processing
 - Les reads en entrée sont rognés afin d'éliminer des extrémités de mauvaises qualités.
 - En fonction de la capacité de l'outil à faire des alignements locaux ou globaux et de la qualité intrinsèque des données, cette étape peut être cruciale.
 - Risque: peu de reads alignés
- Quelques logiciels existants
 - Sickle-trim (sliding window-based trimming)
 - FASTX-Toolkit (cut a defined number of nucleotides)
 - Trimmomatic
 - Cutadapt



Principe de sickle

- Objectif:
 - **Supprimer** les extrémités de mauvaise qualité.
- Solution:
 - Parcourir le read avec un **fenêtre coulissante** de droite à gauche. Calculer la **qualité moyenne** dans chaque fenêtre
 - Si la valeur de qualité chute en dessous d'une **valeur seuil q** , déléter l'extrémité 3'.
 - Si la taille restante du read est inférieure à une **longueur seuil l** , déléter le read.

ACTCGCTCGCTGGTTAATCGATGATCGTGCAGTCGTA



L'interface de sickle

- Sickle contient plusieurs **sous-commandes**: **pe** et **se**.

```
$ module load sickle-trim/1.33
```

```
$ sickle -h
```

```
Usage: sickle <command> [options]
```

Command:

```
pe  paired-end sequence trimming
```

```
se  single-end sequence trimming
```

```
--help, display this help and exit
```

```
--version, output version information and exit
```

```
$ sickle se --help # Obtenir de l'aide sur la sous-commande se.
```

Exercice (noté)

- Créez un **répertoire trimmed** au même niveau dans l'arborescence que fastq.
- Déplacez vous dans ce répertoire.
- Invoquez l'aide de sickle (se)
- Construisez une commande qui combine les options suivantes:
 - Fournissez à **sickle** le fichier d'entrée **siNT_ER_E2_r3_chr21.fastq**.
 - Qualité de type "Sanger", seuils de qualité et de longueur tous deux à 20.
 - Demandez à sickle se de produire un fichier de sortie que vous nommerez **siNT_ER_E2_r3_chr21_trim.fastq** et qui devra être créé dans le dossier trimmed.
 - Rediriger la sortie standard dans un fichier que vous nommerez **siNT_ER_E2_r3_chr21_sicke_log.txt** placé dans le dossier trimmed.
- Comptez le nombre de lignes présentes dans les fichiers fastq avant et après utilisation de sickle (commande wc -l).
- Lisez le contenu du fichier log. Obtenez-vous le même résultat ?

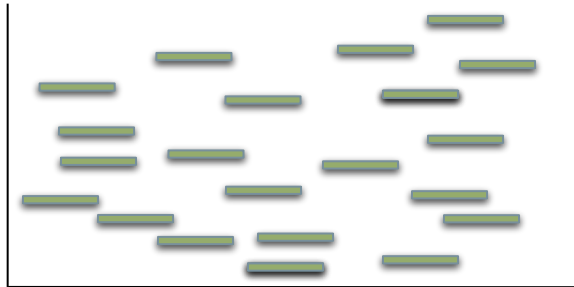
Corrigé

```
$ cd ..          # On remonte d'un niveau dans l'arborescence
$ mkdir trimmed  # On créé un répertoire
$ cd trimmed     # On se déplace dans ce répertoire
$ # On lance sickle
$ # Ici le \ indique un retour à la ligne mais vous n'êtes pas censé le
$ # taper et aller à la ligne
$ # 2> redirige la sortie d'erreur
$ srun sickle se -f ../fastq/siNT_ER_E2_r3_chr21.fastq \
    -t sanger -o siNT_ER_E2_r3_chr21_trim.fastq \
    2> siNT_ER_E2_r3_chr21_sickle.log
$ # le nombre de lignes présentes dans les fichiers fastq
$ wc -l ../fastq/siNT_ER_E2_r3_chr21.fastq    # Données brutes
$ wc -l siNT_ER_E2_r3_chr21_trim.fastq       # Données nettoyées
```

Aligner les reads

- Objectif
 - Trouver la région du génome qui a produit les read.
 - Trouver dans le génome le mot correspondant au read


Ref. Genome

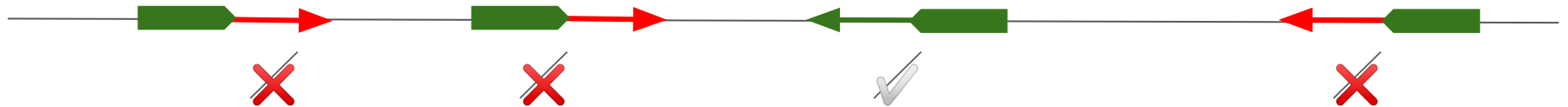


Reads

- Une position dans le génome
- Plusieurs positions (éléments répétés, région de basse complexité)

L'approche de bowtie: *seed and extend*

- Une extrémité du read est interrogée (la graine) 
- On cherche ses régions correspondantes sur le génome (à l'aide d'un index créé initialement) avec ou sans mismatch.
- On teste si le reste du read s'aligne avec la séquence



Aligner les reads

- Pour l'alignement nous utiliserons **Bowtie 2**.
- Bowtie 2 nécessite de préparer un **index**.
 - Cet index permettra une recherche optimisée de la position d'un mot w dans le génome.
 - Des index pour les génomes utilisés classiquement sont disponibles sur le [site de bowtie 2](#).
 - Ici nous voulons restreindre le génome au chromosome 21, nous devons donc construire cet index.

```
# Créez un répertoire pour y stocker l'index dans chip-seq/  
$ cd ..  
$ mkdir index  
$ cd index
```

Création de l'index

- Ne faire qu'une seule fois par génome d'intérêt !
- Allez sur le site de l'UCSC à l'adresse suivante
 - <https://genome.ucsc.edu/>
- Cliquez sur Downloads > Genome Data > human > hg38 > Data set by chromosome.
- Recherchez le fichier **chr21.fa.gz**
- Cliquez bouton droit "Copy link address"

```
$ # Téléchargez l'index avec wget
$ srun wget http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/chr21.fa.gz
$ # décompression
$ srun gunzip chr21.fa.gz
$ module load bowtie2/2.3.4.3 samtools/1.9      # ici on charge 2 outils à la fois
$ # Construction de l'index
$ srun bowtie2-build chr21.fa chr21_hg38
```

Alignement

- On crée un répertoire de travail et on se positionne dans celui-ci
- On lancera l'alignement dans depuis le dossier 'bam'.

```
# Create a directory
```

```
$ mkdir -p ../bam
```

```
# Change directory
```

```
$ cd ../bam
```

Alignement

- On indique à srun de réserver 4 processeurs.

Attention cette ligne de commande est en cours de construction

```
$ srun --cpus=4
```

Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.

```
# Perform alignment
```

```
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log
```


Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.
- Ce flux de texte peut être redirigé (|) vers '**samtools view -hbS**' (-h: header, -b output is BAM, -S: input is SAM) pour produire une version compressée (format **bam**).

```
# -bS (sortie en bam, entrée en sam)
```

```
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log | samtools view -hbS
```

Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.
- Ce flux de texte peut être redirigé (|) vers '**samtools view -hbS**' (-h: header, -b output is BAM, -S: input is SAM) pour produire une version compressée (format **bam**).
- On sélectionne le sous-ensemble des reads pour lequel la mapping quality (-q: quality) est au moins égale à 30.

```
# -q 30 (quality 30)
```

```
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log | samtools view -hbS -q 30
```

Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.
- Ce flux de texte peut être redirigé (**|**) vers '**samtools view -hbS**' (-h: header, -b output is BAM, -S: input is SAM) pour produire une version compressée (format **bam**).
- On sélectionne le sous-ensemble des reads pour lequel la mapping quality (-q: quality) est au moins égale à 30.
- Le flux de texte est redirigé (**|**) vers '**samtools sort**' (trie par coordonnées génomiques).

```
# Trie l'alignement
```

```
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log | samtools view -hbS -q 30 | samtools sort
```

Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.
- Ce flux de texte peut être redirigé (`|`) vers '`samtools view -hbS`' (-h: header, -b output is BAM, -S: input is SAM) pour produire une version compressée (format **bam**).
- On sélectionne le sous-ensemble des reads pour lequel la mapping quality (-q: quality) est au moins égale à 30.
- Le flux de texte est redirigé (`|`) vers '`samtools sort`' (trie par coordonnées génomiques).
- Le flux de texte est redirigé dans un fichier (`>`)

'`>`' est un opérateur de redirection

```
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log | samtools view -hbS -q 30 | samtools sort \
> siNT_ER_E2_r3_chr21_trim.bam
```

Alignement

- On indique à srun de réserver 4 processeurs.
- L'alignement est réalisé avec **bowtie2**, qui produit un flux de texte au format **sam** (texte), **volumineux**.
- Ce flux de texte peut être redirigé (`|`) vers '`samtools view -hbS`' (-h: header, -b output is BAM, -S: input is SAM) pour produire une version compressée (format **bam**).
- On sélectionne le sous-ensemble des reads pour lequel la mapping quality (-q: quality) est au moins égale à 30.
- Le flux de texte est redirigé (`|`) vers '`samtools sort`' (trie par coordonnées génomiques).
- Le flux de texte est redirigé dans un fichier (`>`)
- Le fichier est indexé pour optimiser la recherche la position dans le BAM.

```
# Perform alignment
# -bS (sortie en bam, entrée en sam)
$ srun --cpus=4 bowtie2 -x ../index/chr21_hg38 -U ../trimmed/siNT_ER_E2_r3_chr21_trim.fastq \
2> siNT_ER_E2_r3_chr21_trim_bowtie2.log | samtools view -hbS -q 30 | samtools sort \
  > siNT_ER_E2_r3_chr21_trim.bam
$ srun samtools index siNT_ER_E2_r3_chr21_trim.bam
```

Fichier bam

- SAM: ‘Sequence Alignment/Map’
- BAM: binary/compressed version of SAM
- Stocke les informations liées à l’alignement
 - Coordonnées du read aligné
 - Mapping quality
 - CIGAR String
 - Bitwise FLAG
 - read paired, read mapped in proper pair, read unmapped, ...
 - ...

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

2 Sep 2016

Visualiser le contenu du fichier bam

- Le fichier bam est compressé.
- On peut voir son contenu avec la commande `samtools`.

```
# Visualiser le contenu du fichier bam
# On utilise l'argument -h pour visualiser aussi le 'header'.
# On renvoie le flux de texte dans less.
# On ajoute le paramètre -S pour tronquer les lignes qui excèdent
# la largeur de l'écran
$ srun samtools view -h siNT_ER_E2_r3_chr21_trim.bam | less -S
```

Bitwise flag

- De nombreuses informations sont stockées dans la colonne 2 du fichier SAM/BAM
 - read pairs
 - reads mapped in proper pairs
 - reads unmapped
 - mates unmapped
 - reads reverse strand
 - mates reverse strand
 - first in pair
 - second in pair
 - not primary alignment
 - ...

Bitwise flag

- 000000000001 → $2^0 = 1$ (read paired)
- 000000000010 → $2^1 = 2$ (read mapped in proper pair)
- 000000000100 → $2^2 = 4$ (read unmapped)
- 000000001000 → $2^3 = 8$ (mate unmapped) ...
- 000000010000 → $2^4 = 16$ (read reverse strand)
- 000000010001 → $2^0 + 2^3 = 9$ → (read paired, mate unmapped)
- 000000011001 → $2^0 + 2^2 + 2^3 = 13$...
- ...

Voir [Explained SAM flags](#)

The extended CIGAR string

- Quelques exemples de drapeaux (flag)
 - M match ou mismatch...
 - I Insertion par rapport à la référence
 - D Délétion par rapport à la référence
 - N Espace dans l'alignement (Gap)
- <http://samtools.sourceforge.net/SAM1.pdf>

ATTCAGATGCAGTA
ATTCA--TGCAGTA

5M2D7M

Pourquoi filtrer sur la qualité ?

- Sommes-nous plus confiants
 - dans l'alignement du read **1** ?
 - dans l'alignement read **2** ?

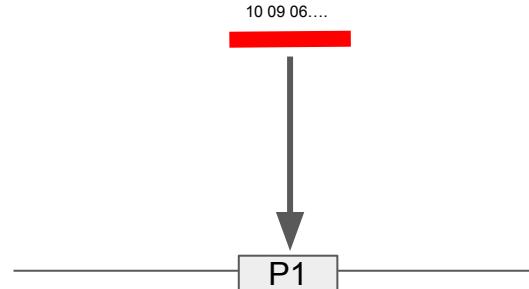
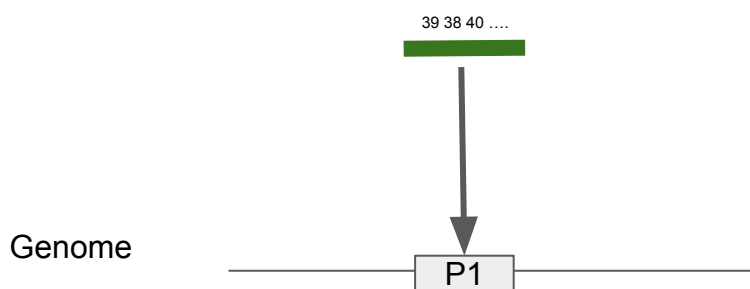


I'm a
uniread

I'm a
multiread

Pourquoi filtrer sur la qualité ?

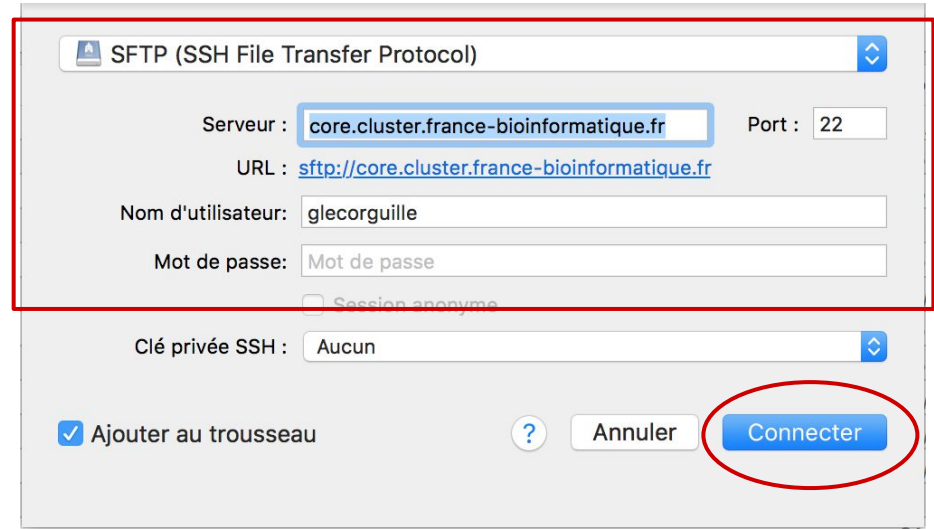
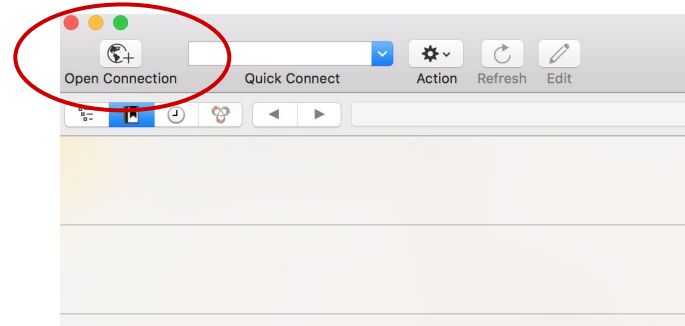
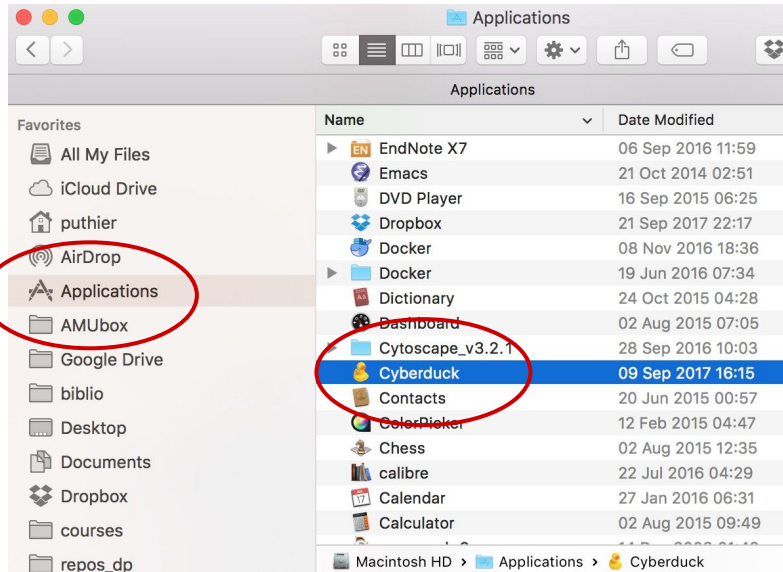
- Sommes-nous plus confiants
 - dans l'alignement **1** ?
 - Si la moyenne de qualité des nucléotides séquencés dans le read est 40
 - dans l'alignement **1'** ?
 - Si la moyenne de qualité des nucléotides séquencés dans le read est 10 ?



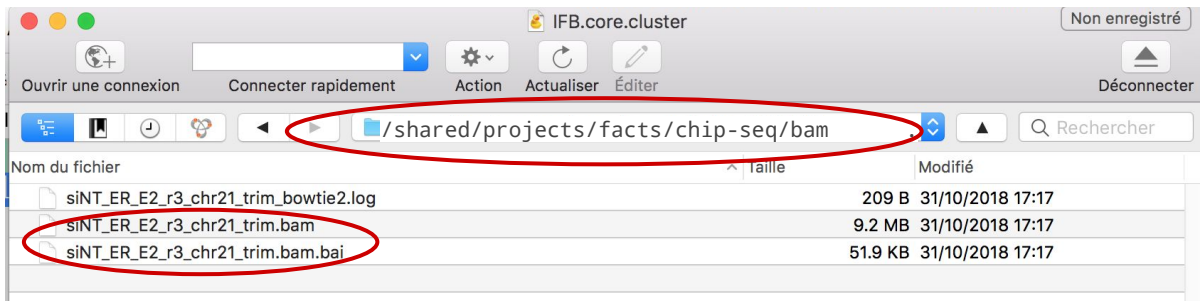
Filtering for Mapping Quality (MAPQ)

- Mapping quality is a score that integrates both the quality of the read itself and the number of positions it maps
- Mapping quality score is computed from the probability that alignment is wrong:
 - takes mappability and sequence quality into account
 - $-10 \cdot \log_{10}(\text{Prob}(\text{alignment is wrong}))$
 - $p=0.01 \rightarrow \text{MAPQ: } 20$
 - $p=0.001 \rightarrow \text{MAPQ: } 30$
 - $p=0.0001 \rightarrow \text{MAPQ: } 40$
 - ...

Télécharger les bams avec Cyberduck (OSX)



Résultats de bowtie2



The screenshot shows a web interface for IFB.core.cluster. The URL bar displays `/shared/projects/facts/chip-seq/bam`, which is circled in red. Below the URL bar is a table of files with columns for 'Nom du fichier', 'taille', and 'Modifié'. Two files are circled in red: `siNT_ER_E2_r3_chr21_trim.bam` and `siNT_ER_E2_r3_chr21_trim.bam.bai`.

Nom du fichier	taille	Modifié
siNT_ER_E2_r3_chr21_trim_bowtie2.log	209 B	31/10/2018 17:17
siNT_ER_E2_r3_chr21_trim.bam	9.2 MB	31/10/2018 17:17
siNT_ER_E2_r3_chr21_trim.bam.bai	51.9 KB	31/10/2018 17:17

Charger le bam dans IGV (i)

IGV
Integrative Genomics Viewer



IGV desktop application



IGV in a web browser



Hi-C contact map viewer



Integrative Genomics Viewer

- Home
- Downloads
- Documents

Hosted Genomes
FAQ
IGV User Guide
File Formats
Release Notes
IGV for iPad
Credits
Contact

Search website

BROAD
INSTITUTE

© 2013-2016 Broad Institute

Home

Integrative Genomics Viewer

Overview

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Citing IGV

Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General

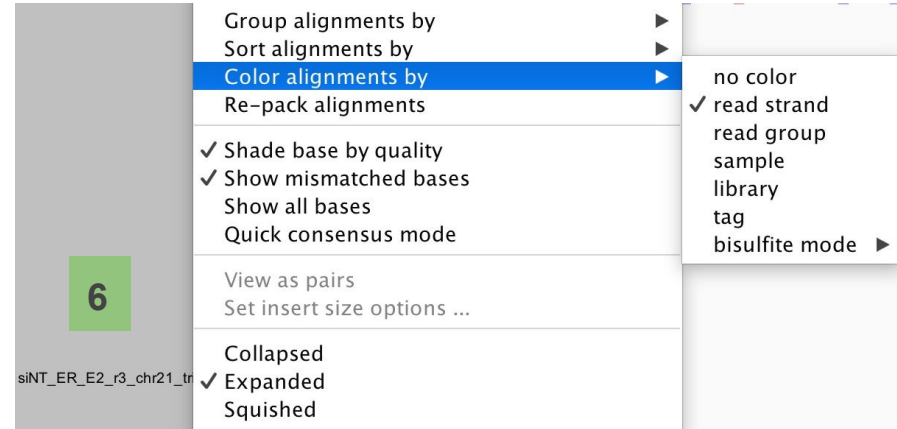
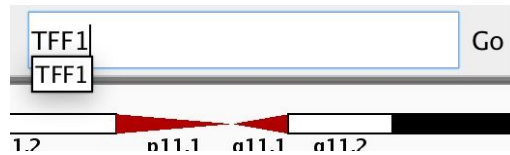
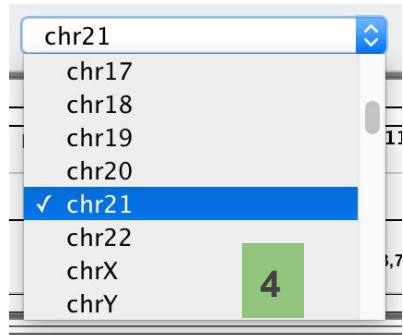
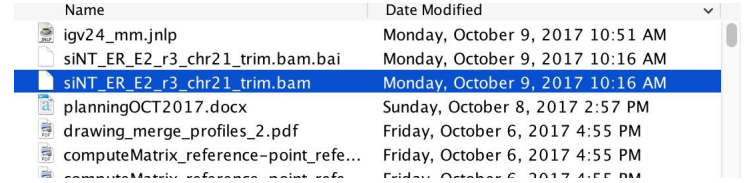
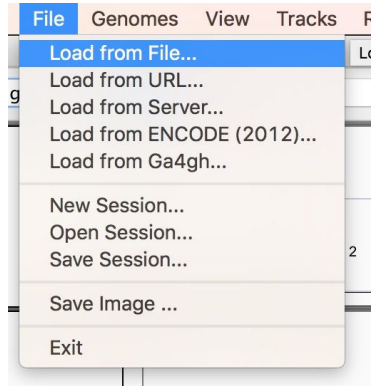
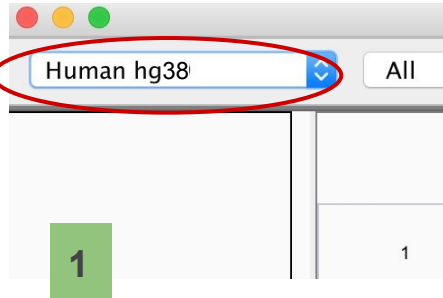
Downloads

Download the IGV desktop application and igvtools.

Funding



Charger le bam dans IGV (ii)

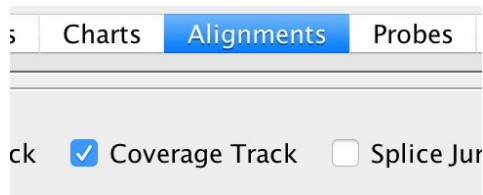
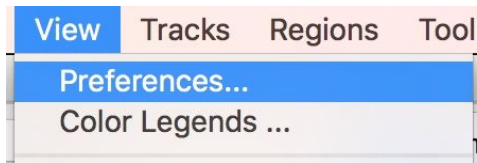


Charger le bam dans IGV (iii)



Augmenter la région visible du BAM

- Attention, très gourmand en RAM...



Visibility range threshold (kb):

Range at which alignments become visible

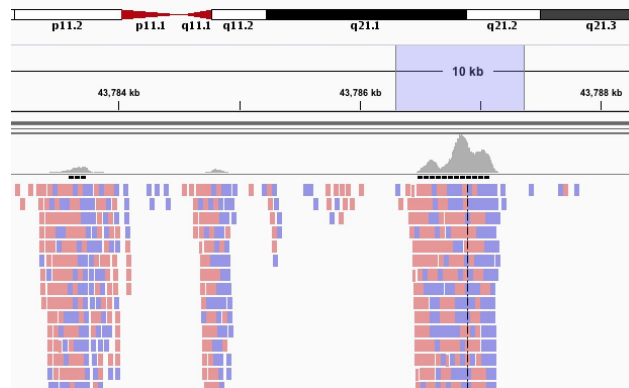
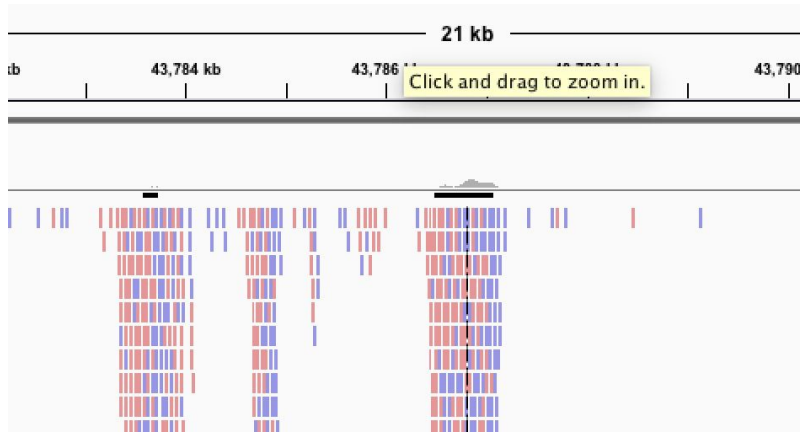
OK



Click + to zoom in, - to zoom out

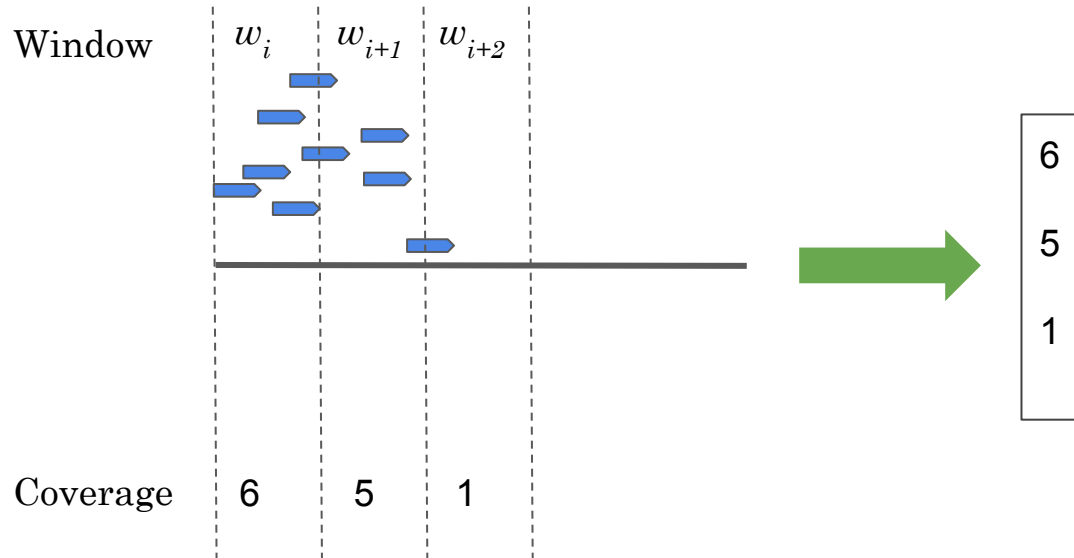
Zoomer dans IGV

- Effectuer un 'click and drag' (cliquer et glisser) dans le panel supérieur.



Produire un fichier de couverture (BigWig)

- Un fichier BAM est volumineux (≥ 11 colonnes)
- On va créer un fichier plus léger contenant 1 colonne (*i.e.* le nombre de reads dans une fenêtre)



Exercice

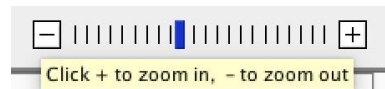
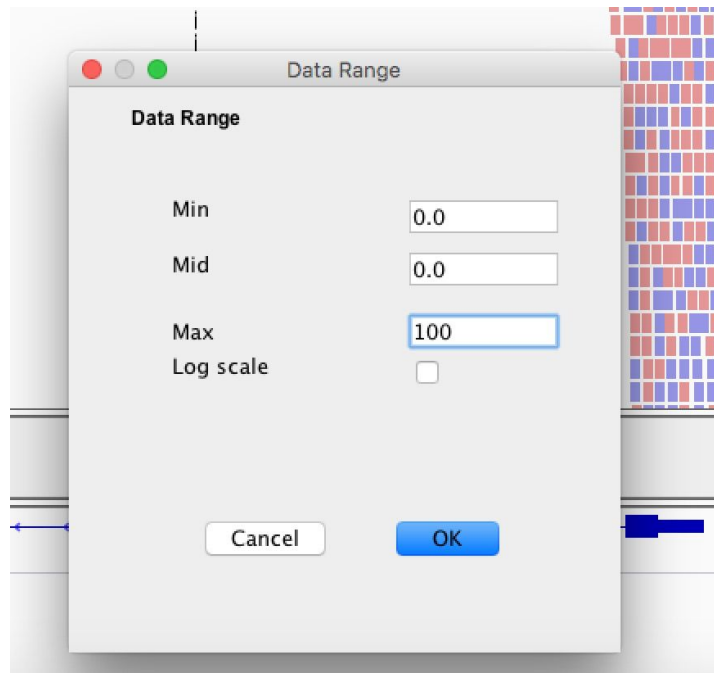
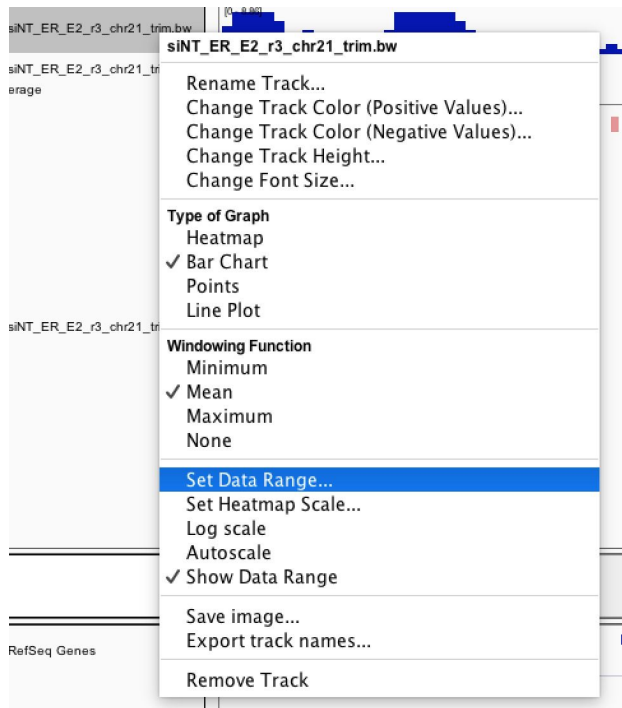
- Créez un répertoire `bwig` au même niveau que le répertoire `fastq`
- Demandez l'aide sur l'outil `bamCoverage` (suite `deeptools`). En utilisant l'argument `-h`.
- Produisez un fichier de couverture au format **bigWig** (`.bw` ou `.bigwig`) en utilisant 1 processeur (regardez bien les arguments disponibles), en fixant la taille des fenêtres (bins) à 25

Exercice

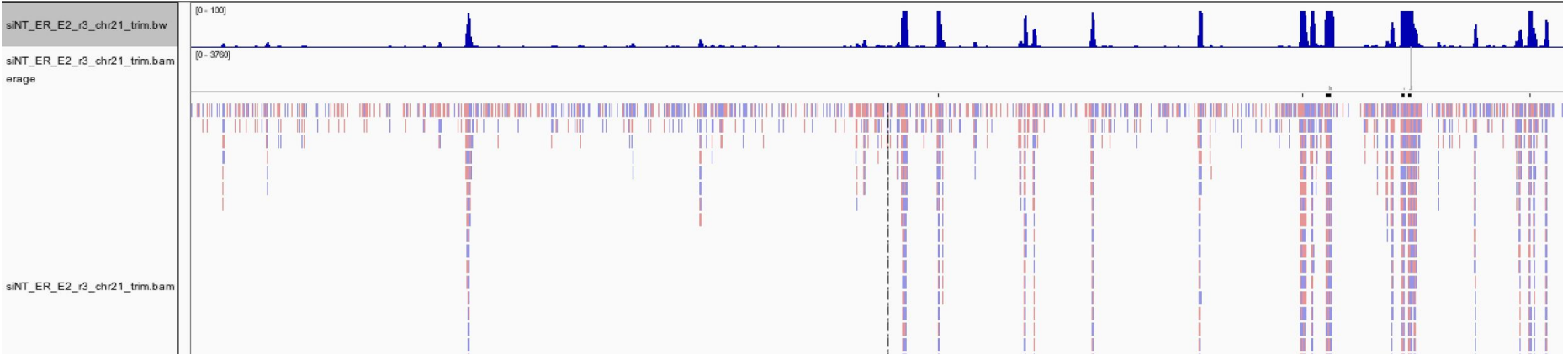
- Créez un répertoire `bwig` au même niveau que le répertoire `fastq`
- Demandez l'aide sur l'outil `bamCoverage` (suite `deeptools`). En utilisant l'argument `-h`.
- Produisez un fichier de couverture au format **bigWig** (`.bw` ou `.bigwig`) en utilisant 1 processeurs (regardez bien les arguments disponible), en fixant la taille des fenêtres (bins) à 25 et en considérant que la taille des fragments était de 300bp.

```
$ mkdir ../bwig          # Create a directory
$ cd ../bwig           # Go into that directory
$ module load deeptools/3.2.0
$ srun --cpus=10 bamCoverage -p 10 -bs 25 -b ../bam/siNT_ER_E2_r3_chr21_trim.bam \
    -o siNT_ER_E2_r3_chr21_trim.bw 2> siNT_ER_E2_r3_chr21_trim.bw.log
```

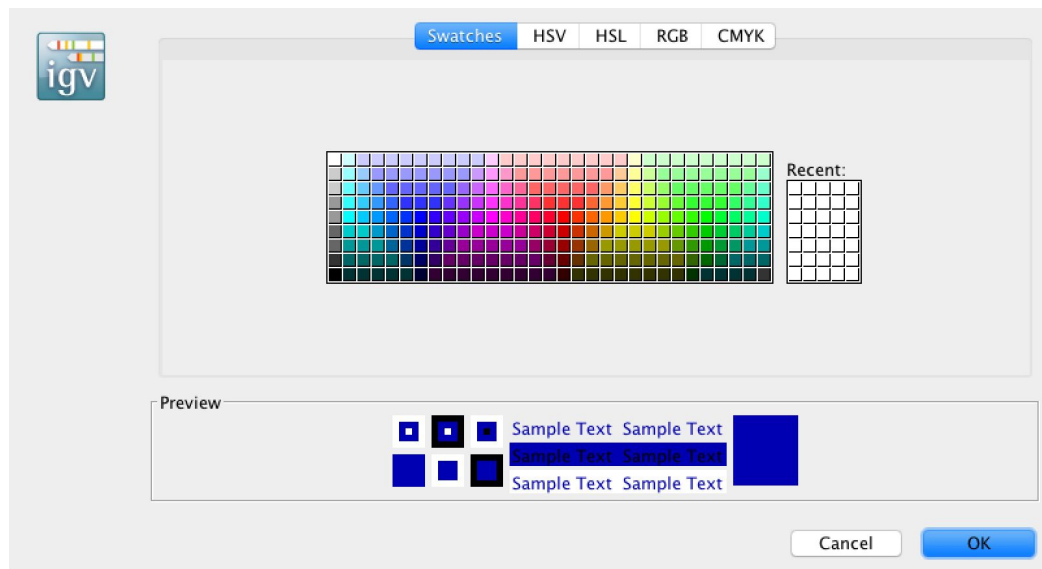
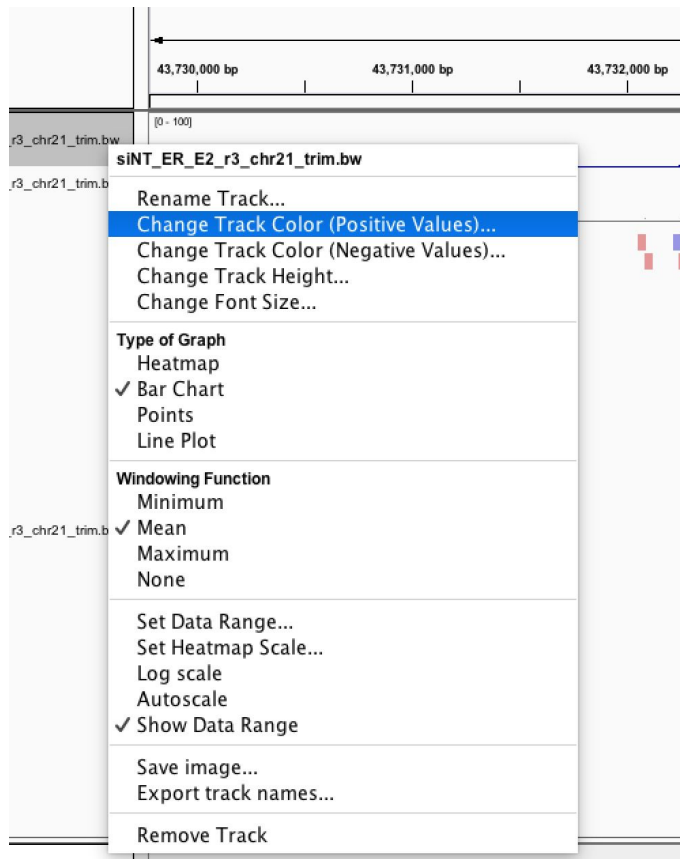
Changer l'échelle d'une piste bigWig



Changer l'échelle d'une piste bigWig



Changer la couleur d'une piste bigWig



Exercice facultatif

- Réalisez le même traitement avec l'échantillon contrôle ([input](#)).

Merci pour votre attention.

**Remerciements à toute l'équipe
pédagogique et technique**

IGV: application web (<https://igv.org/app/>)

The image shows the IGV web application interface. At the top, there are navigation menus for 'Genome', 'Tracks', 'Session', 'Share', 'Bookmark', 'Save SVG', and 'Help'. Below these is a search bar with 'hg38', 'chr21', and coordinates '1:42,361,282-42,367,535' (6,254 bp). A track labeled 'TFF1' is highlighted with a green box and the number '4'. A context menu is open over the track, listing options: 'Set track name', 'Set track height', 'Set track color', 'Color by', 'read strand' (checked), 'tag', 'Show all bases', 'Show soft clips', 'Set visibility window', and 'Remove track'. A green box with the number '5' points to the 'Color by' option, with the text 'Changer la couleur des reads'. Other callouts include '1' pointing to the 'Tracks' menu, '2' to the 'Local File ...' button, and '3' to the 'Session' menu. The main view shows a genomic track with a red double-headed arrow indicating a zoomed-in region. Below this is a read alignment track with blue and red reads. At the bottom, a 'Refseq Genes' track shows the 'TFF1' gene structure.

1

2

3

4 TFF1

5 Changer la couleur des reads

Genome Tracks Session Share Bookmark Save SVG Help

IGV hg38 chr21 1:42,361,282-42,367,535 6,254 bp

Cursor Guide Center Line Track Labels

42,362 kb 42,363 kb 42,364 kb 42,365 kb 42,366 kb 42,367 kb

siNT_ER_E2_r3_chr21_trim....

596

0

Refseq Genes

TFF1

- Set track name
- Set track height
- Set track color
- Color by
- read strand
- tag
- Show all bases
- Show soft clips
- Set visibility window
- Remove track