

Formation à l'analyse de données RNA-seq Galaxy

Liens utiles

Données publiques :



European Nucleotide Archive

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 :



Galaxy is an open, web-based platform for data intensive biomedical research.

Whether on this free public server or [your own instance](#), you can perform, reproduce, and share complete analyses.



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/>



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/>



Integrative Genomics Viewer

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
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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 (GAIIx, 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 Galaxy.



Pour réaliser l'ensemble de ces exercices, connectez-vous avec votre utilisateur genotoul <http://sigenae-workbench.toulouse.inra.fr/> depuis un navigateur.

Les données que nous utiliserons sont accessible à cette adresse : http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/

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

Exercice n°1: Data Management

Quelques liens:

- EMBL-ENA (European Nucleotide Archive) : <http://www.ebi.ac.uk/ena/>

Étude des données publiques disponibles à EMBL ENA:

- données correspondantes aux runs ERR022486 et ERR022488
- sujet de l'étude :
ERR022486 : RNA from Zebrafish 1day, embryo (31 504 560 reads)
ERR022488 : RNA from Zebrafish 3 days, embryo (24 920 613 reads)
- type de séquenceur utilisé : Illumina Genome Analyzer II
- type de librairie de séquençage (protocole) utilisé :
PAIRED

Abstract: **Paired-end sequence** data has been generated using **polyA selected RNA** from a range of zebrafish tissues and developmental stages using the **Illumina Genome Analyzer**. These data have been used to improve the gene annotation of the zebrafish genome. Study description: Zebrafish total RNA was extracted from embryonic and adult tissue, then polyA selected. After fragmentation and reverse transcription Illumina sequencing libraries were prepared. Paired-end sequence runs were performed with 36, 37, 54 and **76 base reads** on the Illumina Genome Analyzer.

Exercice n°2: Qualité et Nettoyage

- Lancer fastQC sur l'échantillon ERR022486.
- Lancer sickle sur tous les échantillons avec les options suivantes :
 - Minimal length of 20
 - Minimal mean quality of 20
 - No N in seq
 - No 5' trimming

- Lancer fastQC sur les échantillons nettoyés, combien de séquences ont-été supprimée pour l'échantillon 86 ?

On passe de 850245*2 (brutes) à 817391*2 (nettoyées)

Sickle (version 1.0.0)

Single-End or Paired-End reads?:
 Paired-End

Paired-End Forward Strand FastQsanger Reads:
 1: ERR022486_R1

Paired-End Reverse Strand FastQsanger Reads:
 2: ERR022486_R2

Quality Threshold:
 20

Length Threshold:
 20

Disable 5'-end trimming:

Discard any sequence with any number of Ns:

Execute

Exercice n°3: 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>

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, vérifier si cette référence existe en utilisant tophat.



Si votre génome d'intérêt n'existe pas veuillez faire une demande auprès du support ou (seulement si le génome est petit) charger le fichier fasta et utiliser dans tophat un index « from my history ».

- 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 ERR022486, ERR022488 contre la référence nommée « Danio rerio Zv9 62 chr 22 » en

Tophat for Illumina (version 1.0.0)

Your RNA-Seq FASTQ file (read 1):
 12: 88_R1_sickle

Your RNA-Seq FASTQ file (read 2):
 13: 88_R2_sickle

Select a reference genome:
 Danio rerio Zv9 62 chr 22

Number of threads used to align reads:
 4

Maximum intron length:
 5000

Expected (mean) inner distance between mate pairs:
 200

Your RNA-seq FASTQ file are zipped:
 Yes
 Please check this option if your files are zipped.

GTF file available:
 Yes
 Do you have a gtf file available ?

Your GTF file:
 24: http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/data/reference/Danio_rerio_chr22.Zv9.62.gtf

Execute

Vous obtenez en résultat 3 fichiers (pouvant être renommés) :

- Fichier de jonction (bed)
 - Fichier d'alignement (bam)
 - Fichier des reads non alignées (unmapped.bam)
- Utiliser « samtools flagstat » sur le fichier bam de chaque alignement, pour obtenir un résumé des statistiques d'alignement.
 - Indexer le 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 top-hat (bam et bed) et le fichier d'indexation (bai)
 - Renommer ces fichiers ERR022488.bam, ERR022488.bed, ERR022488.bam.bai

Visualisation des résultats :

- Utiliser IGV pour visualiser les résultats sur votre poste de travail.
- Lancer IGV depuis « download » du site web de la formation (en bas de la page):
<http://www.broadinstitute.org/software/igv/download>
 Le génome zebrafish est déjà intégré dans IGV
- Vous pouvez également charger les annotations correspondant au chromosome

22 en tant que fichier (disponible à http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/data/reference/Danio_rerio_chr22.Zv9.62.gtf.gz).

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

Pour info :

Lors d'une analyse future, il se pourra que les noms des chromosomes soient différents entre ceux intégrés dans IGV et ceux de votre gtf

Dans ce cas, il faut utiliser un fichier de correspondance ; par exemple : Zv9_alias.tab à mettre dans le répertoire :



/home/...../igv/genomes

<http://www.broadinstitute.org/software/igv/LoadData/#aliasfile>

Ce fichier devra alors contenir les correspondances entre les noms utilisés par IGV et les noms de votre GTF :

```
1   chr1
2   chr2
3   chr3
4   chr4
5   chr5...
```

....

Exercice n°4: mesure d'expression brute au niveau gène/transcripts :

Manipulation du GTF, se familiariser avec sa référence :

- A partir du fichier référence du chromosome 22 :
Danio_rerio_chr22.Zv9.62.gtf
- Combien y a-t-il de gènes (voir outil d'analyse de GTF)?
- Combien y a-t-il de transcrits ?

*** Ensemble GTF statistics (version 1.0.0)**

Your GTF file on which you would count genes:

17: Danio_rerio_chr22.Zv9.62.gtf

Execute

This tool count uniq line or column.

Your GTF contains :

Number of genes : 1276

Number of transcripts : 2133

Quantification au niveau gènes à l'aide du gtf de référence et Htseq-count :

- trier les alignements à l'aide de samtools sort selon les read name (pré-requis pour htseq-count sur des données paired-ends) (penser à cocher yes)

*** Samtools sort (version 1.0.0)**

Your accepted hits bam file:

12: {ERR022488_read1}-Tophat.bam

Sort by read names rather than by chromosomal coordinates:

Yes

Use this option if you want to sort by read names

Execute

This tool sort alignments by leftmost coordinates. File out.prefix.bam will be created.

Command : samtools sort -n in.bam out.prefix

OPTION: -n Sort by read names rather than by chromosomal coordinates

- lancer htseq count sachant que les données ne sont pas strand-spécifique et que l'on veut les intersections de gène non vide, on utilisera comme attribut le gene_id (par défaut c'est le transcript_id).

htseq (version 1.0.0)

Your accepted hits bam file (SORTED BY NAME):

Your gtf or gff file:

Use this option if you want to skip all reads with alignment quality lower than the given minimum value (default: 0):

Use this option to feature type (3rd column in GFF file) to be used, all features of other type are ignored:

GFF attribute to be used as feature ID (default,suitable for Ensembl GTF files: gene_id):

Select whether the data is from a strand-specific assay. Specify 'yes', 'no', or 'reverse' (default: yes). 'reverse' means 'yes' with reversed strand interpretation:

Select mode to handle reads overlapping more than one feature(choices: union, intersection-strict, intersection-nonempty; default: union):

- afin d'avoir un fichier de comptage complet, utiliser l'outil « Merge tabulated files» qui concatène les résultats de chaque colonne pour chaque ligne.

Merge tabulated files (version 1.0.0)

Your first file to merge:

Associated name for first file (You should enter either all or no fields names.):

Column of each files to extract, first column is 0, default 1:

Your second file to merge:

Associated name for second file:

Datasets

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

- lancer sigcufflinks avec le GTF de référence sur chaque échantillon séparément (sans recherche de nouveaux transcrits).

*** Sigcufflinks (version 1.0.0)**

Your accepted hits bam file:

Your gtf file:

G or g ?:

- Afin d'avoir un fichier de comptage complet, utiliser l'outil « Merge sigcufflinks » qui concatène les résultats de chaque colonne pour chaque ligne.

*** Merge sigcufflinks (version 1.0.0)**

Select a reference genome (if your genome of interest is not listed, please contact Siganae Team):

Your annotation file (defined feature to count):

Your first raw transcripts tsv file from sigcufflinks:

Your first raw transcripts tsv file name:

Datasets

Dataset 1

Other raw transcripts tsv file from sigcufflinks:

Other raw transcripts tsv file name:

Pour info :

Quelques outils pour l'analyse d'expression différentielle:

Il existe toute une batterie de package Bioconductor disponible pour l'analyse différentielle de données RNA-seq. Ces outils sont en plein développement et ne sont pas encore mature (difficulté dans le choix de la méthode à appliquer, choix de normalisation, évolution rapide des versions avec de fort changements d'une version à l'autre souvent...)

1. DESeq:

<http://bioconductor.org/packages/release/bioc/html/DESeq.html>

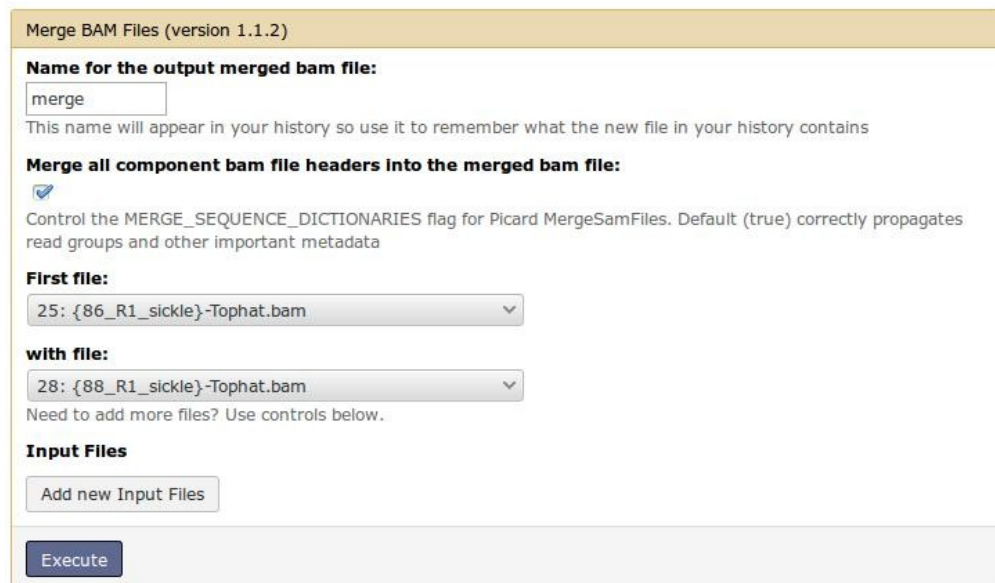
2. EdgeR

<http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html>

Voir publication : <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0103207>

Exercice 5 : Recherche de nouveaux transcrits :

- Créer un fichier bam contenant tous les alignements de tous les échantillons (avec Merge BAM Files)



Merge BAM Files (version 1.1.2)

Name for the output merged bam file:

 This name will appear in your history so use it to remember what the new file in your history contains

Merge all component bam file headers into the merged bam file:

 Control the MERGE_SEQUENCE_DICTIONARIES flag for Picard MergeSamFiles. Default (true) correctly propagates read groups and other important metadata

First file:

with file:

 Need to add more files? Use controls below.

Input Files

- Lancer cufflinks avec le GTF de référence, avec le fichier bam complet (afin d'obtenir un gtf complet correspondant à nos échantillons) en prenant l'option recherche de nouveaux transcrits (use ref as guide).

Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:

Max Intron Length:

Min Isoform Fraction:

Pre MRNA Fraction:

Perform quartile normalization:

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:

Reference Annotation:

Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):

- Lancer sigcufflinks avec le GTF issu de l'étape précédente (assembled transcripts), chaque échantillon séparément (sans rechercher de nouveaux transcrits : option quantitate against ref).

*** Sigcufflinks (version 1.0.0)**

Your accepted hits bam file:

Your gtf file:

G or g ?:

- lancer les comptages bruts Htseq-count comme précédemment.

En fin de formation:

Veuillez, en fin de TP, nettoyer votre compte de formation ("Delete permanently" de l'ensemble des "histories" créés).

Pour répondre à vos questions à propos de Galaxy:

- Mail : sigenae-support@listes.inra.fr
- Une FAQ et un manuel utilisateur sont disponibles depuis la page d'accueil de l'instance Sigenae de Galaxy.
- Les formations de la plateforme BIOINFO GENOTOUL sont disponibles sur <http://sig-learning.toulouse.inra.fr>

Enchaînement des logiciels

