

Analyses RNA-seq

TP

Vonick SIBUT, PhD
Bioinformatique et Biostatistique
vonick.sibut@univ-rennes.fr



Plan du cours

01

Question scientifique,
design expérimental et jeu
de données

02

Données brutes : contrôle
qualité, data cleaning

03

Alignement des données
brutes, qualité des
alignements, visualisation
IGV

04

Comptage des reads
mappés

05

Analyse différentielle avec
DESeq2



01

Question scientifique,
design expérimental et
jeu de données



La question scientifique

- **Question** : Quels sont les **gènes qui différencient** les sous-populations **cellulaires stromales** chez les individus sains ?
- **Hypothèse** : les expressions de certains transcrits sont significativement différentes entre les sous-populations cellulaires étudiées
- **Population** : individus sains (homo sapiens)
- Trois types de cellules stromales :
 - ✓ **FDC** : cellules dendritiques folliculaires , situées dans les centres germinatifs et les follicules primaires
 - ▶ marquées par GP38+ CD21Lpos
 - ✓ **FRC49a** : cellules réticulaires fibroblastiques, situées dans la zone T
 - ▶ marquées par GP38+ CD49a+
 - ✓ **DN** : double négative : cellules les plus immatures, proches des cellules progénitrices, situées sur le péricyte des cellules endothéliales des amygdales
 - ▶ marquées par GP38neg CD21Lneg

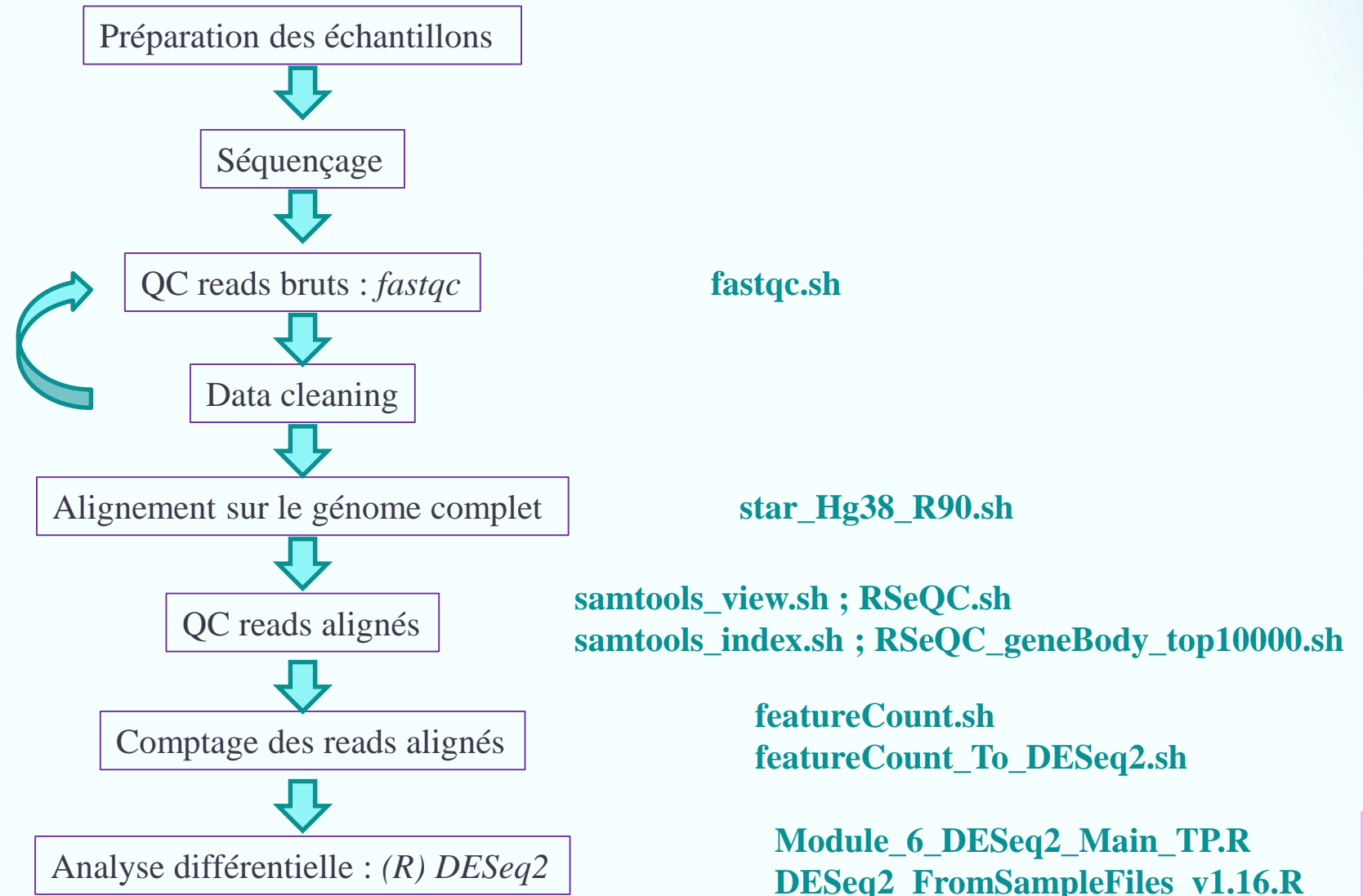
Le design expérimental

- Le **matériel** prélevé consiste en :
 - ✓ les cellules stromales **triées** des 3 sous-populations FDC, FRC49a et DN (contrôle)
 - ✓ provenant de 4 amygdales de sujets sains
 - ✓ kit de préparation des librairies : SMARTer Stranded RNA-Seq Kit
 - ✓ RNA-seq :
 - ✓ ILLUMINA Hiseq 2500
 - ✓ HiSeq - Rapid Run - PE100
 - ✓ paired end
 - ✓ 2 runs

PIPELINE

G
E
N
O
U
E
S
T

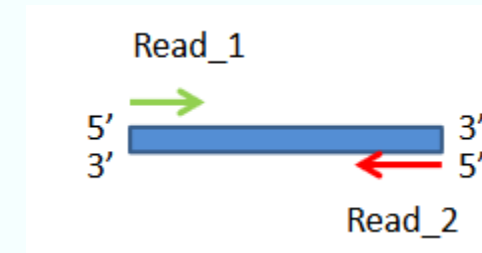
local



Préparation des librairies et séquençage

Projet Cellules stromales

- Préparation des librairies :
 - ✓ SMARTer Stranded RNA-Seq Kit
 - ✓ Double stranded
- Séquençage :
 - ✓ Illumina HiSeq 2500
 - ✓ paired-end reads .R1, .R2
 - ✓ format des fichiers : fastq
 - ✓ description 4 lignes
 - @SEQ identifier
 - Sequence {A,T,G,C}
 - + description
 - Sequence quality controls (pour chaque base)
 - ✓ head xxx.fastq



Premiers pas sur Genouest - 1

Quelques commandes Unix

- **pwd** (*print working directory*) affiche le chemin du répertoire courant.
- **ls** (*list*) affiche le contenu du répertoire courant.
 - **-ls -a** affiche également les fichiers cachés qui commencent par un **.** sous linux
 - **-ls -l** affiche des informations supplémentaires comme la date et la taille des fichiers
 - **-ls -t** trie les fichiers par ordre de dernière modification
- **cd** (*change directory*) permet de se déplacer dans le système de fichiers en changeant de répertoire courant.
 - **cd monrep** fait du répertoire **monrep** le répertoire courant.
 - **cd ..** permet de remonter d'un niveau dans l'arborescence.
 - **cd /** permet de revenir à la racine de l'arborescence.

Premiers pas sur Genouest - 2

Quelques commandes Unix

- **mkdir** (*make directory*) crée un nouveau répertoire.
- **cp** (*copy*) copie les fichiers ou les répertoires.
 - **cp fic1.txt monrep/** copie le fichier fic1.txt dans le répertoire monrep.
 - **cp fic1.txt fic2.txt** duplique le fichier fic1.txt sous le nom fic2.txt.
- **mv** (*move*) déplace ou renomme des fichiers ou des répertoires.
 - **mv fic1.txt monrep/** déplace le fichier fic1.txt dans le répertoire monrep.
 - **mv fic1.txt fic2.txt** renomme le fichier fic1.txt en fic2.txt.
- **rm** (*remove*) supprime des fichiers. **rm -r** supprime des répertoires.
 - **rm fic1.txt** supprime le fichier fic1.txt.
 - **rm -r monrep** supprime le répertoire monrep ainsi que tout son contenu.

Premiers pas sur Genouest - 3

Quelques commandes Unix

- **df -h** affiche la taille de l'espace disque occupée et la taille de l'espace du disque libre.
- **du -h /home/genouest/inserm_u1236/vsibut**
affiche la taille d'un répertoire et de tous les sous répertoires récursifs qu'il contient (disk usage)
- **man <cmd>** affiche la documentation d'une commande unix avec toutes les options
 - pour sortir du **man**, taper **q** pour **quit** : **q <enter>**

Scripts shell

Vous travaillerez chacun à partir de votre propre répertoire sur genouest = répertoire de travail personnel

- créer votre répertoire `/home/genouest/tp_gnf_rnaseq_40965/tp600XX`
- copier tous les scripts .sh du répertoire `/TP_RNAseq` dans votre répertoire
=> commandes unix `cd [] /home/genouest/ tp_gnf_rnaseq_40965/tp600XX`
`cp [] /home/genouest/inserm_u1236/vsibut/TP_RNAseq/*.sh [] .`
- créer 4 sous-répertoires sous votre répertoire de travail
`cd [] /home/genouest/tp_gnf_rnaseq_40965/tp600XX`
`mkdir [] 2_fastQC`
`mkdir [] 3_Align_STAR`
`mkdir [] 4_RSeQC`
`mkdir [] 5_featureCounts`

[] Représente un espace

Scripts shell

A vous de jouer

- Soumettre un script

sbatch **--cpus-per-task=8** **--mem=50G** **-o** **nom_du_job.out** **mon_script.sh**

- Vérifier le statut du job

squeue **-u** **tp600XX**

```
[vsibut@genossh:~] $ squeue -u vsibut
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST (REASON)
-------	-----------	------	------	----	------	-------	-------------------

- « Killer » un job

scancel **jobId**

 Représente un espace

Scripts shell

Quelques commandes Unix

<https://fr.wikipedia.org/wiki/Chmod>

- **chmod** change les permissions d'accès d'un fichier ou d'un répertoire
exemple: `chmod 777 vsibut` : tous les droits à tous [user group others]

r (4) autorisation de lecture
w (2) autorisation d'écriture
x (1) autorisation d'exécution

Correspondances de représentation des droits		
Droit	Valeur alphanumérique	Valeur octale
aucun droit	---	0
exécution seulement	--x	1
écriture seulement	-w-	2
écriture et exécution	-wx	3
lecture seulement	r--	4
lecture et exécution	r-x	5
lecture et écriture	rw-	6
tous les droits (lecture, écriture et exécution)	rwX	7

Scripts shell

Quelques commandes Unix

- **chmod** [OPTIONS] [u g o a] [+ - =] [r, w, x] fichier

Le première groupe de paramètres définit à quelle catégorie vous appliquez les permissions.

Les catégories	Description
u (user)	Affecter les permissions à l'utilisateur
g (group)	Affecter les permissions au groupe
o (other)	Affecter les permissions à autre
a (all)	Pour appliquer à toutes les catégories

Les appartenances de fichiers sur Linux

Le deuxième ensemble d'options définit l'action, si vous ajoutez ou supprimez des permissions.

Drapeaux	Description
-	Supprime les autorisations de fichier à partir d'un utilisateur spécifié.
+	Ajoute des autorisations à un utilisateur spécifié.
=	Attribue des autorisations distinctes des utilisateurs spécifiés et supprime les autorisations précédentes du segment utilisateur.

Genouest – Aide en ligne - 1

<https://www.genouest.org/>

The screenshot shows the Genouest bioinformatics website. The header includes the site name and a navigation menu with links: HOME, TOOLS, HOW TO, EXPERTISE AND TRAINING, ABOUT US, and ACCOUNT. A red box highlights the 'Computing resources' link in the 'TOOLS' dropdown menu, with a red arrow pointing to it. A callout box on the right provides instructions: 'Tools > Computer Ressources', 'Puis dans le paragraphe 'cluster' cliquer sur **here**', and a list of services: '* procédure de connexion', '* data storage', '* software', and '* launching jobs on the cluster'. The main content area features a large 'G' logo and the text 'Development, expertise and resources for bioinformatics'. At the bottom, there are three sections: 'Computing resources' (with a server rack image), 'Expertise and Training' (with a code snippet image), and a chat button labeled 'Chat with a GenOuest agent'.

GenOuest bioinformatics

HOME TOOLS HOW TO EXPERTISE AND TRAINING ABOUT US ACCOUNT

Computing resources

GoDocker

Galaxy@GenOuest

Cloud

CeSGO : main portal

Hosted resources and tools

Tools > Computer Ressources

Puis dans le paragraphe 'cluster' cliquer sur **here**

- * procédure de connexion
- * data storage
- * software
- * launching jobs on the cluster

Development, expertise and resources for bioinformatics

Computing resources

We offer several ways to carry out your data analysis on our

Expertise and Training

We can provide support for your

Chat with a GenOuest agent

Genouest – Aide en ligne - 2

STORAGE

Each user has access to a home directory (limited to 100 GB) and to two temporary directories (/omaha, initially limited to 120 GB but extensible on request and /scratch, limited to 250 GB).

For scientific teams another directory named /groups is available (to access to this directory ask to your team leader).

An online storage is also available at <https://genostack-data.genouest.org>, a Dropbox like service. It can be used via web API or using swift command line client from the internet and from the cluster and can also be used to share some files publicly or privately with external users.

/home and /groups have snapshots and are not intended for computation. /omaha and /scratch have no snapshots and are intended for computation. Files on omaha and scratch are not supposed to be stored for a long period and may be deleted.

Genouest – Aide en ligne - 3

Softwares installés sur le serveur Genouest

Software

- un grand nombre de programmes sont pré-installés dans */softs/local*
- aller sur **software manager** sur le site de genouest pour accéder à la liste des softwares
<http://www.genouest.org/outils/softwaremanager/>

Genouest – Aide en ligne - 4

www.genouest.org/outils/softwaremanager/ 90 % Rechercher

Trello BiblioInserm cpt Google ENT Mail ENT NAS OMIC analysis Genouest HIPC tools Chipster GenOuest account R RNA-seq Analysis Outils Web CYTOF analysis MAF Analysis Alt Splicing Libreoffice

[Login](#)
[How does it work?](#)

Filter by keyword (english terms) _____
 or resource name : _____
 🔍

OR

Filter by operation : _____

[remove filters](#)

[A](#) - [B](#) - [C](#) - [D](#) - [E](#) - [F](#) - [G](#) - [H](#) - [I](#) - [J](#) - [K](#) - [L](#) - [M](#) - [N](#) - [O](#) - [P](#) - [Q](#) - [R](#) - [S](#) - [T](#) - [U](#) - [V](#) - [W](#) - [X](#) - [Y](#) - [Z](#)

Name	Description	Link	EDAM category
FRC_align	FRCbam is a tool able to evaluate and analyze de novo assembly/assemblers. The tool has been already successfully applied in several de novo . FRCbam: tool to compute Feature Response Curves in order to validate and rank assemblies and assemblers		
FastMe	Distance Based Phylogeny Reconstruction Packages		
FastQ Screen	FastQ Screen is a simple application which allows you to search a large sequence dataset against a panel of different genomes to determine from where the sequences in your data originate. It was built as a QC check for sequencing pipelines but may also be useful in characterising metagenomic samples. When running a sequencing pipeline it is useful to know that your sequencing runs contain the types of sequence they're supposed to. Your search libraries might contain the genomes of all of the organisms you work on, along with PhiX, Vectors or other contaminants commonly seen in sequencing experiments.		
FastQC	A quality control tool for high throughput sequence data.		
FindPeaks	The FindPeaks application can be used for several purposes: Converting Eland, Maq (.map), BED or other files into WIG files (See Supported Input formats) Identifying areas of enrichment (ChIP-Seq analysis) Noise estimation in quantifying enrichment (in progress)		
Flexbar	Flexbar is a software to preprocess high-throughput sequencing data efficiently. It demultiplexes barcoded runs and removes adapter sequences. Moreover, trimming and filtering features are provided. Flexbar increases mapping rates and improves genome and transcriptome assemblies. It supports next-generation sequencing data from Illumina, Roche 454, and the SOLID platform. Recognition is based on exact overlap sequence alignment.		
FreeBayes	FreeBayes is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms)		

Genouest – Aide en ligne - 5

[Login](#)
[Back to the list](#)

FastQC

Description	A quality control tool for high throughput sequence data.
Version(s)	0.8 0.10 0.11.7
Environment(s)	/local/env/envfastqc-0.11.7.sh
Directory	/local/FastQC/
Web version	
Comment	

Pour utiliser un programme installé, il faut charger son environnement dans le script d'exécution.
Ceci configurera automatiquement PATH, les librairies dans votre environnement shell.
Ici fastQC sera exécuté avec la version 0.11.7

Le répertoire où se trouve le programme est indiqué ici

ls /local/FastQC

La liste des environnements s'affiche par **ls /softs/local/env/env***

Premiers pas sur Genouest – Premier script shell - 1

1er script shell qui affiche Bonjour !

- écriture du script shell : **l'éditeur de texte vi**

Editeur vi <insert> pour écrire
 <ESC> pour passer en mode commande
 :wq! sauver et sortir de l'éditeur de texte
 :q! sortir sans sauvegarder le fichier
 dd supprimer de la ligne où se trouve le curseur

- exécution du script sur le cluster de calcul
- affichage du status du job

<http://www.linux-france.org/prj/support/outils/vi.html>

Premiers pas sur Genouest – Premier script shell - 1

A vous de jouer

- aller sous votre répertoire sous /groups/inserm_u1236/Formation
- taper **vi test.sh** : le fichier s'ouvre puis sortir sans sauver
- vérifier que le fichier **test.sh** n'a pas été créé


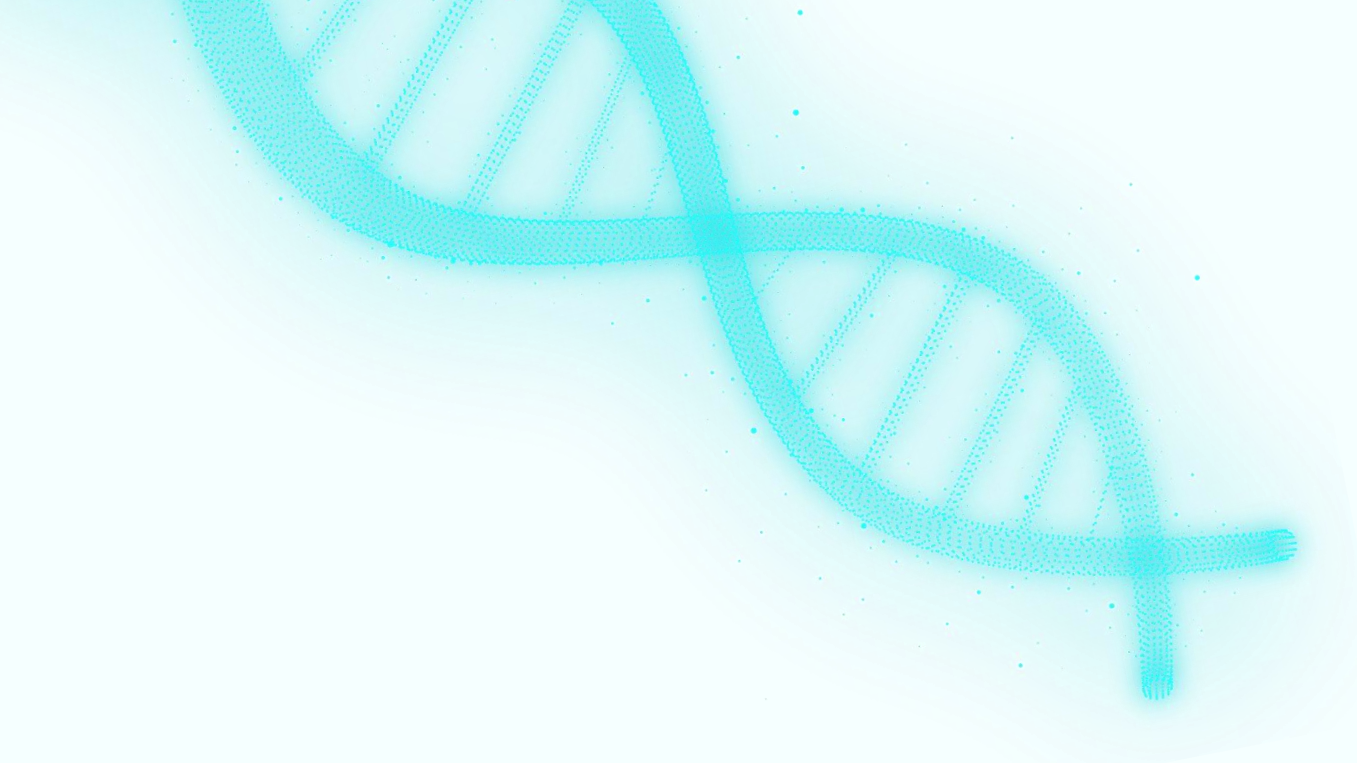
Premiers pas sur Genouest – Premier script shell - 2

A vous de jouer

- [illegible]


[illegible]

```
[drossill@genossh:Module_6] $ more test.sh
#!/bin/bash
echo Bonjour!
```



02

Données brutes : contrôle qualité, data cleaning



Données brutes - Contrôle qualité des reads - 1

Utilisation de fastQC

Vous effectuerez tout le TP sur votre répertoire */home/genouest/tp_gnf_rnaseq_40965/tp600XX*

- **Documentation** <https://wiki.gacrc.uga.edu/wiki/FastQC>

- **Utilisation**

fastqc [-o output dir] [-f fastq|bam|sam] seqfile1... seqfileN

-o --outdir creates all output files in the specified output directory

-f --format bypasses the normal sequence file format detection and forces the program to use the specified format

-j --java provides the full path to the java binary you want to use to launch fastqc

- **Paired-end sequencing**

- ✓ deux reads .R1 et .R2 pour chaque échantillon

- ✓ les deux fichiers sont à analyser avec fastqc pour chaque échantillon

Données brutes - Contrôle qualité des reads - 2

A vous de jouer

- ouvrir le script **fastqc.sh**, le modifier en indiquant votre répertoire personnel sous */home/genouest/tp_gnf_rnaseq_40965/tp600XX* comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

```
vsibut@genossh:/groups/inserm_u1236/Module_6/TP_RNA-seq
#!/bin/bash

# initialisation environnement : aucun pour FastQC
# attention: java-1.7.0_01 et java-1.8.0 changent les car normaux en car speciaux sur les figures de rendu
. /softs/local/env/envjava-1.6.0.sh

#####
# Contrôle qualité des raw reads avec fastQC
#####

date

# QC Read 1 de l'échantillon T05
/local/FastQC/FastQC/fastqc /groups/tp40734/TP_RNAseq/1_Brut/A837-A838T05.R1.fastq -o /groups/tp40734/Nom_Prénom/2_fastQC

# QC Read 2 de l'échantillon T05
/local/FastQC/FastQC/fastqc /groups/tp40734/TP_RNAseq/1_Brut/A837-A838T05.R2.fastq -o /groups/tp40734/Nom_Prénom/2_fastQC

date
~
~
```

Données brutes - Contrôle qualité des reads - 3

Résultat de **fastqc.sh** pour A837-A838T05.R1_fastq

```
[vsibut@genossh:2_FastQC] $ ll
total 664
-rwxr-xr-x 1 vsibut inserm_u1236 299865 Oct  5 23:14 A837-A838T05.R1_fastqc.html
-rwxr-xr-x 1 vsibut inserm_u1236 376792 Oct  5 23:14 A837-A838T05.R1_fastqc.zip
[vsibut@genossh:2_FastQC] $
```


Données brutes - Contrôle qualité des reads - 4

fastqc.out

```
[drossill@genossh:drossill] $ more fastqc.sh.o5866424
Thu Apr 26 15:12:52 CEST 2018
Analysis complete for A837-A838T05.R1.fastq
Thu Apr 26 15:16:10 CEST 2018
```

Données brutes - Contrôle qualité des reads - 5

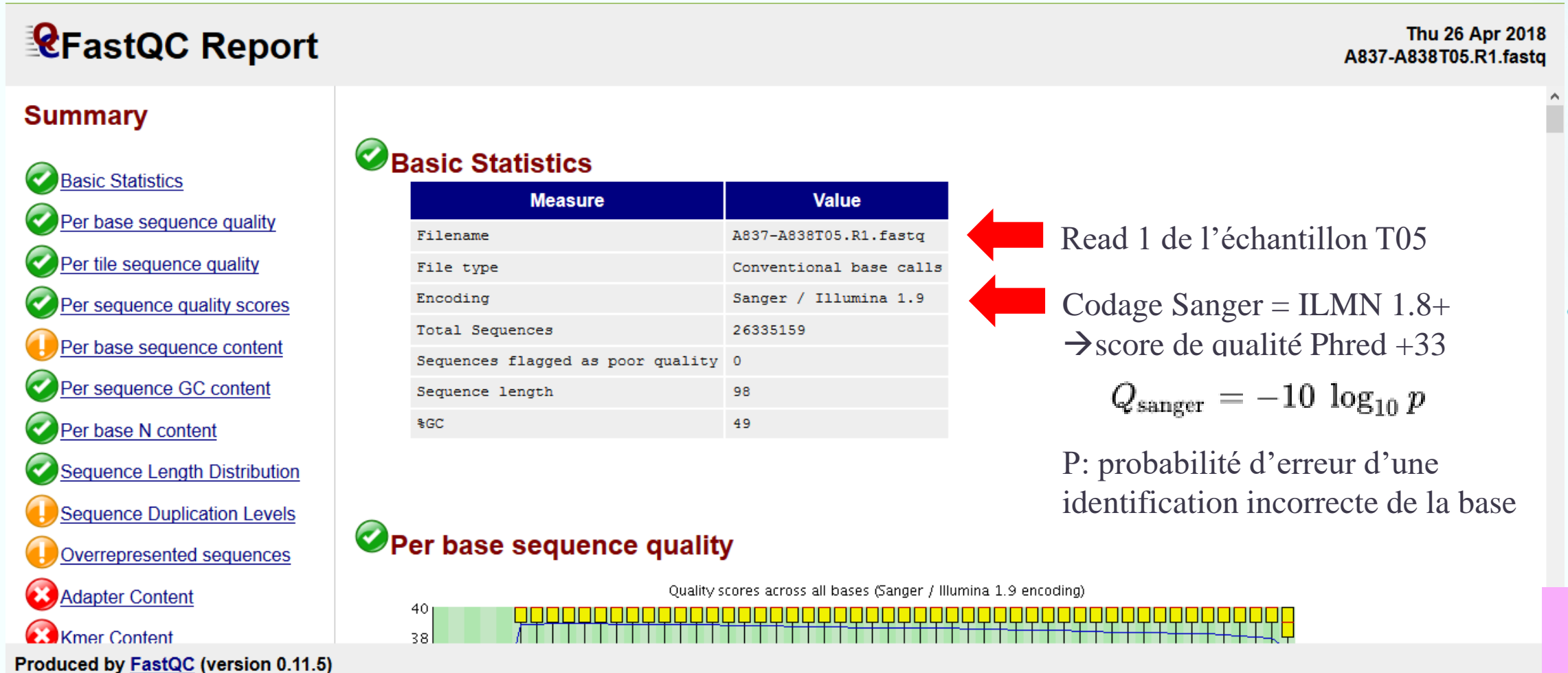
A vous de jouer

Il n'est pas possible de visualiser les résultats html sous Genouest.

- télécharger les fichiers générés sur votre ordinateur local (Filezilla)
- ouvrir le fichier .html sur votre ordinateur local

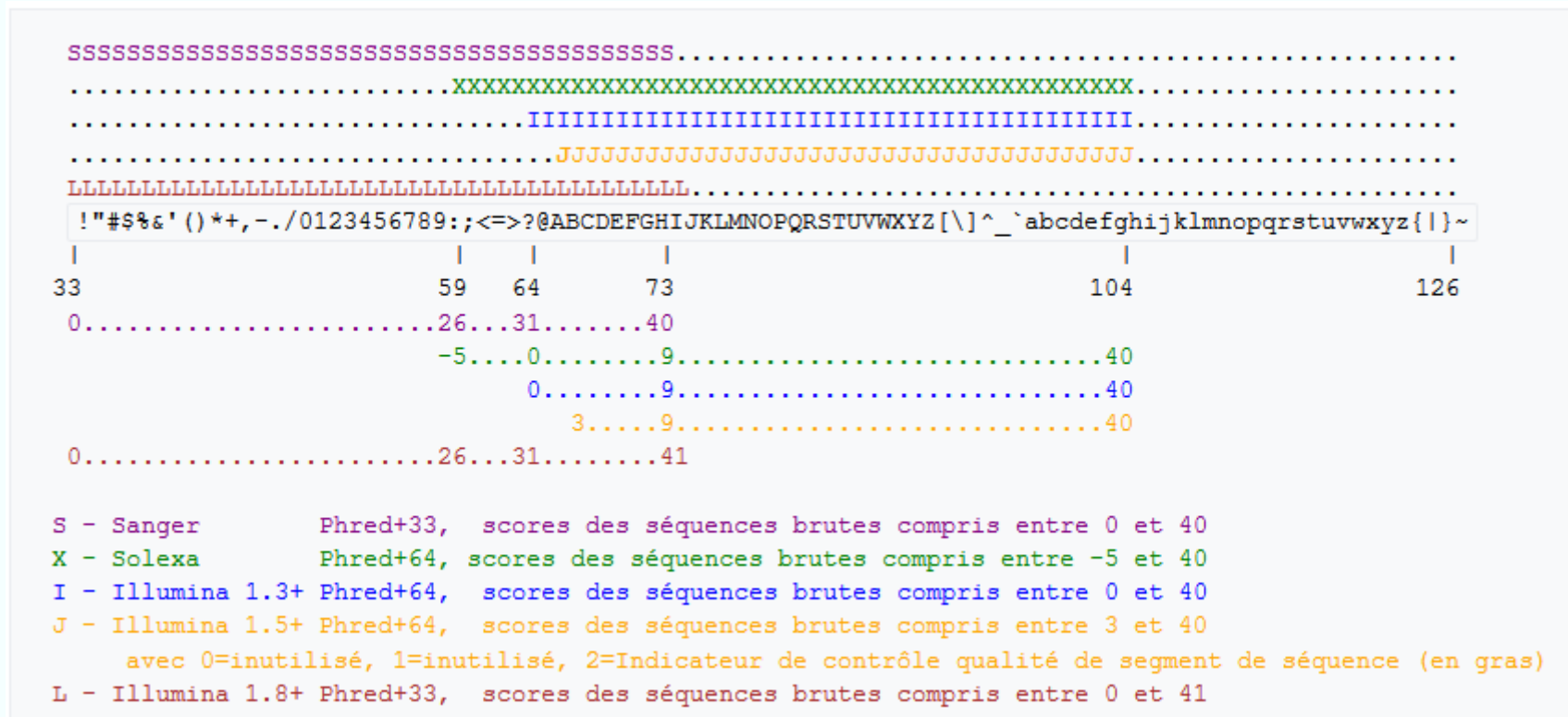
```
[vsibut@genossh:2_FastQC] $ ll
total 664
-rwxr-xr-x 1 vsibut inserm_u1236 299865 Oct  5 23:14 A837-A838T05.R1_fastqc.html
-rwxr-xr-x 1 vsibut inserm_u1236 376792 Oct  5 23:14 A837-A838T05.R1_fastqc.zip
[vsibut@genossh:2_FastQC] $
```

Données brutes - Contrôle qualité des reads - 6

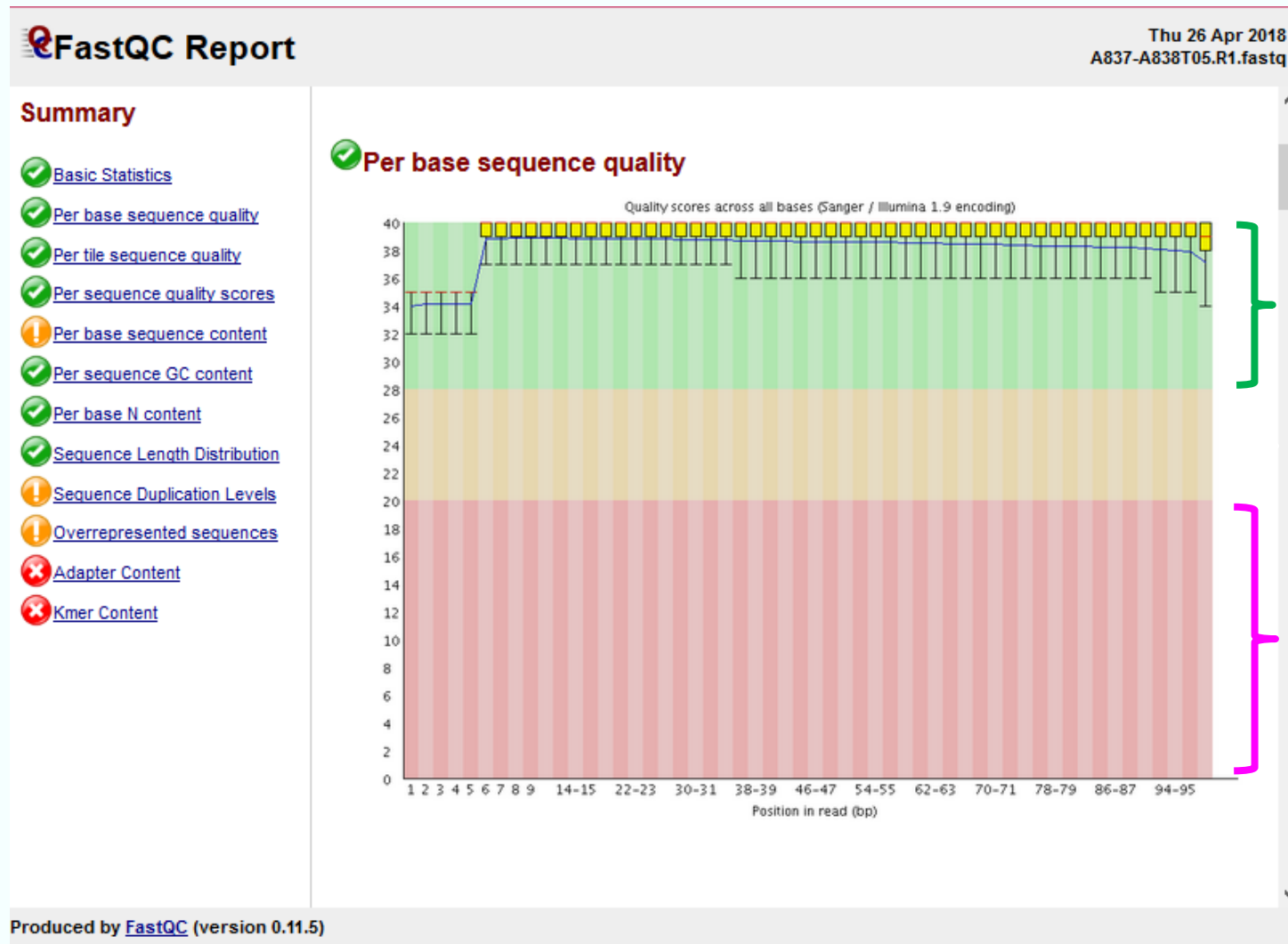


31

[https://fr.wikipedia.org/wiki/FASTQ#Codage du score de qualit%C3%A9](https://fr.wikipedia.org/wiki/FASTQ#Codage_du_score_de_qualit%C3%A9)



Données brutes - Contrôle qualité des reads - 8



Très bonne qualité

Qualité moyenne

Qualité pauvre

Rouge : médiane

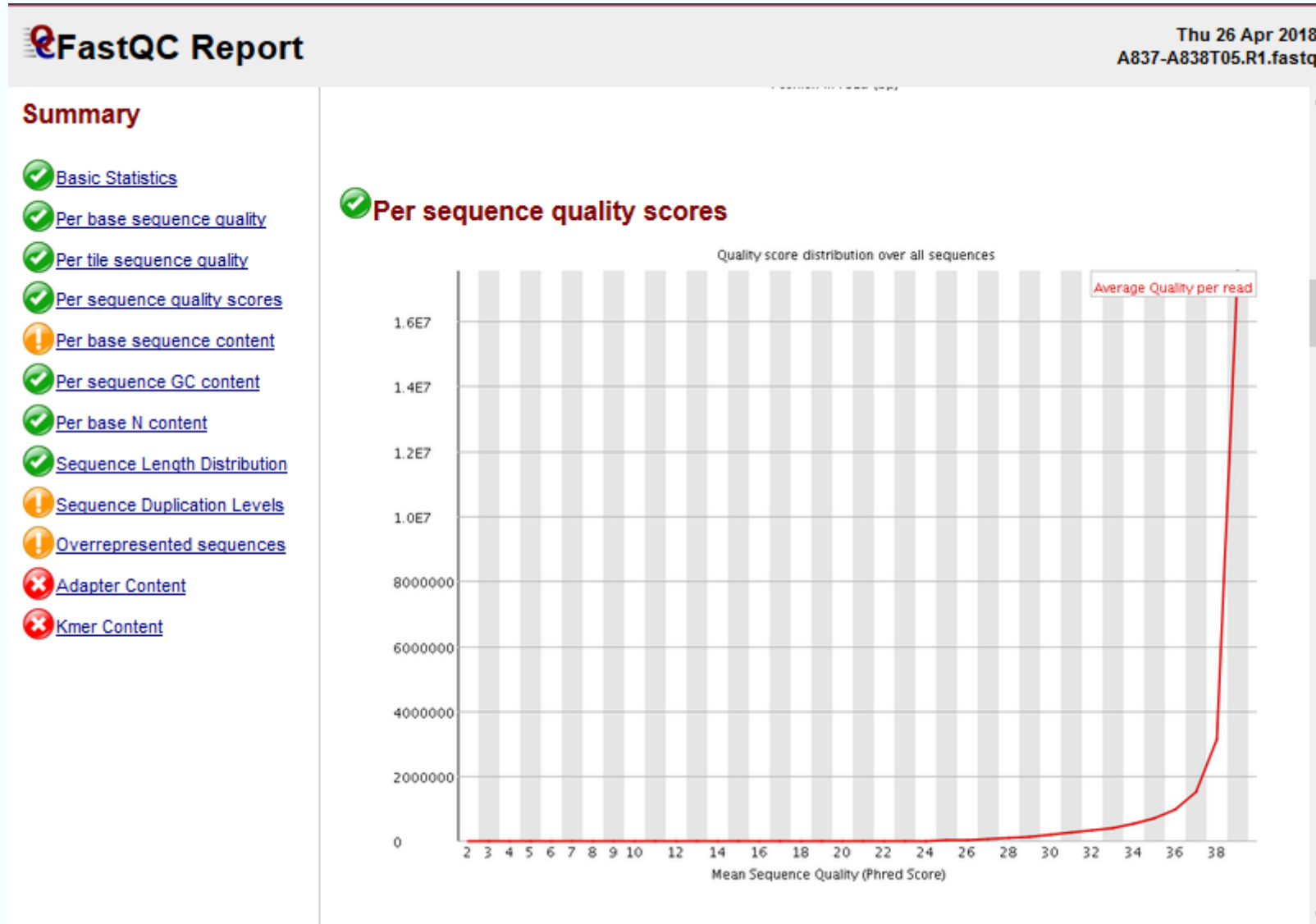
Bleu : moyenne

Boite à moustaches :

-moustaches : 10%, 90%

-**Jaune** : 25%, 75%

Données brutes - Contrôle qualité des reads - 9



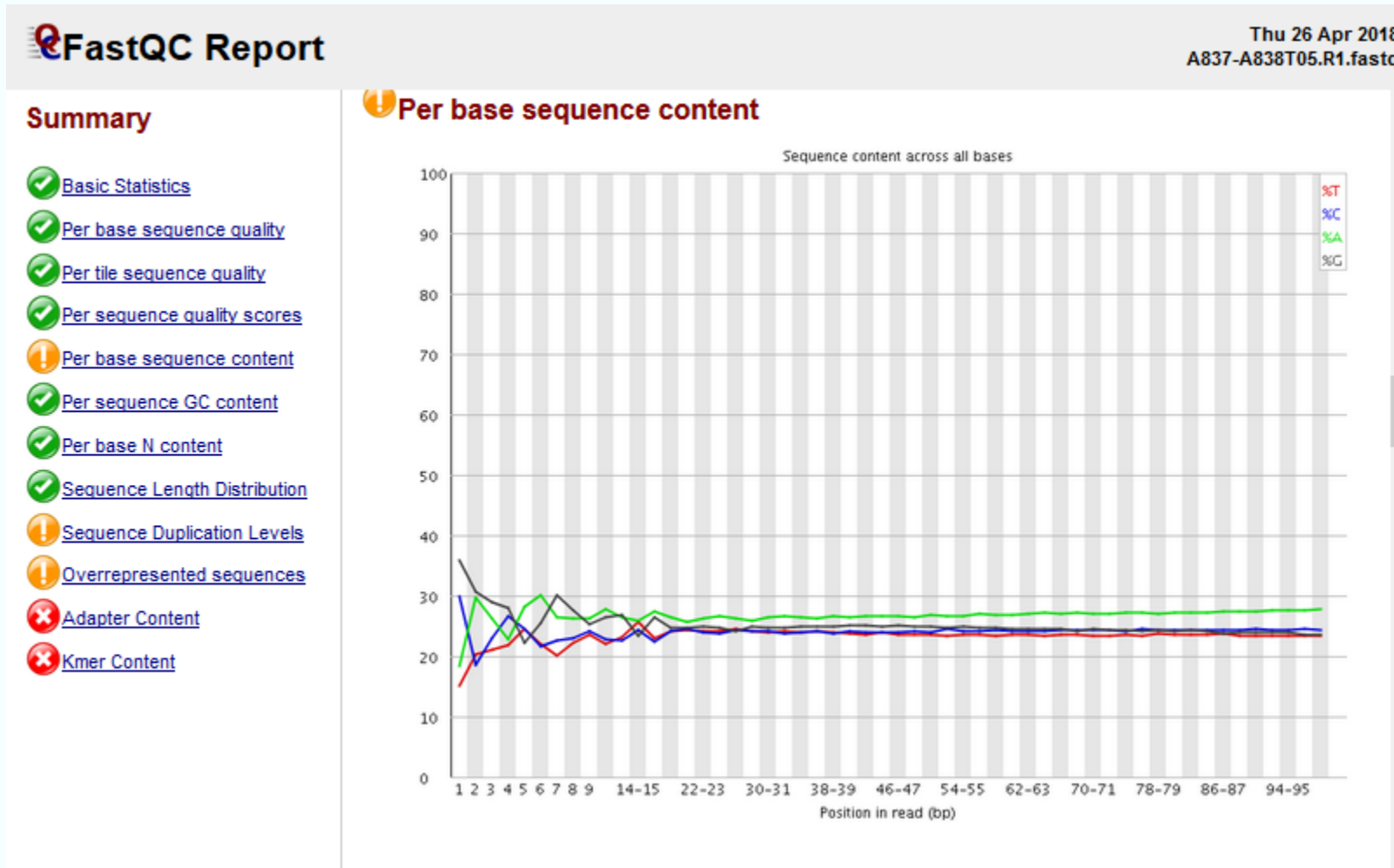
Score Phred >30

Soit $p < 10^{(-\text{Phred}/10)} = 0.001$

La probabilité que la base ait été identifiée incorrectement est inférieure à 1 sur 1000.

=> Précision de l'identification de la base = 99.9%

Données brutes - Contrôle qualité des reads - 10




On s'attend à avoir une répartition similaire des différentes bases soit **environ 25%**. Toute séquence sur-représentée peut être une contamination (overrepresented sequences)

Cas des premières **12 bases** = caractéristique des séquences Illumina : du à la méthode d'amorces randomisées (random priming process) qui n'est pas totalement aléatoire.



03

Alignement des données brutes,
qualité des alignements,
visualisation IGV



Alignement RNA-seq sur le génome Hg38.R90 - 1

Utilisation de STAR

- **Documentation**

<http://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/STAR.posix/doc/STARmanual.pdf>

- **Utilisation**

star [-o output dir] [-f fastq|bam|sam] seqfile1... seqfileN

--runMode alignReads option to align to the genome

--runThreadN nombre de threads à utiliser

--genomeDir chemin du répertoire où sont stockés les fichiers d'indexation du génome de référence

--readFilesIn nom des fichiers contenant les reads à mapper (format fastq ou fasta)

--outFileNamePrefix préfixe des fichiers de sortie de résultats

--outSAMtype BAM SortedByCoordinate résultats au format bam triés par coordonnées génomiques

--quantMode TranscriptomeSAM GeneCounts résultats au format bam triés par coordonnées des transcrits

Alignement RNA-seq sur le génome Hg38.R90 - 2

Utilisation de STAR

• Options avancées

--outFilterMultimapNmax: nombre maximal d'alignements multiples (sur différent loci) reporté pour un read. Si supérieur à la valeur alors le read est considéré non mappé (défaut = 20)

--outFilterMismatchNmax : nombre maximal de mismatches accepté par paire (défaut=999)

--alignIntronMin, --alignIntronMax, --alignMatesGapMax : taille min et max d'un intron, et taille max de l'espace non mappé entre les paired end reads

--outFilterIntronMotifs :

défaut=none : aucun filtrage

RemoveNoncanonical : supprime les alignements qui contiennent des jonctions non canoniques

RemoveNoncanonicalUnannotated : supprime les alignements qui contiennent des jonctions non canoniques non annotés quand on utilise une base de données annotées des jonctions de splice. Les jonctions non canoniques annotées seront conservées

Alignement RNA-seq sur le génome Hg38.R90 - 3

A vous de jouer

- ouvrir le script **star_Hg38_R90.sh**, le modifier en indiquant votre répertoire personnel sous */home/genouest/tp_gnf_rnaseq_40965/tp600XX* comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

```
vsibut@genossh:/groups/inserm_u1236/Module_6/TP_RNA-seq
#!/bin/bash

# initialisation environnement
. /local/env/envstar.sh

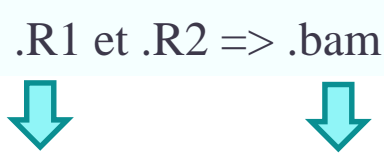
# command line

date

STAR
  --runMode alignReads
  --runThreadN 8
  --genomeDir /groups/tp40734/TP_RNAseq/Annottaions/Hg38_R90
  --readFilesIn /groups/tp40734/TP_RNAseq/1_Brut/A837-A838T05.R1.fastq /groups/tp40734/TP_RNAseq/1_Brut/A837-A838T05.R2.fastq
  --outFilterMultimapNmax 1
  --outFilterMismatchNmax 2
  --outFilterIntronMotifs RemoveNoncanonicalUnannotated
  --outFileNamePrefix /groups/tp40734/Nom_Prenom/3_Align/START05_R90_
  --outSAMtype BAM SortedByCoordinate
  --quantMode TranscriptomeSAM GeneCounts

date
~
~
~
```

.R1 et .R2 => .bam



Alignement RNA-seq sur le génome Hg38.R90 - 4

Résultat de **star_Hg38_R90.sh** pour A837-A838T05.R1_fastq et A837-A838T05.R2_fastq

```
drwxr-xr-x 1 vsibut inserm_u1236 452 Oct  5 23:17 STAR
drwxr-xr-x 1 vsibut inserm_u1236  0 Oct  5 15:46 Tophat2
[vsibut@genossh:3_Align] $ cd STAR/
[vsibut@genossh:STAR] $ ll
total 4818232
-rwxr-xr-x 1 vsibut inserm_u1236 2881744630 Oct  5 23:16 T05_R90_Aligned.sortedByCoord.out.bam
-rwxr-xr-x 1 vsibut inserm_u1236  3012040 Oct  5 23:16 T05_R90_Aligned.sortedByCoord.out.bam.bai
-rwxr-xr-x 1 vsibut inserm_u1236 2040443275 Oct  5 23:16 T05_R90_Aligned.toTranscriptome.out.bam
-rwxr-xr-x 1 vsibut inserm_u1236  1860 Oct  5 23:16 T05_R90_Log.final.out
-rwxr-xr-x 1 vsibut inserm_u1236  24159 Oct  5 23:16 T05_R90_Log.out
-rwxr-xr-x 1 vsibut inserm_u1236  836 Oct  5 23:16 T05_R90_Log.progress.out
-rwxr-xr-x 1 vsibut inserm_u1236 1351373 Oct  5 23:16 T05_R90_ReadsPerGene.out.tab
-rwxr-xr-x 1 vsibut inserm_u1236 7271459 Oct  5 23:16 T05_R90_SJ.out.tab
drwx----- 1 vsibut inserm_u1236  80 Oct  5 23:17 T05_R90__STARtmp
[vsibut@genossh:STAR] $
```

```
[drossill@genossh:drossill] $ more star_Hg38_R90.sh.o5866459
Thu Apr 26 15:25:07 CEST 2018
Apr 26 15:25:07 ..... started STAR run
Apr 26 15:25:07 ..... loading genome
Apr 26 15:29:45 ..... started mapping
Apr 26 15:36:18 ..... started sorting BAM
Apr 26 15:37:52 ..... finished successfully
Thu Apr 26 15:37:53 CEST 2018
[drossill@genossh:drossill] $
```

star_Hg38_R90.out

Contrôle qualité de reads alignés - 1

Questions

- Est-ce que la plupart des reads ont été alignés sur le génome de référence ?
- Que contient le fichier BAM?
- Peut-on visualiser les alignements sur le génome de manière interactive ?
- Y a-t-il des biais potentiels du RNA-seq ?
- Est-ce que les réplicats sont similaires ?

QC – statistiques d'alignement

T05_R90_Log.final.out

Est-ce que la plupart des reads ont été alignés sur le génome de référence ?

On considère un alignement RNAseq réussi si le taux d'alignements > 70%

Ici STAR : ***_Log.final.out** indique que

- ✓ 75.3 % des reads sont alignés de manière unique
- ✓ 13.96% sont unmapped
- ✓ 10.74% sont alignés sur plusieurs loci

```
[vsibut@genossh:STAR] $ more T05_R90_Log.final.out
      Started job on | Feb 21 11:47:55
      Started mapping on | Feb 21 11:49:32
      Finished on | Feb 21 11:54:42
Mapping speed, Million of reads per hour | 305.83

      Number of input reads | 26335159
      Average input read length | 199
      UNIQUE READS:
      Uniquely mapped reads number | 19831213
      Uniquely mapped reads % | 75.30%
      Average mapped length | 194.95
      Number of splices: Total | 6959064
Number of splices: Annotated (sjdb) | 6803403
      Number of splices: GT/AG | 6894579
      Number of splices: GC/AG | 57469
      Number of splices: AT/AC | 4786
      Number of splices: Non-canonical | 2230
      Mismatch rate per base, % | 0.28%
      Deletion rate per base | 0.02%
      Deletion average length | 1.55
      Insertion rate per base | 0.02%
      Insertion average length | 1.32
      MULTI-MAPPING READS:
      Number of reads mapped to multiple loci | 0
      % of reads mapped to multiple loci | 0.00%
      Number of reads mapped to too many loci | 2829295
      % of reads mapped to too many loci | 10.74%
      UNMAPPED READS:
      % of reads unmapped: too many mismatches | 0.00%
      % of reads unmapped: too short | 13.91%
      % of reads unmapped: other | 0.05%
      CHIMERIC READS:
      Number of chimeric reads | 0
      % of chimeric reads | 0.00%
Démarrer [vsibut@genossh:STAR] $
```

QC – Visualisation du contenu du fichier BAM - 1

Que contient le fichier BAM?

Le fichier de résultat **T05_R90_Aligned.sortedByCoord.out.bam** n'est pas lisible avec la commande **more**.

A vous de jouer

- ouvrir le script **samtools_view.sh**, le modifier avec votre e-mail et en indiquant votre répertoire personnel sous /home/genouest/tp_gnf_rnaseq_40965/tp600XX comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

```
vsibut@genossh:/groups/inserm_u1236/Module_6/TP_RNA-seq  
#!/bin/bash
```

```
# initialisation environnement  
./local/env/envsamtools-1.3.sh
```

```
# visualisation des premières lignes du .bam
```

```
samtools view /groups/tp40734/Nom_Prénom/3_Align/STAR/T05_R90_Aligned.sortedByCoord.out.bam | head
```

```
# compter le nombre d'alignements générés par STAR RNA-seq aligner
```

```
samtools view -c /groups/tp40734/Nom_Prénom/3_Align/STAR/T05_R90_Aligned.sortedByCoord.out.bam
```

samtools_view.out

Visualisation des premières lignes du .bam

```
[drossill@genossh:Formation] $ more samtools.sh.o5866957  
HISEQ:~837:H7H3HBCX2:1:1209:19366:8495    163      1          12006   255       91M10S =         182521  170610 GCAGGTGCTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCC  
ATGCCTCCCAGATCGG DDDDDHHIIIIIHHIIIIIHHIEHHIIIIIHIIGGHCHHHIIIIIHFIIIIIHHIIIIIHHIIIIIHHII IIIIIEHII NH:i:1 HI:i:1 AS:i:182 nM:i:0  
HISEQ:~838:H7GFJBCX2:1:1101:12209:61723     99        1      13380    255       98M =         183837  170553 CTCCTGCCTTTTCTTCTTCCCTTAGACCTCCACCACCCCGAGATCACATTCTCTACTGCCTTTTTGTCTGCCCCAGTTTCTACTAGAAG  
TAGGCCTCATCT BDDCDHHIHIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHGFFEEE@EHCHCGHHIIHHICHIBGE@GHHGHGCElFhl<CHHHHHIHI NH:I:l HI:I:l AS:I:186 nM:I:2  
HISEQ:~837:H7H3HBCX2:1:2103:14152:6318      163        1      13381    255       41M4t48M8S =         183900  170605 TCCTGCCTTTTCTTCTTCCCTTAGAGCCTCCACCACCCCGAGAGAGATCACATTCTCTACTGCCTTTTGTCTGCCCCAGT  
TTCACCAGAAGTAGGCAGATCGGA DDCDHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHNH NH:I:l HI:I:l AS:I:151 nM:I:0  
HISEQ:~838:H7GFJBCX2:2:1211:2392:67499      163        1      14400    255       12S87M2S =         184916  170576 TTTTTTTTTTTTTTGGTTTCTGCTCAGTTCTTTATTGATTGGTGTGCCGTTTTCTCTGGAAGCCTCTTAAGAACACA  
GTGGCGCAGGCTGGGTGGAGCCCA DDDDHHIIIIIHDCHIHCf?ElF?EHE@HHHlc@llDclGl<l<@FDHHlHlCeL<lD@lFlClcGlFl1<ll<<cDc?EDChClFG=<DDL<FEll NH:I:l HI:I:l AS:I:143 nM:I:0  
HISEQ:~837:H7H3HBCX2:2:1116:9868:52621      163        1      14410    255       2S99M =         14561   249 G CCTCAGTCTTTTATTGATTGGTGTGCCGTTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCC  
CCATGGAGCAGCAGCA BBBBCHIIIIHHIIIIIHHHHlHHHHFHlHHHHCGHHHHIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHNH NH:I:l HI:I:l AS:I:195 nM:I:0  
HISEQ:~837:H7H3HBCX2:2:2104:6967:17506      163        1      14440    255       101M =         14536   194 TCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCC  
CCAGCTGTGTGGCCTC DDDDDHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHNH NH:I:l HI:I:l AS:I:197 nM:I:0  
HISEQ:~837:H7H3HBCX2:2:2116:15328:46570      163        1      14443    255       101M =         185260  171055 CTGGTAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAAAAGTCCCCGCCCA  
GCTGTGTGGCCTCAAG DDDDHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHNH NH:I:l HI:I:l AS:I:193 nM:I:2  
HISEQ:~837:H7H3HBCX2:1:2101:11442:54432      163        1      14448    255       5S96M =         185260  171050 CTGTAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAAAAGTCCCCGCCCA  
GCTGTGTGGCCTCAAG DDDBBHGHHHHHIIIIIHHGHGHHIHHIIIIIHHIGIGIGHIIHHIIIIIHHGHIIHHIIIGGHFH HHIIIGIIIIIHHGHIIIGIIIIIHHHH=EHGHH? NH:I:l HI:I:l AS:I:188 nM:I:2  
HISEQ:~838:H7GFJBCX2:1:2208:18332:51256     163        1      14456    255       32S69M =         14672   297 GTGTGCCATTTTCTCTGGAAGCCTCTTTAGAGAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCA  
GACAGAAgtccccacc @DDDDHFEHHIIIIIHIEHHIIIIIHHIG?GHHHHEHH?FGCHHHHEDHIIHHIIEHHHIIIGEBHHIIHHCCldGHHCHHHIHEH@gH@gLEGHeLDCEE NH:I:l HI:I:l AS:I:144 nM:I:2  
HISEQ:~837:H7H3HBCX2:2:2104:6967:17506      83        1      14536    255       98M =         14440 -194 GCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCCTGGTTC CGT C ACC CCCT CCCA AG GAAG TAGGTCTGAGC  
AGCTGTGCTTGGC I II IHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHHIIIIIHDDDCD NH:I:l HI:I:l AS:I:197 nM:I:0
```

39662426



Nombre d'alignements générés par STAR = 39662426

QC - Visualisation avec IGV - 1

Peut-on visualiser les alignements sur le génome de manière interactive ?
logiciel **IGV (Broad Institute)**

- **Documentation** <http://software.broadinstitute.org/software/igv/>
- IGV prend deux fichiers en entrée
 - ✓ un fichier .bam
 - ✓ un fichier d'indexation .bam.bai du fichier .bamOu
 - ✓ un fichier .sam (.bam non compressé)
 - ✓ un fichier d'indexation qui se génère automatiquement dans IGV

QC - Visualisation avec IGV - 1

A vous de jouer

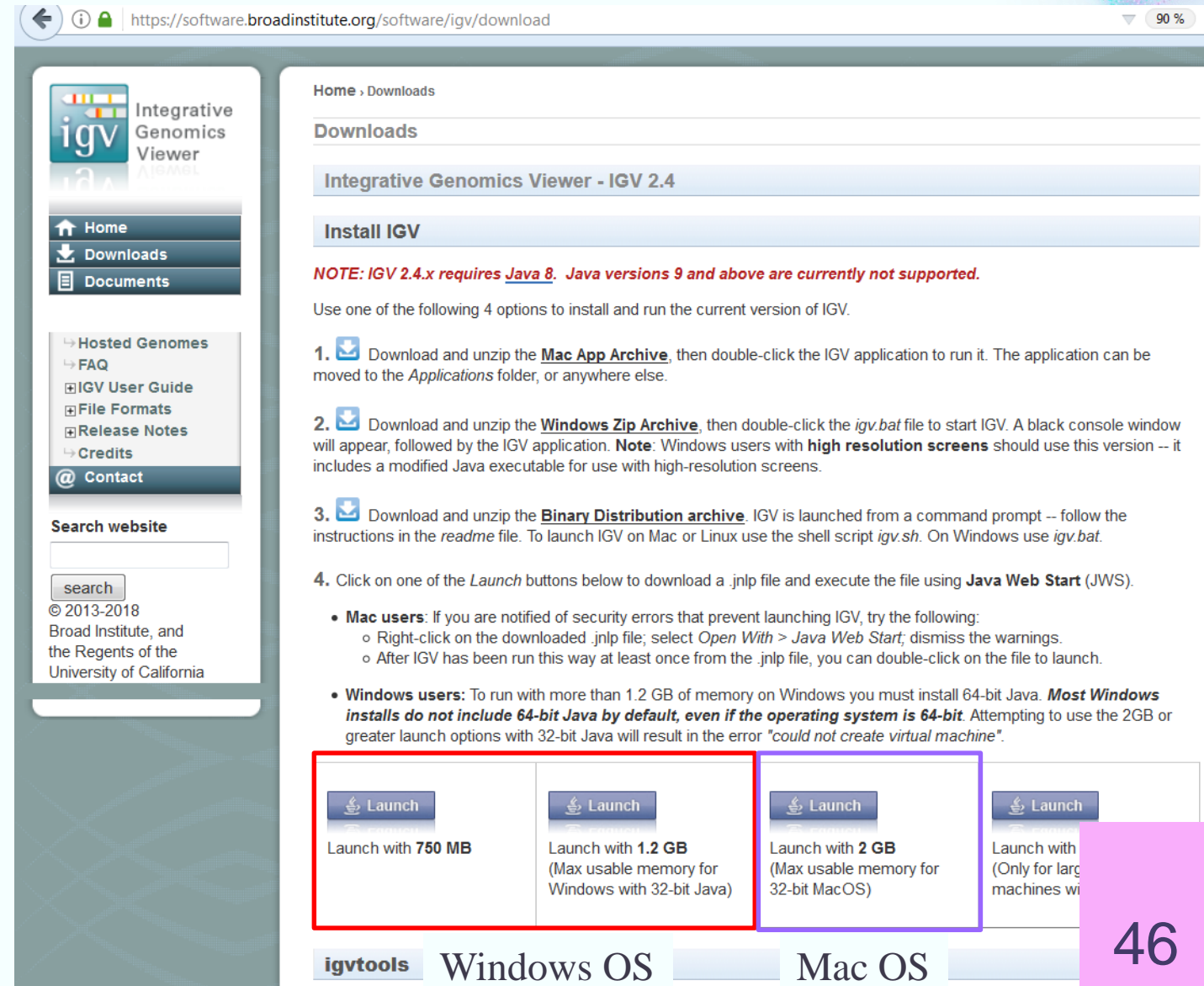
- ouvrir le script **samtools_index.sh**, le modifier avec votre e-mail et en indiquant votre répertoire personnel sous */home/genouest/tp_gnf_rnaseq_40965/tp600XX* comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

QC - Visualisation avec IGV - 2

Sur votre ordinateur local

- copier les fichiers .bam et .bam.bai (Filezilla)
- ouvrir IGV en allant sur

<http://software.broadinstitute.org/software/igv/download> .



The screenshot shows the IGV download page. The left sidebar contains navigation links: Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, Credits, and Contact. The main content area is titled 'Downloads' and features a section for 'Integrative Genomics Viewer - IGV 2.4'. It includes a note that IGV 2.4.x requires Java 8 and that Java versions 9 and above are not supported. Below this, there are four numbered steps for installation: 1. Download and unzip the Mac App Archive, 2. Download and unzip the Windows Zip Archive, 3. Download and unzip the Binary Distribution archive, and 4. Click on one of the Launch buttons. The Launch buttons are arranged in a row, each with a 'Launch' button and a description of the memory requirements. The first button is highlighted with a red box and labeled 'Windows OS' below it. The second button is highlighted with a purple box and labeled 'Mac OS' below it. The third button is labeled 'Launch with 2 GB (Max usable memory for 32-bit MacOS)'. The fourth button is labeled 'Launch with (Only for large machines wi'. At the bottom, there are labels for 'igvtools', 'Windows OS', and 'Mac OS'.

Home , Downloads

Downloads





Integrative Genomics Viewer - IGV 2.4

Install IGV

NOTE: IGV 2.4.x requires Java 8. Java versions 9 and above are currently not supported.

Use one of the following 4 options to install and run the current version of IGV.

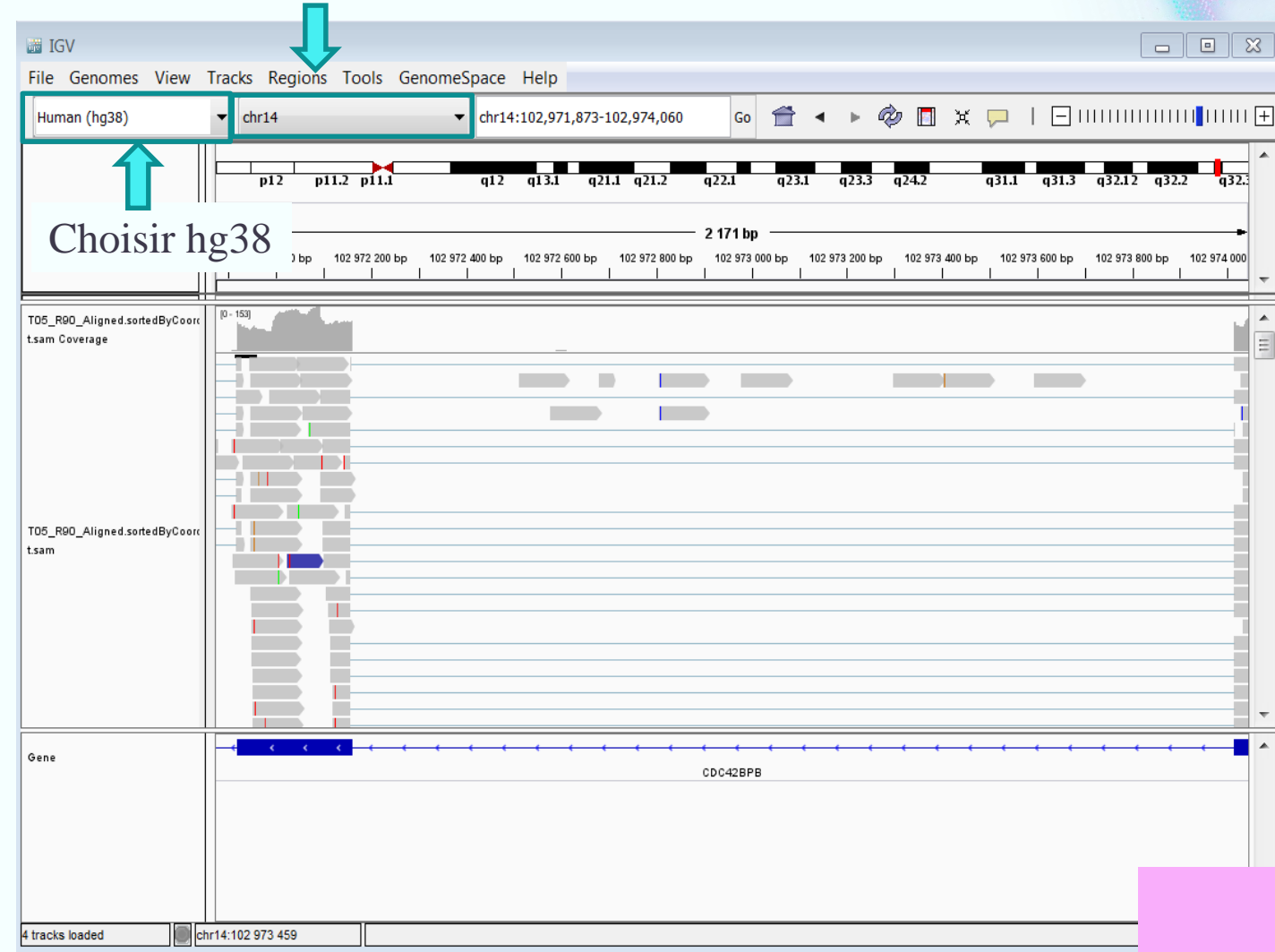
1. Download and unzip the **Mac App Archive**, then double-click the IGV application to run it. The application can be moved to the *Applications* folder, or anywhere else.
2. Download and unzip the **Windows Zip Archive**, then double-click the *igv.bat* file to start IGV. A black console window will appear, followed by the IGV application. **Note:** Windows users with **high resolution screens** should use this version -- it includes a modified Java executable for use with high-resolution screens.
3. Download and unzip the **Binary Distribution archive**. IGV is launched from a command prompt -- follow the instructions in the *readme* file. To launch IGV on Mac or Linux use the shell script *igv.sh*. On Windows use *igv.bat*.
4. Click on one of the *Launch* buttons below to download a .jnlp file and execute the file using **Java Web Start (JWS)**.
 - **Mac users:** If you are notified of security errors that prevent launching IGV, try the following:
 - Right-click on the downloaded .jnlp file; select *Open With > Java Web Start*; dismiss the warnings.
 - After IGV has been run this way at least once from the .jnlp file, you can double-click on the file to launch.
 - **Windows users:** To run with more than 1.2 GB of memory on Windows you must install 64-bit Java. **Most Windows installs do not include 64-bit Java by default, even if the operating system is 64-bit.** Attempting to use the 2GB or greater launch options with 32-bit Java will result in the error "could not create virtual machine".

 Launch with 750 MB	 Launch with 1.2 GB (Max usable memory for Windows with 32-bit Java)	 Launch with 2 GB (Max usable memory for 32-bit MacOS)	 Launch with (Only for large machines wi
--	--	--	---

igvtools Windows OS Mac OS

QC - Visualisation avec IGV - 2

Choisir chr1



QC avec RSeQC

Utilisation de RSeQC

- **Documentation** <http://rseqc.sourceforge.net/>
- **Utilisation**
 - ✓ RSeQC nécessite l'environnement Python (les fonctions sont écrites en Python) et l'environnement R (pour la génération des résultats, .pdf et autres)
 - ✓ chaque fonction .py réalise un contrôle qualité spécifique

QC avec RSeQC

A vous de jouer

- ouvrir le script **RSeQC.sh**, le modifier avec votre e-mail et en indiquant votre répertoire personnel sous */home/genouest/tp_gnf_rnaseq_40965/tp600XX* comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

QC avec RSeQC – bam_stat.py - 1

`bam_stat.py`

Objectif

Statistiques sur le nombre de reads mappés

Input

Input BAM/SAM file Alignment file in BAM/SAM format.

Output

Summary

Total Reads (Total records) = {Multiple mapped reads} + {Uniquely mapped}

Uniquely mapped Reads = {read1} + {read2} (if paired end)

Uniquely mapped Reads = {Reads map to '+'} + {Reads map to '-'}

Uniquely mapped Reads = {Splice reads} + {Non-splice reads}

QC avec RSeQC – bam_stat.py - 2

A vous de jouer :

Comparer les statistiques RSeQC avec celles de STAR

```
#####  
#All numbers are READ count  
#####  
Total records:                39662426  
  
QC failed:                    0  
Optical/PCR duplicate:        0  
Non primary hits              0  
Unmapped reads:              0  
mapq < mapq_cut (non-unique): 0  
  
mapq >= mapq_cut (unique):    39662426  
Read-1:                      19831213  
Read-2:                      19831213  
Reads map to '+':            19831213  
Reads map to '-':            19831213  
Non-splice reads:            33209958  
Splice reads:                6452468  
Reads mapped in proper pairs: 39662426  
Proper-paired reads map to different chrom:0  
Wed May  2 13:07:45 CEST 2018
```

Input : T05_R90_Aligned.sortedByCoord.out.bam

← Multiple mapped reads = 0

← Uniquely mapped reads = 19831213 (x2 = 39662426 paired-end)
* Même nombre de reads mappés sur brin sens et anti-sens
* Tous les reads sont alignés par paire

Pourcentage de reads mappés =
 $(\text{\#multiple mapped reads} + \text{\#uniquely mapped reads}) / \text{total raw reads (fastqc)}$
 $= (0 + 19831213) / 26335159 = 75.30\%$
→ Mêmes résultats que ceux rendus par STAR

QC avec RSeQC – infer_experiment.py - 1

infer_experiment.py

Objectif

Inférence sur le type de séquençage utilisé (orienté, non-orienté)

Input

Input BAM/SAM file Alignment file in BAM/SAM format.

Reference gene model Gene model in BED format.

Output

Librairie paired-end : deux catégories

1++, 1--, 2+-, 2--

le read1 aligné sur le brin + indique un gène parental sur le brin +

le read1 aligné sur le brin - indique un gène parental sur le brin -

le read2 aligné sur le brin + indique un gène parental sur le brin -

le read2 aligné sur le brin - indique un gène parental sur le brin +

1+-, 1-+, 2++, 2—

QC avec RSeQC – infer_experiment.py - 2

A vous de jouer

il faut que les chromosomes soient appelés de la même manière dans les fichiers BAM et BED : soit 'chr1' soit '1'.

A vérifier en utilisant : pour le fichier .bed : head hg38.UCSC.bed
pour le fichier .bam : samtools_view pour le fichier .bam

```
This is PairEnd Data
Fraction of reads failed to determine: 0.0424
Fraction of reads explained by "1++,1--,2+-,2-+": 0.8068
Fraction of reads explained by "1+-,1-+,2++,2--": 0.1508
```

Input : T05_R90_Aligned.sortedByCoord.out.bam

- ✓ reconnaissance de la librairie paired-end
- ✓ 4.24% des reads totaux sont mappés sur des régions du génome pour lesquelles l'orientation des transcrits ne peut être déterminées (tq des régions avec les deux brins transcrits)
- ✓ 80.68% sont expliqués par 1++ 1- 2+- 2-+ et 15.08% par 1+- 1-+ 2++ 2--
⇒ la disproportion des pourcentages indique que les reads sont orientés : l'orientation des reads 'strandness of reads' est dépendante de l'orientation des transcrits 'strandness of transcripts'
⇒ le script retrouve donc bien que le séquençage est orienté.

QC avec RSeQC – inner_distance.py - 1

inner_distance.py

Objectif

distribution des distances *inner* entre paire de reads

Input

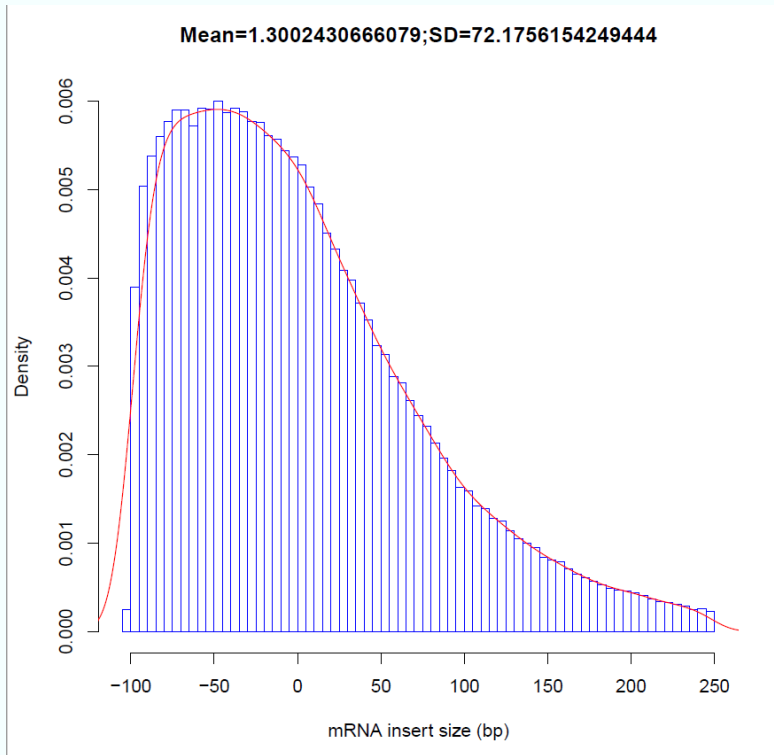
- Input BAM/SAM file Alignment file in BAM/SAM format.
- Reference gene model Gene model in BED format.

Output

- output.inner_distance.txt
 - ✓ read ID
 - ✓ inner distance
 - ✓ how paired reads were mapped: PE_reads_overlap, dist=genomic...
- output..inner_distance_freq.txt
 - ✓ inner distance starts
 - ✓ inner distance ends
 - ✓ number of read pairs
- output.inner_distance_plot.r R script to generate histogram
- output.inner_distance_plot.pdf histogram plot

QC avec RSeQC – inner_distance.py - 2

Quelle est la distribution des distances *inner* entre paire de reads? inner_distance.py



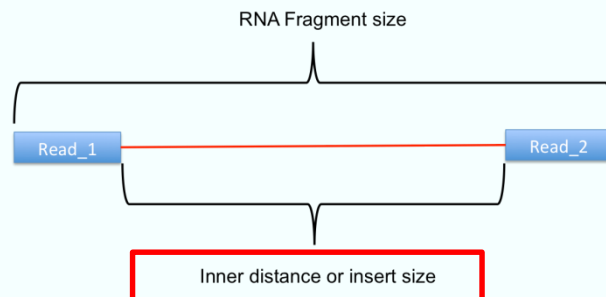
Input : T05_R90_Aligned.sortedByCoord.out.bam

- Séquenceur Illumina
 - ✓ librairie paired-end
 - ✓ taille des reads = 150bp

➔ Distance *inner* est négative

Mean = 1.30

Médiane = - 13



QC avec RSeQC – épissages alternatifs - 1

junction_annotation.py

Objectif

La profondeur de séquençage est-elle suffisante pour étudier les épissages alternatifs ?

Input

- Input BAM/SAM file Alignment file in BAM/SAM format.
- Reference gene model Gene model in BED format.

Output

- output.junc.anno.junction.xls
 - ✓ chrom ID
 - ✓ start position of junction
 - ✓ end position of junction
 - ✓ number of splice events supporting this junction
 - ✓ 'annotated', 'complete_novel' or 'partial_novel'
- output.anno.junction_plot.r R script to generate pie chart
- output.splice_junction.pdf plot of splice junctions
- output.splice_events.pdf plot of splice events

QC avec RSeQC – épissages alternatifs - 2

`junction_saturation.py`

Objectif

Annotation des jonctions d'épissage alternatif

Input

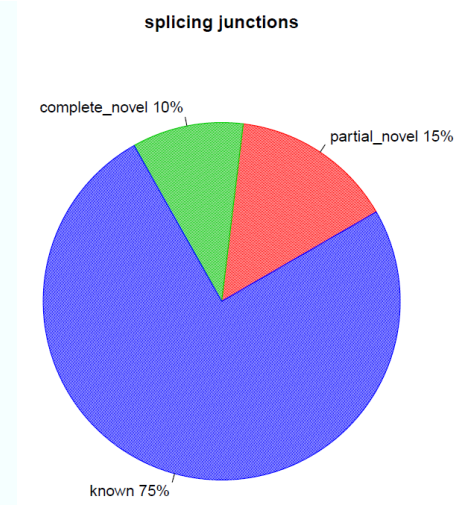
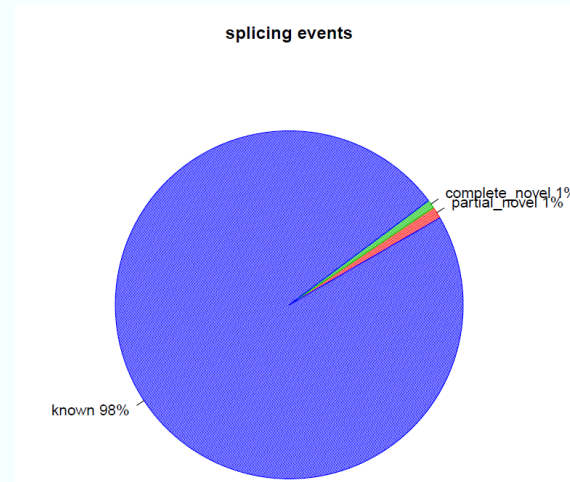
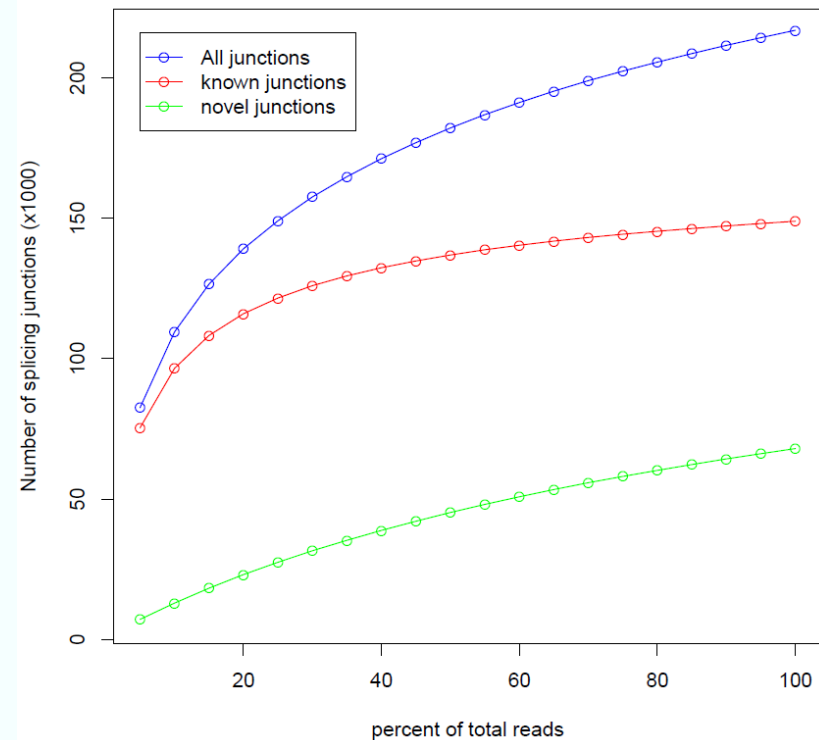
- Input BAM/SAM file Alignment file in BAM/SAM format.
- Reference gene model Gene model in BED format.

Output

- `output.junctionSaturation_plot.r` R script to generate plot
- `output.junctionSaturation_plot.pdf`

QC avec RSeQC – épissages alternatifs - 3

Epissages alternatifs : `junction_annotation.py`, `junction_saturation.py`



Input :

T05_R90_Aligned.sortedByCoord.out.bam

- La majorité des épissages alternatifs sont connus :

- ✓ 98% known splicing events
- ✓ 75% known splicing junctions

- Pas tout à fait à saturation des jonctions connues : on pourrait vouloir augmenter la profondeur du séquençage.

QC avec RSeQC – read_distribution.py - 1

read_distribution.py

Objectif

Distribution du nombre de reads mappés selon le type génomique (CDS exon, 5'UTR exon, 3' UTR exon, Intron, Intergenic regions).

Input

- Input BAM/SAM file Alignment file in BAM/SAM format.
- Reference gene model Gene model in BED format.

Output

- Summary
 - ✓ Total Reads
 - ✓ Total Tags: reads spliced once will be counted as 2 tags, reads spliced twice will be counted as 3 tags, etc. And because of this, "Total Tags" \geq "Total Reads"
 - ✓ Total Assigned Tags: number of tags that can be unambiguously assigned to the 10 groups

QC avec RSeQC – read_distribution.py - 23

Quel est la répartition des reads alignés selon le type génomique? `read_distribution.py`

Total Reads	39662426		
Total Tags	47589651		
Total Assigned Tags	44794191		
=====			
Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	38564553	20501699	531.62
5'UTR_Exons	32001905	1555427	48.60
3'UTR_Exons	59522509	6160086	103.49
Introns	1476696484	15416084	10.44
TSS_up_1kb	25113591	69717	2.78
TSS_up_5kb	113529464	222217	1.96
TSS_up_10kb	207732729	362164	1.74
TES_down_1kb	27059805	146575	5.42
TES_down_5kb	117943776	529394	4.49
TES_down_10kb	211293094	798731	3.78

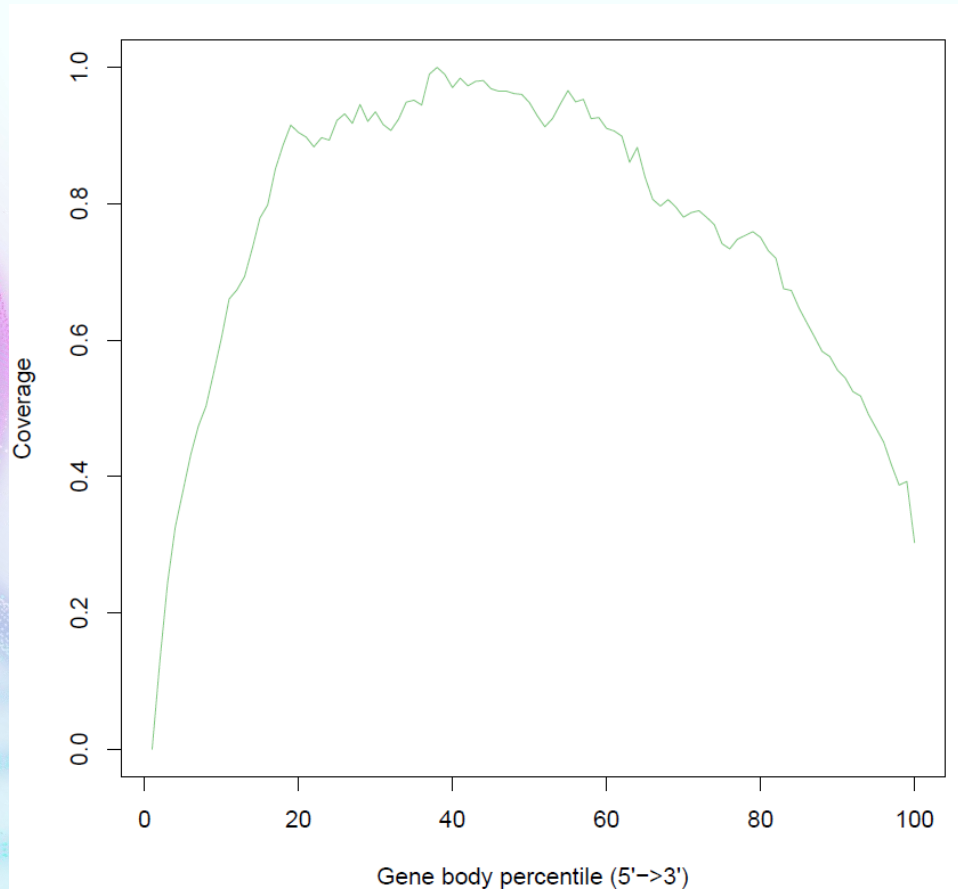
Input :

T05_R90_Aligned.sortedByCoord.out.bam

- CDS exons = 20 501 699 tags
- introns = 15 416 084 tags

QC avec RSeQC – geneBody_coverage.py - 1

Évaluation des biais de couverture des reads à 5' et 3' : `geneBody_coverage.py`



A vous de jouer

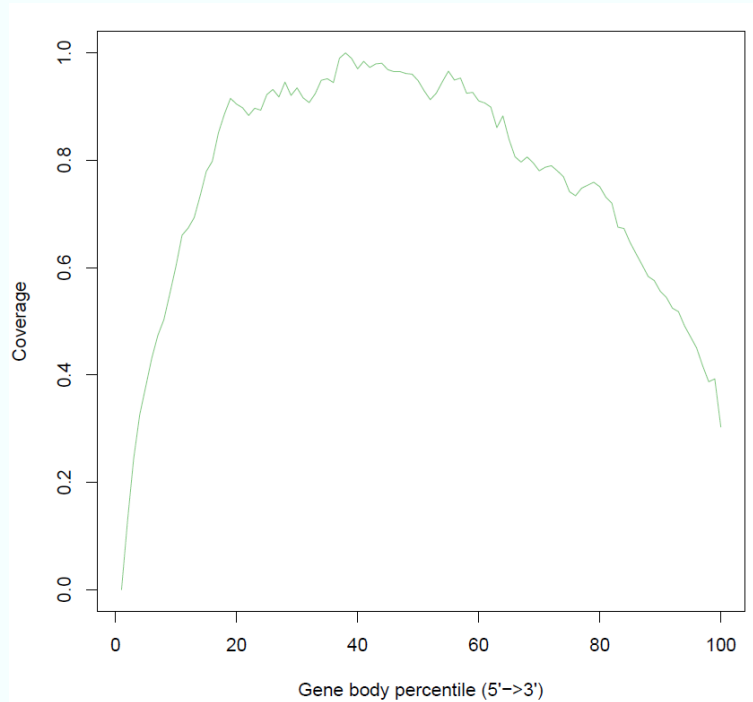
- modifier le fichier
`/home/genouest/tp_gnf_rnaseq_40965/tp600XX/RSeQC_geneBody_1bam_top10000bed.sh.`
- soumettre le script au cluster de calcul, vérifier le status du job et les fichiers générés en cours d'exécution

Input

- T05_R90_Aligned.sortedByCoord.out.bam
- hg38.UCSC.top10000.bed : top 10000 transcrits de hg38.UCSC.bed
- Fichier d'indexation .bai généré à partir du fichier .bam -> utilisation de `samtools_index.sh` (avant le lancement de `RSeQC_geneBody_1bam_top10000bed.sh`)

QC avec RSeQC – geneBody_coverage.py - 2

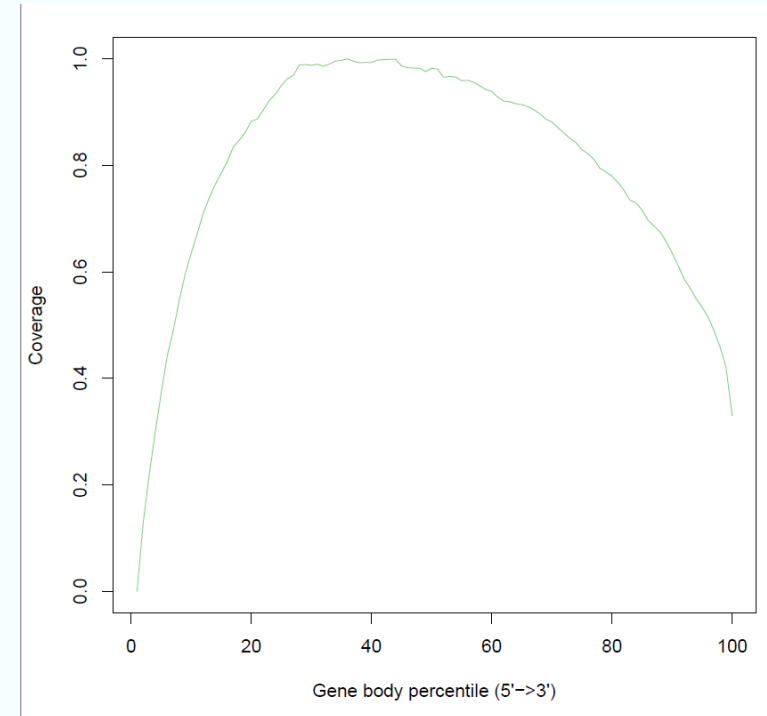
Évaluation des biais de couverture des reads à 5' et 3' : `geneBody_coverage.py`



Input

T05_R90_Aligned.sortedByCoord.out.bam
hg38.UCSC.top10000.bed

Temps d'exécution = 15 min



Input

T05_R90_Aligned.sortedByCoord.out.bam
hg38.UCSC.bed

Temps d'exécution = 17 heures



04

Comptage des reads mappés

Comptage des reads mappés - 1

Utilisation de `featureCounts` {subread package}

➔ `featureCounts` compte les reads mappés par type génomique : gènes, exons, promoteurs, localisations chromosomiques...

- Documentation <http://subread.sourceforge.net/>
- Utilisation : to summarize a BAM format dataset



`featureCounts [-t exon] [-g gene_id|transcript_id] [-a annotation file] [-o output files] bamfile`

`-t <exon>` feature type
`-g <gene_id|transcript_id>` meta-feature type : the aggregation of a set of features
`-a </path/to/annotation/GTF/file>` annotation file
`-o <output file>` name of the output file

Comptage des reads mappés - 2

A vous de jouer

- ouvrir le script **featureCount.sh**, le modifier avec votre e-mail et en indiquant votre répertoire personnel sous */home/genouest/tp_gnf_rnaseq_40965/tp600XX* comme répertoire où seront sauvés les résultats.
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

Comptage des reads mappés - 3

```
vsibut@genossh:/groups/inserm_u1236/Module_6/TP_RNA-seq
#!/bin/bash

# initiation environnement
. /local/env/envpython-2.7.sh
. /local/env/envR-3.2.3.sh
. /local/env/envsubread-1.4.6.sh

#####
# comptage des reads alignes avant analyse differentielle
#####

date

echo 'Subread featureCounts - Raw counts of fragments'

# variables
InIndDeb=/groups/tp40734/Nom_Prénom/3_Align/STAR/T # chemin où se trouve les fichiers .bam à analyser
InIndFin=( _Hgxx_Rxx_Aligned.sortedByCoord.out.bam ) # extension commune des fichiers .bam à analyser
OutIndDeb=/groups/tp40734/Nom_Prénom/5_FeatureCounts/T
OutfeatureFin=( _countGene_Hgxx_Rxx.txt ) # extension des fichiers résultats de count sur les gènes
OutfeatureFinTranscript=( _countTranscript_Hgxx_Rxx.txt ) # extension des fichiers résultats de count sur les transcripts

for i in {05};
do
    echo $individuDeb$i$individuFin

    # créer l'index des .bam
    ### summarize paired-end reads and count fragments (instead of reads) feature : Exon --> meta-feature : Gene
    featureCounts -t exon -g gene_id -p -s 0 -a /groups/tp40734/TP_RNAseq/Annotations/Homo_sapiens.GRChxx.xx.chr.gtf -o $OutIndDeb$i$OutfeatureFin $InIndDeb$i$InIndFin

    ### summarize paired-end reads and count fragments feature : Exon --> meta-feature : Transcript
    featureCounts -t exon -g transcript_id -p -s 0 -a /groups/tp40734/TP_RNAseq/Annotations/Homo_sapiens.GRChxx.xx.chr.gtf -o $OutIndDeb$i$OutfeatureFinTranscript $InIndDeb$i$InIndFin
done

date
```

Comptage des reads mappés - 4

Résultat de **featureCount.sh** pour les deux fichiers BAM T05_R90 et T06_R90

```
[vsibut@genossh:TP_RNA-seq] $ cd 5_FeatureCounts/
[vsibut@genossh:5_FeatureCounts] $ ll
total 60348
-rw----- 1 vsibut inserm_u1236 28766687 Oct  5 23:40 T05_countGene_Hg38_R90.txt
-rw----- 1 vsibut inserm_u1236      396 Oct  5 23:38 T05_countGene_Hg38_R90.txt.summary
-rw----- 1 vsibut inserm_u1236 31935040 Oct  5 23:44 T05_countTranscript_Hg38_R90.txt
-rw----- 1 vsibut inserm_u1236      396 Oct  5 23:40 T05_countTranscript_Hg38_R90.txt.summary
-rw----- 1 vsibut inserm_u1236 1079163 Oct  5 23:29 T05.txt
[vsibut@genossh:5_FeatureCounts] $
```

Comptage des reads mappés - 5

featureCount.out

```
[drossill@genossh:drossill] $ more featureCount.sh.o5868709
Thu May  3 14:56:41 CEST 2018
Subread featureCounts - Raw counts of fragments
/groups/inserm_u1236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam
exon to gene
Thu May  3 15:02:10 CEST 2018
exon to transcript
/groups/inserm_u1236/Module_6/2_Align_Hg38_R90/T06_R90_Aligned.sortedByCoord.out.bam
exon to gene
Thu May  3 15:13:34 CEST 2018
exon to transcript
Thu May  3 15:19:31 CEST 2018
[drossill@genossh:drossill] $
```

```
[drossill@genossh:drossill] $ more featureCount.sh.e5868709

=====
=====
=====
=====
=====
=====
v1.4.6

SUBREAD

//===== featureCounts setting =====\\
||
||      Input files : 1 BAM file
||                      P /groups/inserm_u1236/Module_6/2_Align_Hg38 ...
||
||      Output file : /groups/inserm_u1236/Module_6/Formation/dros ...
||      Annotations : /groups/inserm_u1236/Commun/Annotations/Homo ...
||
||      Threads : 1
||      Level : meta-feature level
||      Paired-end : yes
||      Strand specific : no
||      Multimapping reads : not counted
||      Multi-overlapping reads : not counted
||
||      Chimeric reads : counted
||      Both ends mapped : not required
||
\\===== http://subread.sourceforge.net/ =====\\

//===== Running =====\\
||
|| Load annotation file /groups/inserm_u1236/Commun/Annotations/Homo_sapi ...
||      Features : 1199596
||      Meta-features : 58243
||      Chromosomes : 25
||
|| Process BAM file /groups/inserm_u1236/Module_6/2_Align_Hg38_R90/T05_R9 ...
||      Paired-end reads are included.
||      Assign fragments (read pairs) to features...
||      Found reads that are not properly paired.
||      (missing mate or the mate is not the next read)
||      0 read has missing mates.
||      Input was converted to a format accepted by featureCounts.
||      Total fragments : 19831213
||      Successfully assigned fragments : 10926934 (55.1%)
||      Running time : 5.22 minutes
||
||
||      Read assignment finished.
```


Comptage des reads mappés - 6

T05_countGene_Hg38_R90.txt.summary

```
Status /groups/inserm_ul236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam
Assigned 10926934
Unassigned_Ambiguity 613555
Unassigned_MultiMapping 0
Unassigned_NoFeatures 8290724
Unassigned_Unmapped 0
Unassigned_MappingQuality 0
Unassigned_FragmentLength 0
Unassigned_Chimera 0
Unassigned_Secondary 0
Unassigned_Nonjunction 0
Unassigned_Duplicate 0
```

Assigned : nber of reads assigned to a gene

Unassigned_Ambiguity : overlapping with ≥ 2 features or meta-features

Unassigned_NoFeatures : not overlapping with any features included in the annotation

T05_countGene_Hg38_R90.txt

```
[vsibut@genossh:5_FeatureCounts] $ head T05_countGene_Hg38_R90.txt
# Program:featureCounts v1.6.0; Command:"featureCounts" "-t" "exon" "-g" "gene_id" "-p" "-s" "0" "-a" "/groups/inserm_ul236/Commun/Annotations/Homo_sapiens.GRCh38.90.chr.gtf" "-o" "/groups/inserm_ul236/Module_6/Formation/drossill/4_featureCounts/T05_countGene_Hg38_R90.txt" "/groups/inserm_ul236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam"
Geneid Chr Start End Strand Length /groups/inserm_ul236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam
ENSG00000223972 1;1;1;1;1;1;1;1 11869;12010;12179;12613;12613;12975;13221;13221;13453 12227;12057;12227;12721;12697;13052;13374;14409;13670 +;+;+;+;+;+;+;+ 1735 1
ENSG00000227232 1;1;1;1;1;1;1;1 14404;15005;15796;16607;16858;17233;17606;17915;18268;24738;29534 14501;15038;15947;16765;17055;17368;17742;18061;18366;24891;29570 -;-;-;-;-;-;-
;-;-;-;-;-;-;- 1351 28
ENSG00000278267 1 17369 17436 - 68 0
ENSG00000243485 1;1;1;1;1 29554;30267;30564;30976;30976 30039;30667;30667;31109;31097 +;+;+;+;+ 1021 1
ENSG00000284332 1 30366 30503 + 138 0
ENSG00000237613 1;1;1;1;1 34554;35245;35277;35721;35721 35174;35481;35481;36073;36081 -;-;-;-;-;- 1219 0
ENSG00000268020 1 52473 53312 + 840 0
ENSG00000240361 1;1;1;1 57598;58700;62916;62949 57653;58856;64116;63887 +;+;+;+ 1414 0
[vsibut@genossh:5_FeatureCounts] $
```

Col1 Geneid
Col7 counts



Comptage des reads mappés - 7

A vous de jouer

- ouvrir le script **featureCount_to_DESeq2.sh**, le modifier avec votre e-mail et en indiquant votre répertoire personnel sous `/home/genouest/tp_gnf_rnaseq_40965/tp600XX` comme répertoire où seront sauvés les résultats.
- quelle est la structure du fichier d'entrée ? Du fichier résultat?
- soumettre le script au cluster de calcul, vérifier le statut du job et les fichiers générés en cours d'exécution

```
[vsibut@genossh:5_FeatureCounts] $ head T05_countGene_Hg38_R90.txt
# Program:featureCounts v1.6.0; Command:"featureCounts" "-t" "exon" "-g" "gene_id" "-p" "-s" "0" "-a" "/groups/inserm_u1236/Commun/Annotations/Homo_sapiens.GRCh38.90.chr.gtf" "-o" "/groups/inserm_u1236/Module_6/Formation/drossill/4_featureCounts/T05_countGene_Hg38_R90.txt" "/groups/inserm_u1236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam"
Geneid Chr Start End Strand Length /groups/inserm_u1236/Module_6/2_Align_Hg38_R90/T05_R90_Aligned.sortedByCoord.out.bam
ENSG00000223972 1;1;1;1;1;1;1;1 11869;12010;12179;12613;12613;12975;13221;13221;13453 12227;12057;12227;12721;12697;13052;13374;14409;13670 +;+;+;+;+;+;+ 1735 1
ENSG00000227232 1;1;1;1;1;1;1;1 14404;15005;15796;16607;16858;17233;17606;17915;18268;24738;29534 14501;15038;15947;16765;17055;17368;17742;18061;18366;24891;29570 -;-;-;-;-;-;-
;-;-;-;-;-;-;- 1351 28
ENSG00000278267 1 17369 17436 - 68 0
ENSG00000243485 1;1;1;1;1 29554;30267;30564;30976;30976 30039;30667;30667;31109;31097 +;+;+;+;+ 1021 1
ENSG00000284332 1 30366 30503 + 138 0
ENSG00000237613 1;1;1;1;1 34554;35245;35277;35721;35721 35174;35481;35481;36073;36081 -;-;-;-;-;- 1219 0
ENSG00000268020 1 52473 53312 + 840 0
ENSG00000240361 1;1;1;1 57598;58700;62916;62949 57653;58856;64116;63887 +;+;+;+ 1414 0

[vsibut@genossh:5_FeatureCounts] $ head T05.txt
ENSG00000223972 1
ENSG00000227232 28
ENSG00000278267 0
ENSG00000243485 1
ENSG00000284332 0
ENSG00000237613 0
ENSG00000268020 0
ENSG00000240361 0
ENSG00000186092 0
ENSG00000238009 2

[vsibut@genossh:5_FeatureCounts] $
```



05

Analyse différentielle avec DESeq2

Analyse différentielle avec DESeq2 - 2

Script **DESeq2_FromSampleFiles_v1.16.R** à charger pour l'utilisation de ses fonctions.

Fonction principale : *DESeq2_FromSampleFiles()*

Input

table	sample info with association sample names & Condition (fichier .txt)
dir	sample files directory (one file = raw counts of one sample)
pval	pvalue for filtering DE genes
FC	FC for filtering DE genes
analysis	analysis type = "All" "Ctrl"
condCtrl	control condition (NULL by default)
PreFilt	pre-filtering : NULL integer : keep only rows that have at least PreFilt reads total (NULL per default : no pre-filtering)
Filt	HTS Filtering = TRUE FALSE (FALSE per default)
nbCountMin	min number of normalized counts in 100% samples (NULL per default)
paired	paired analysis = TRUE FALSE (FALSE per default)
correction	multiple testing correction = "bonferroni" "fdr"
record	enregistrer les resultats DESeq2 sans filtrage par p ou FC = TRUE FALSE (FALSE per default)

Analyse différentielle avec DESeq2 - 3

- **Input : counts bruts** avec `DESeq2_FromSampleFiles_v1.16.R`
- **Format :** un fichier .txt par échantillon associés à une table d'annotation

T05..20.txt

```
ENSG00000223972 1
ENSG00000227232 28
ENSG00000278267 0
ENSG00000243485 1
ENSG00000284332 0
ENSG00000237613 0
ENSG00000268020 0
ENSG00000240361 0
```

Aucune entête au fichier !
Col1 identifiant du gène
Col2 counts bruts

Module_6_TP_table.txt - Bloc-notes		
Fichier Edition Format Affichage ?		
Sample	File	Condition
T05	T05.txt	DN
T06	T06.txt	DN
T07	T07.txt	DN
T08	T08.txt	DN
T09	T09.txt	FDC
T10	T10.txt	FDC
T11	T11.txt	FDC
T12	T12.txt	FDC
T17	T17.txt	FRC49a
T18	T18.txt	FRC49a
T19	T19.txt	FRC49a
T20	T20.txt	FRC49a

Avec entête !
Col2 noms des sampleFiles_x.txt à importer
Col3 Condition pour les tests de comparaison
+/- Col4 Appariement utilisé si les échantillons sont appariés

Analyse différentielle avec DESeq2 - 4

```
# --- répertoire de travail
setwd("d:/home/drossill/bureau/ENSEIGNEMENT/2018 Formation R/MODULE 6/Module_6_TP/5_DESeq2")
dir <- getwd()

#####
##### DESeq2 from sample files #####
#####

##### chargement des fonctions DESeq2 pour sample files
source("DESeq2_FromSampleFiles_v1.16.R")

##### creer la table des informations sur les echantillons
# a faire avec blocnotes ou excel -> .txt
# exemple: sampleinfo_ForSampleFiles.txt

##### cas : toutes les comparaisons possibles
# --- setting parameters - All

table="Module_6_TP_table.txt"
dir <- getwd()
pval <- 0.01          # seuil de filtrage sur p (si adjP actif alors p=adjp)
FC <- 2               # seuil de filtrage sur FC (FC lineaire et non logFC)
analysis = "All"
condCtrl= "DN"        # condition de reference : non prise en compte si analysis = All
PreFilt = 10          # pre-filtrage : conservation des genes ayant au min PreFilt reads au total
Filt = T              # HTSfilter actif uniquement si NbCountMin = NULL
NbCountMin <- NULL    # min nb counts normalises par gene pour un echantillon
paired <- F           # analyse non appariee
correction="bonferroni" # correction bonferroni pour adjp
record = FALSE        # pas d'enregistrement ds resultats DESeq2 sans filtrage

# --- un fichier log sera genere
log <- paste("log_Module_6_TP_", format(sys.time(), '%Y%m%d_%Hh%M'), ".txt", sep="")
sink(log, type="output")
cat(log, "\n")
cat(date(), "\n")

DESeq2_FromSampleFiles(table, dir, pval, FC, analysis=analysis, condCtrl=condCtrl,
                        PreFilt=PreFilt, Filt=Filt, NbCountMin=NbCountMin,
                        paired=paired, correction=correction, record=record)

cat(date(), "\n")
sink()
```

Module_6_TP_table.txt - Bloc-notes

Fichier	Edition	Format	Affichage ?
Sample	File	Condition	
T05	T05.txt	DN	
T06	T06.txt	DN	
T07	T07.txt	DN	
T08	T08.txt	DN	
T09	T09.txt	FDC	
T10	T10.txt	FDC	
T11	T11.txt	FDC	
T12	T12.txt	FDC	
T17	T17.txt	FRC49a	
T18	T18.txt	FRC49a	
T19	T19.txt	FRC49a	
T20	T20.txt	FRC49a	

Output

Graph_report_condition.pdf

Condition_Counts_global.xlsx

Condition_FDC_vs_DN.xlsx

Condition_FRC49a_vs_DN.xlsx

Condition_FRC49a_vs_FDC.xlsx

Log_Module_6_TP_<...>.txt

Analyse différentielle avec DESeq2 - 5

Script **DESeq2_FromSampleFiles_v1.16.R** à charger pour l'utilisation de ses fonctions.

Fonction principale : *DESeq2_FromSampleFiles()*

Output

Log_Module_6_TP_<date_heure>.txt Log de l'exécution du script

Graph_report_Condition.pdf Ensemble des graphes des analyses

- Global dispersion : ce graphe permet de valider la bonne normalisation par DESeq2

Après pré-filtrage mais avant filtrage des counts normalisés:

- Plot PCA, Heatmap
- Normalised read threshold computed by HTSFilter si filtrage HTSFilter demandé :
ce graphe donne le seuil de filtrage = nombre de counts normalisés minimum requis

Après pré-filtrage et filtrage des counts normalisés : impact du filtrage sur PCA et Heatmap

Pour chaque comparaison (condition_j versus condition_i)

- Plot logFC versus mean(normalized counts) avant modération du logFC
- Plot logFC versus mean(normalized counts) après modération du logFC
- Plot expressions géniques <condition_j> versus <condition_i> informant sur les gènes DE +/- filtrés par FC

Analyse différentielle avec DESeq2 - 6

Script **DESeq2_FromSampleFiles_v1.16.R** à charger pour l'utilisation de ses fonctions.

Fonction principale : *DESeq2_FromSampleFiles()*

Output

Condition_Counts_global.xlsx Résultats du **test global LRT**

Après pré-filtrage des counts bruts

- Counts_GlobalAnalysis counts normalisés : ligne= ENSG
col= échantillons
- Pval_GlobalAnalysis test global LRT :
ligne= ENSG
col= baseMean(counts normalisés)
pvalue
adjp

Après Filtrage des counts normalisés

- Counts_GlobalAnalysis
- Pval_GlobalAnalysis

Analyse différentielle avec DESeq2 - 7

Fonction principale : *DESeq2_FromSampleFiles()*

Output

Condition_<condition1>_vs_<condition2>.xlsx

Résultats du **test de Wald**

- Counts <xx> Condition_<condition1>_vs_<condition2> counts normalisés : ligne=ENSG, col=échantillons
- pval <xx> Condition_<condition1>_vs_<condition2> test de Wald <condition1> versus <condition2>

ligne= ENSG

col= baseMean(counts normalisés)

log2FoldChange (logFC modéré)

stat

pvalue

adjp

FC (FC linéaire modéré)

- Contents information récapitulatif des nombres ENSG selon filtrage

<xx> DE résultats filtrés par adjp < paramètre adjp

FC résultats filtrés par adj < paramètre adjp et $|FC| > \text{paramètre FC}$

<vide> - si paramètre record = TRUE – résultats non filtrés (seulement pré-filtrage pris en compte)

Analyse différentielle avec DESeq2 - 1

A vous de jouer

- utiliser le script **Module_6_DESeq2_Main_TP.R** pour faire l'analyse différentielle des populations FDC, FRCCD49a et DN à partir des données RNAseq, ceci pour toutes les comparaisons 2x2.

Analyse différentielle avec DESeq2 - 8

A vous de jouer

- modifier le script **DESeq2_Main_TP.R** pour faire l'analyse différentielle de toutes les comparaisons par rapport à la condition contrôle = DN, avec une correction de tests multiples = FDR et un filtrage manuel des counts normalisés (nombre min de counts normalisés par gène pour chaque échantillon = 100).