Formation à l'analyse de données RNA-seq Galaxy

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 :

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/

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 (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 narcisse 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:

- sur le site recherchez les entrées correspondantes aux runs ERR022486 et ERR022488 (en une seule requête)
- Quel est le sujet de l'étude, quels tissus sont étudiés ?
- Quel est le type de séquenceur utilisé?
- Quel est le type de librairie de séquençage (protocole) utilisé

Dans ng6 :

- Récupérer les données re-formatées pour l'étude du chromosome 22 dans NG6 (les 4 fichiers fastq dans RawData et la référence au format gtf dans Analyse Annotation).

Avec FileZilla :

- copier/coller le chemin d'accès aux 4 fichiers fastq récupérés sur votre /work/user/

Dans galaxy :

- Uploader les 4 fichiers fastq
- Visualiser le contenu de chacun des fichiers pour vérifier leur chargement.
- Renommer vos datasets.

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 supprimé pour l'échantillon 86 ?

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

- 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 tophat (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...
….

<u>Exercice n°4:</u> 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 ?

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)

- 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) sur chacun des échantillons.

- afin d'avoir un fichier de comptage complet (pour les deux echantillons), utiliser l'outil « Merge tabulated files» qui concatène par ligne les résultats de chaque colonne (du fichier htseq-count.txt).

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

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

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 apliquer, 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)
- 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).
- Lancer cufflinks avec le GTF issu de l'étape précédente (assembled transcripts), chaque échantillon séparément (sans rechercher de nouveaux transcrits : option quantitate against ref).
- lancer les comptages bruts Htseq-count comme précédemment.

En fin de formation:
Veuiller, 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