Mars 2016

Formation à l'analyse de données RNA-seq

Exercices

Liens utiles

Données publiques :

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
http://www.ebi.ac.uk/ena/

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
http://www.ensembl.org/index.html

Logiciels utilisés :

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis. *http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/*

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.
http://tophat.cbcb.umd.edu/

STAR is a Spliced Transcripts Alignment to a Reference.
https://github.com/alexdobin/STAR

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols. http://cufflinks.cbcb.umd.edu/

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. http://samtools.sourceforge.net/

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types including sequence alignments, microarrays, and genomic annotations.
http://www.broadinstitute.org/igv/

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development.
http://bioconductor.org/

R is a free software environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
http://www.r-project.org/

Objectifs:

Cette formation a pour but de vous aider à traiter les séquences issues des SGS (Seconde Generation Sequencing) en particulier les plates-formes Illumina HiSeq. Vous y découvrirez les nouveaux formats de séquences, les biais connus et mettrez en œuvre des logiciels d'alignement épissé sur génome de référence, la recherche de nouveaux gènes, de nouveaux transcrits et la quantification de l'expression de ces gènes et transcrits.

Pré-requis: savoir utiliser un environnement Unix.

Pour réaliser l'ensemble de ces exercices, connectez-vous sur votre **compte « genotoul »** en utilisant « putty » depuis un poste windows ou la commande ssh depuis un poste linux.

Vous pouvez également utiliser un des comptes formation : anemone aster bleuet iris muguet narcisse pensee rose tulipe violette

Pour les traitements « lourds » utilisez le cluster avec la commande **« qlogin » ou « qrsh »**.

Sur genotoul, créer, dans votre répertoire work, un répertoire de travail : **tp_rnaseq.**

```
ssh -X user@genotoul.toulouse.inra.fr   OU Putty
qlogin OU qrsh
mkdir tp_rnaseq
cd tp_rnaseq/
```

Exercice n°1: Data Quality

- Récupérer les données re-formatées pour l'étude du chromosome 6 de la Tomate depuis la page web de la formation: http://bioinfo.genotoul.fr/index.php?id=119.

```
wget http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reads/MT_rep1_1_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reads/MT_rep1_2_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reads/WT_rep1_1_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reads/WT_rep1_2_Ch6.fastq.gz
```

Analyse de la qualité des données (en se connectant sur un noeud):

```
fastqc MT_rep1_1_Ch6.fastq.gz
fastqc MT_rep1_2_Ch6.fastq.gz
fastqc WT_rep1_1_Ch6.fastq.gz
fastqc WT_rep1_2_Ch6.fastq.gz
```

- Quelle est la longueur des lectures ?
  101
- Quelle est la qualité du séquençage ?
  Bonne
- Regarder les résultats concernant les biais décrits lors du cours, lesquels retrouve-t-on ?

  - Hexamer random primer (Per base sequence content & Per base GC content)

Exercice n°2: alignement/visualisation

Quelques liens:

- Tophat: *http://tophat.cbcb.umd.edu/*
- Samtools: *http://samtools.sourceforge.net/*
- Bowtie*: http://bowtie-bio.sourceforge.net/index.shtml*
- FTP download de Ensembl*:* http://www.ensembl.org/info/data/ftp/index.html
- STAR: https://github.com/alexdobin/STAR

Aujourd'hui nous allons nous focaliser sur l'alignement sans transcriptome de référence avec les paramètres de base. Pour lancer l'alignement il vous faut une référence.

- Créer un répertoire bowtie2-index.

```
mkdir  bowtie2-index
```

- Depuis la page de la formation sur le web, récupérer la séquence du chromosome 6 (ITAG2.3_genomic_Ch6.fasta) sur genotoul.

```
wget
http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reference/ITAG
2.3_genomic_Ch6.fasta
```

- Générer l'index bowtie2 des séquences fasta (suivre les indications suivantes)
  syntaxe : bowtie2-build [options]* <reference_in> <bt2_index_base>

  a) Saisir bowtie2-build sans paramètres pour obtenir l'aide.

```
bowtie2-build
```

  b) Lancer la commande sur le fichier fasta précédemment téléchargé en spécifiant comme nom d'index 'bowtie2-index/tomato_chr6'.

```
bowtie2-build ITAG2.3_genomic_Ch6.fasta bowtie2-build/tomato_chr6
```

  c) Lister le contenu du répertoire bowtie2-index. A quoi correspondent les fichiers *.bt2 ?

```
ls bowtie2-index/
tomato_chr6.1.bt2  tomato_chr6.3.bt2  tomato_chr6.rev.1.bt2
tomato_chr6.2.bt2  tomato_chr6.4.bt2  tomato_chr6.rev.2.bt2
```

Sur le serveur genotoul les génomes sont déjà indexés pour vous dans /bank/bowtie2db/ . Vous pouvez directement les utiliser pour réaliser l'alignement.

- Réaliser les alignements épissés (suivre les indications suivantes):
  syntaxe : tophat [options] <bowtie_index> <reads1.1[,reads2.1,...]>\
  [reads1.2[,reads2.2,...]]

  a) Saisir tophat pour obtenir l'aide du logiciel d'alignement.

```
tophat
```

  b) Quelle version de tophat est utilisé ? (tophat -v)

```
tophat --version
TopHat v2.0.14
```

  Quelle est la dernière version de tophat disponible sur internet ?

```
TopHat 2.1.1 release 2/23/2016
```

  Quelle est la version la plus récente disponible sur genotoul ? (lister le répertoire /usr/local/bioinfo/src/tophat/)

```
version 2.0.14
ls /usr/local/bioinfo/src/tophat/
current                    README.txt  test_data.tar.gz  tophat-1.4.0.tar.gz
tophat-2.0.14.tar.gz       tophat-2.0.5.Linux_x86_64.tar.gz  tophat-
2.0.8b.tar.gz
How_to_use_tophat-2.0.8b  test_data   tophat-1.4.0      tophat-2.0.14
tophat-2.0.5.Linux_x86_64  tophat-2.0.8b
```

b) Lancer tophat sur 4 CPU :

- en paired-end
- avec une taille d'insert de 200bp
- et une taille maximale d'intron de 5000bp
- pour les jeux de données WT et MT contre la l'index nommé 'tomato_chr6', nommer respectivement le répertoire de sortie aln_tophat_wt et aln_tophat_mt

```
qsub -N tophat_wt -pe parallel_smp 4 -b Y 'tophat2 -o aln_tophat_wt --max-intron-length
5000 --mate-inner-dist 200 bowtie2-index/tomato_chr6 WT_rep1_1_Ch6.fastq.gz
WT_rep1_2_Ch6.fastq.gz '

qsub -N tophat_mt -pe parallel_smp 4 -b Y 'tophat2 -o aln_tophat_mt --max-intron-length
5000 --mate-inner-dist 200 bowtie2-index/tomato_chr6 MT_rep1_1_Ch6.fastq.gz
MT_rep1_2_Ch6.fastq.gz '
```

Rappel :
Pour lancer une commande sur le cluster en réservant 4 CPU utiliser la commande :
qsub -N job_name -pe parallel_smp 4 -b Y 'ma commande'

Pour vérifier l'avancement des calculs utiliser la commande :
qstat -u nom_utilisateur

c) Vérifier que votre job tourne sur le cluster et est lancé sur 4 CPU (qstat)

```
qstat -u LoginUser
```

En option, si vous souhaitez pendant que les calculs tournent, réaliser un alignement STAR suivre les indications suivantes, sinon passer cette section grisée.
Voir le manuel : https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
Penser à vérifier la version du logiciel. (STAR --version)
Créer et se déplacer dans un repertoire nommé star.

Indexer la référence :

```
mkdir star-index
STAR --runMode genomeGenerate --genomeDir star-index --genomeFastaFiles
../ITAG2.3_genomic_Ch6.fasta
```

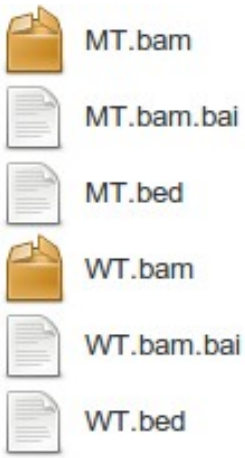
nb. Si vous avez le GTF de référence il est recommandé de l'utiliser pour le TP nous ne l'utiliserons pas.

Aligner le WT contre l'index créé :

```
STAR --runThreadN 4 --genomeDir star-index --readFilesIn ../WT_rep1_1_Ch6.fastq.gz
../WT_rep1_2_Ch6.fastq.gz --readFilesCommand zcat --alignIntronMax 5000
--outFileNamePrefix ./aln_star_wt --outSAMtype BAM SortedByCoordinate
```

Important si votre librairie n'est pas brin spécifique et que vous souhaitez faire de la découverte de nouveaux transcrits avec cufflinks, il faut ajouter l'option –outSAMstrandField intronMotif

- Visualiser le contenu du fichier align_summary.txt dans chacun des répertoires de sortie.
```
more aln_tophat_wt/align_summary.txt
more aln_tophat_mt/align_summary.txt
```
- Utiliser « samtools flagstat » sur les fichiers accepted_hits.bam pour obtenir le nombre de read alignés.
Syntaxe : samtools flagstat <in.bam>
```
qsub -N flagstat_wt -b Y 'samtools flagstat
aln_tophat_wt/accepted_hits.bam >
aln_tophat_wt/accepted_hits.flagstat'
qsub -N flagstat_mt -b Y 'samtools flagstat
aln_tophat_mt/accepted_hits.bam >
aln_tophat_mt/accepted_hits.flagstat'
```



- Quelle sont les différences entre le fichier align_summary.txt et ces résultats ?
```
Le nombre reads mappés est le même.
Le fichier accepted_hit.bam ne contient que les lectures alignés flagstat
donne donc un taux de mapping de 100 %.
```
- Indexer le fichier bam avec samtools (samtools index) pour pouvoir ensuite le visualiser avec IGV sur votre ordinateur.
```
qsub -N index_wt -b Y 'samtools index aln_tophat_wt/accepted_hits.bam'
qsub -N index_mt -b Y 'samtools index aln_tophat_mt/accepted_hits.bam'
```

- Télécharger sur votre ordinateur les fichiers de résultats de tophat (bam et junctions.bed) et le fichier d'indexation (bai)
```
Utiliser filezilla
```

- En local, renommer ces fichiers WT.bam, WT.bed, WT.bam.bai

Visualisation des résultats :

- Utilisez IGV pour visualiser les résultats sur votre poste de travail.
- Lancez IGV depuis « download » du site web de la formation (en bas de la page): http://www.broadinstitute.org/software/igv/download

- Chargez les annotations (fichier gtf mis à disposition dans http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reference/)
- Chargez les .bam, .bed

- Explorez l'interface, utiliser le clic-droit (pour visualiser toutes les isoformes, les paires de lectures associées....)

- Regardez les régions montrées dans le cours ainsi que les régions suivantes :

SL2.40ch06:2,786,806-2,807,064
SL2.40ch06:22,595-27,402
SL2.40ch06:38,480,842-38,484,938
SL2.40ch06:10,694,176-10,704,838
SL2.40ch06:9,839,693-9,862,815
Solyc06g009140.2.1

–

Exercice n°3: mesure d'expression brute au niveau gène/transcripts :

Manipulation du GTF, se familiariser avec sa référence :

- Depuis la page de la formation sur le web, récupérer le gtf ne contenant que le chromosome 6 : ITAG_pre2.3_gene_models_Ch6.gtf
  ```
  wget http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reference/ITAG_pre2.3_gene_models_Ch6.gtf
  ```
- A partir de ce fichier :
  - Combien y a-t-il de gènes ?(utiliser cut sur colonne 9, cut selon « ; », sort -u et wc)?
    ```
    cut -f 9 ITAG_pre2.3_gene_models_Ch6.gtf | cut -d ';' -f 1 | sort -u | wc
    ```
  - Combien y a-t-il de transcrits ?
    ```
    cut -f 9 ITAG_pre2.3_gene_models_Ch6.gtf | cut -d ';' -f 2 | sort -u | wc
    ```

Quantification au niveau transcrits à l'aide du gtf de référence et featureCounts :

- Saisir featureCounts sans options pour obtenir l'aide et la version.
  ```
  FeatureCounts
  ```
- Quelle est la version disponible sur genotoul et la dernière version disponible sur internet ?
  ```
  Genotoul : Version 1.4.5-p1
  Internet :Latest version 1.5.0-p1
  ```
- lancer featureCounts (sur un nœud du cluster) sur les deux bam en sachant que l'on compte les reads :
  - s'alignant sur les exons : `-t exon`
  - à regrouper par gène : `-g gene_id`
  - un read peut être assigné plusieurs fois :`-O`
  - la librairie n'est pas brin spécifique : `-s 0`
  - on ne veut compter que les alignements primaires (`--primary`) et les reads mappés de façon unique (`ne pas mettre -M`) ,
  - les reads ayant un alignement avec une qualité minimun de 20 : `-Q 20` (attention la qualité STAR n'est pas standard)
  - le nombre minimum de bases chevauchantes doit être > 10 : `--minReadOverlap 10`
  - on ne compte que les fragments (les paires) : `-P`
  - la distance entre deux reads doit être en 60 et 600 bp : `-d 60 -D 600`
  - on ne compte que les fragments dont chacun des reads est correctement alignés :`-B`

```
qsub -N featCount -pe parallel_smp 4 -b Y 'featureCounts -a
../ITAG_pre2.3_gene_models_Ch6.gtf -o tomato_count.txt -t exon -g gene_id -Q 20
--primary --minReadOverlap 10 -p  -P -d 60 -D 600 -B
./aln_tophat_mt/accepted_hits.bam ./aln_tophat_wt/accepted_hits.bam -T 4'
```

Exercice 4 : Recherche de nouveaux transcrits :

- Créer un répertoire 'cufflinks' pour l'analyse par cufflinks de l'ensemble du jeu de données.
  ```
  mkdir cufflinks
  ```

- Fusionner les alignements obtenu par échantillons dans un seul fichier.
  Syntaxe : samtools merge merge.bam fichier1.bam fichier2.bam ..
  ```
  samtools merge merge.bam ../aln_tophat_mt/accepted_hits.bam
  ../aln_tophat_wt/accepted_hits.bam
  ```

- Que signifie RABT ? A quoi sert l'option -g du cufflinks?
  ```
  RABT : reference annotation based transcript
  -g permet de fournir un gtf et de realiser un assemblage des transcrits.
  ```

- Quelle version de cufflinks est disponible sur genotoul ? Et sur internet ?
  ```
  Genotoul :cufflinks v2.2.1
  Internet : 2.2.1
  ```
- Lancer cufflinks en utilisant le fichier bam fusionné (afin d'obtenir un gtf complet correspondant à nos échantillons) avec les options suivante :
  - -g pour faire un assemblage RABT
  - library-type : fr-unstranded
  - max-intron-length : 5000
  - si vous souhaitez paralléliser utiliser l'option -p

  ```
  qsub -N cufflinks -pe parallel_smp 6 -b Y 'cufflinks -g
  ITAG_pre2.3_gene_models_Ch6.gtf -o cufflinks --library-type fr-unstranded --max-
  intron-length 5000 -p 6'
  ```

- Combien de transcrits obtenez vous ? Comparer ce résultat au comptage de l'exercice 3.
  nombre de transcrit :
  ```
  cut -f 9 cufflinks/transcripts.gtf | cut -d ';' -f2 | sort -u | wc
  4695
  ```
  nombre de gènes :
  ```
  cut -f 9 cufflinks/transcripts.gtf | cut -d ';' -f1 | sort -u | wc
  3147
  ```
- L'outil cuffcompare permet d'obtenir une comparaison entre deux fichiers d'annotation.
  Syntaxe : cuffcompare -r reference.gft cufflink1.gtf cufflink2.gtf …

  cuffcompare -r ITAG_pre2.3_gene_models_Ch6.gtf cufflinks/transcripts.gtf -o cufflinks/compare

  Extrayez du fichier tmap, les lignes dont la troisième colonne n'est pas '=' et allez voir pour chaque type de transfrag un exemple dans IGV, puis retournez voir les zones citées dans l'exercice 2.
  awk '$3!="=" cufflinks/compare.transcripts.gtf.tmap
- Lancer a nouveau featureCounts avec ce nouveau transcriptome de référence.
  ```
  qsub -N featCountNew -pe parallel_smp 4 -b Y 'featureCounts -a
  cufflinks/transcripts.gtf -o tomato_count_new.txt -t exon -g gene_id -Q 20
  --primary --minReadOverlap 10 -p  -P -d 60 -D 600 -B
  ```

```
./aln_tophat_mt/accepted_hits.bam ./aln_tophat_wt/accepted_hits.bam -T 4'
```

Exercice 5 : Un pas vers les statistiques (en option)

Toutes les informations sur l'étape de biostatistique sont disponibles en bas de la page suivante : http://bioinfo.genotoul.fr/index.php?id=119

- constituer la matrice attendue par le script R :

#gene_id	mt	wt
Solyc06g005000.2.1	240	72
Solyc06g005010.1.1	0	0
Solyc06g005020.1.1	1	0
Solyc06g005030.1.1	0	0
Solyc06g005040.1.1	0	0
Solyc06g005050.2.1	20	5

- Utiliser le script /usr/local/bioinfo/Scripts/bin/Normalization.R sur la matrice de comptage issue de featureCount.

```
/usr/local/bioinfo/src/R/R-3.2.2/bin/Rscript
/usr/local/bioinfo/Scripts/bin/Normalization.R -f tomato_count.R.txt -o norm
```

Copier en local le répertoire de sortie pour visualiser les images.

- Utiliser le script d'expression différentielle /usr/local/bioinfo/Scripts/bin/DEG.R sur un des résultats de la normalisation

```
/usr/local/bioinfo/src/R/R-3.2.2/bin/Rscript /usr/local/bioinfo/Scripts/bin/DEG.R -f
tomato_count.R.txt --norm norm/RLE_info.txt --pool1 mt --pool2 wt -o DEG
```

Copier en local le répertoire de sortie pour visualiser les images.