Octobre 2019

Formation à l'analyse de données RNA-seq

Exercices

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 GEO Profiles database stores gene expression profiles derived from curated GEO DataSets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a DataSet.

Expression Atlas is an open science resource that gives users a powerful way to find information about gene and protein expression across species and biological conditions such as different tissues, cell types, developmental stages and diseases among others.
https://www.ebi.ac.uk/gxa/home

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/*

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. http://cufflinks.cbcb.umd.edu/

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

RSEM: accurate quantification of gene and isoform expression from RNA-Seq data

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 plates-formes de séquençage Illumina HiSeq. Vous y découvrirez les formats de séquences et d'alignement 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 « `genologin` »** en utilisant « `putty` » depuis un poste windows ou la commande ssh depuis un poste linux.

Pour les traitements « lourds » utilisez le cluster avec la commande **«`srun --pty bash`» ou «`srun --x11 --pty bash`» (pour l'interface graphique)**.

Durant la formation utiliser la queue testq : -p testq

A savoir : pour tous **les logiciels de bioinformatique** vous avez sur le site web une description d'utilisation du logiciel.

Exercice n°1: Quality control and cleaning

- Sur `genologin`, créer, dans votre répertoire `work`, un répertoire de travail : **`tp_rnaseq`.**
- Récupérer les lectures re-formatées pour l'étude du chromosome 6 de la Tomate depuis la page http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/data/ (sous répertoire reads puis contient 4 fichiers fastq)

Vous pouvez télécharger les fichiers fastq directement sur votre compte « `genologin` » en utilisant la commande « `wget` » depuis `genologin` (en copiant l'adresse du lien et coller), penser à vous placer dans le répertoire correspondant sur `genologin`.

Lancer les 4 analyses fastqc en parallele

- Vous allez créer un fichier contenant l'ensemble des commandes pour les 4 fichiers fastq, pour cela suivez les étapes suivantes :

a) Trouver le module a charger,
b) Trouver la syntaxe de la commande,
c) Tester la syntaxe de la boucle :

```
for i in `ls *.fastq.gz`
do
echo "module load bioinfo/FastQC_v0.11.7; fastqc $i"
done
```

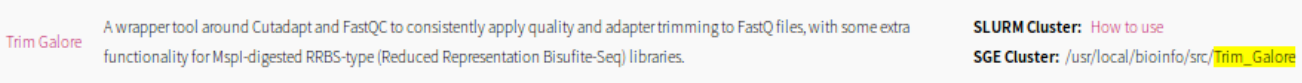
d) Pour générer un fichier contenant une ligne par fastq (module load XXX ; fastqc fichier.fastq.gz), utiliser la boucle en une seule ligne :
```
for i in `ls *.fastq.gz`; do echo "module load bioinfo/FastQC_v0.11.7; fastqc $i"; done > mescommandes.sh
```

e) Visualiser le contenu de `mescommandes.sh`
f) Lancer en une seule commande les 4 fastqc en parallèle (`sarray`)
g) Suivre l'avancement de vos calculs (`squeue`)

- Visualiser les résultats, deux possibilités :
  * sur genologin directement : *`firefox fastqc_report.html`*
  * ou en téléchargeant les résultats avec Filezilla
- Quelle est la longueur des lectures ?
- Quelle est la qualité du séquençage ?
- Regarder les résultats concernant les biais décrits lors du cours, lesquels retrouve-t-on ?

Exercice n°2 : Nettoyage des adaptateurs

a) Aller sur la page de la plateforme http://bioinfo.genotoul.fr/ → ressources → software
Rechercher le logiciel trim_galore, visualiser le contenu du « How_to_use »



b) Regarder le fichier de test situé dans le répertoire « `example_on_cluster` »
c) Créer un fichier de commande en vous inspirant de ce modèle et en spécifiant que le nombre de bases chevauchantes doit être d'au moins 3bp et demander un rapport `fastqc` après le nettoyage.

d) Lancer les commandes sur le cluster.

Nb : si vous souhaitez générer le fichier de commande comme dans la question 1.c vous pouvez récupérer :
- les fichiers paire par paire (`ls *.fastq.gz | paste -d ',' - -`)
- le nom de l'échantillon a partir du nom de fichier (`${i%*.fastq.gz}`)
*- ex : `for i in `ls *.fastq.gz | paste -d ',' - - `; do ech=${i%*.fastq.gz,*} ; echo "$ech: $i" ; done`*

Exercice n°3: Générer l'index STAR

L'aligneur `STAR` :
- Quelle est la version la plus récente disponible sur `genologin` ?
- Quelle est la dernière version de STAR disponible sur internet (« `rnaSTAR` ») ?
- Parcourir le manuel.

Générer l'index `STAR` à partir du fichier **fasta et du gtf :**

- Se connecter a un nœud du cluster en réservant 4 cpu (`-c 4`)
- Créer un répertoire `star-index` et se déplacer dedans.
- Depuis la page de données, récupérer la séquence et l'annotation du chromosome 6 (`ITAG2.3_genomic_Ch6.fasta` & `ITAG_pre2.3_gene_models_Ch6.gtf`).

- Indexer le génome.
- Lister le contenu du répertoire `star-index`. A quoi correspondent les nouveaux fichiers ?
- Se déconnecter du cluster

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

Exercice n°4: Réaliser les alignements épissés

- Se positionner dans `tp_rnaseq`

- Créer un fichier par échantillon contenant une ligne de commande STAR comme vu pendant le cours.

- Lancer l'exécution sur le cluster avec la commande `sbatch`.

- Vérifier que votre job tourne sur le cluster et est lancé sur 4 CPU (`squeue`)
- Combien de read sont alignés de façon unique et de façon multiple ? (voir `Log.final.out`)

Exercice n°5: Visualisation

Préparation pour la visualisation : création des fichiers d'index

- Se connecter sur un nœud.
- Indexer les fichier bam avec samtools (`samtools index`) pour pouvoir ensuite le visualiser avec `IGV` sur votre ordinateur.
- Télécharger sur votre ordinateur les fichiers de résultats de `STAR` (`bam` et `SJ.out.tab`) et le fichier d'indexation (`bai`)

Visualisation des résultats avec IGV 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 le génome (fichier fasta)
- Chargez les annotations (fichier `gtf`)

- Chargez les `*Aligned.sortedByCoord.out.bam, *Signal.Unique.str1.out.wig`
- 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`
  `SL2.40ch06:38,479,173-38,483,269`
  `SL2.40ch06:10,694,176-10,704,838`
  `Solyc06g009140.2.1`
  `SL2.40ch06:7,973,823-7,977,708`

Exercice 6 : Recherche de nouveaux transcrits

- **Manipulation du GTF, se familiariser avec sa référence :**
  A partir du fichier `ITAG_pre2.3_gene_models_Ch6.gtf`, compter combien il y a de transcrits. (Utiliser `cut` sur colonne 9, `cut` selon « ; », `sort` et `wc`)
- Se positionner sur un nœud
- Penser a charger les modules nécessaires pour samtools et cufflinks
- Fusionner les alignements obtenus dans un seul fichier.
  Syntaxe : `samtools merge -@ 4 merge.bam fichier1.bam fichier2.bam ..`
- Quelle version de `cufflinks` est disponible sur `genologin` ? Et sur internet ?
- Lancer `cufflinks` en utilisant le fichier `bam` fusionné (afin d'obtenir un `gtf` complet correspondant à nos échantillons) avec les options suivantes :
  - `-g` pour faire un assemblage guidé
  - `library-type` : fr-unstranded
  - `max-intron-length` : 5000
  - si vous souhaitez utiliser 4 CPU ajouter l'option `-p 4`
- Combien de transcrits obtenez vous ? Comparer ce résultat au comptage de l'exercice 5 (première question)
- L'outil `cuffcompare` permet d'obtenir une comparaison entre deux fichiers d'annotation.
- 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`.
  `Par exemple :`
  - `Solyc06g083520.2.1 (j)`
  - `Solyc06g083580.2.1 (e)`
  - `Solyc06g005250.2.1 (i)`
  - `Solyc06g005220.2.1 (o) ...`

- Vous pouvez aussi retourner voir les zones suivantes citées dans l'exercice 2 :
  - `SL2.40ch06:10,694,176-10,704,838`
  - `SL2.40ch06:2,786,806-2,807,064`

Exercice n°7: estimation de l'expression

Si nous voulons faire la quantification sur le nouveau transcriptome, il faut réaliser a nouveau l'alignement car RSEM utilise les alignements sur le transcriptome (*Aligned.toTranscriptome.out.bam) réalisés par STAR.

a) **Préparation de l'index RSEM**
Pour estimer l'abondance avec `RSEM`, il faut un fichier de référence. Préparer la référence à l'aide du programme `rsem-prepare-reference.`

b) **Lancer la quantification à l'aide des options présentées en cours**
```
rsem-calculate-expression --paired-end --alignments alignment.bam
[...options] rsem_lib quant
```

c) Création de la matrice de comptage
Utiliser le script suivant pour créer la matrice de comptage :
```
/usr/local/bioinfo/Scripts/bin/merge_cols.py -f
QuantMT.genes.results,QuantWT_Quant.genes.results -n MT,WT -c 5 -o
matrice.txt
```

d) Vous pouvez réaliser la quantification brute sur le nouveau transcriptome à l'aide de featureCount :
```
featureCounts -a transcripts.gtf -o featureCounts.txt -Q 20 --minOverlap
10 sample1.bam sample2.bam ...
```

Exercice 8 : Un pas vers les statistiques (en option)

Toutes les informations sur l'étape de biostatistique sont disponibles dans http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/doc/ :

- page nommée StatisticsOnGenotoul.html ou directement sur la biostat page
- documentation de scripts disponibles sur genologin : ScriptsDocumentation.pdf

Une formation spécifique est dispensée par la plateforme Genotoul Biostat. Les documents sont disponible en ligne http://www.nathalievialaneix.eu/teaching/rnaseq.html. Ce TP est largement inspiré de ces ressources.

Constituer la matrice attendu par le script R

Utiliser les script `merge_cols.py` ou faire un `cut` Unix pour ne sélectionner que les colonnes d'intérêt

```
#gene_id   mt    wt
Solyc06g005000.2.1    240   72
Solyc06g005010.1.1    0     0
Solyc06g005020.1.1    1     0
```

- Créer un répertoire `rnaseq_stat`.
- Pour des raisons pratiques nous utiliserons la matrice mise à disposition par la plateforme Biostat.

wget http://www.nathalievilla.org/doc/gz/RNAseq_data.tar.gz

- Décompresser, le fichier que l'on utilisera se trouve dans les sous-répertoires :

`RNAseq_data/count_table_files/count_table.tsv`

Suivre les indications fournis a cette page : http://bioinfo.genotoul.fr/index.php/rnaseq-bioinfobiostats/

A chaque étape télécharger le répertoire de résultats, (le fichier `Rplots.pdf` créé dans le répertoire d'exécution) et explorez les graphiques.