Mai 2017

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 Report	**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/*
cutadapt	**Cutadapt** Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. Cutadapt helps to trim reads by finding the adapter or primer sequences in an error-tolerant way. It can also modify and filter reads in various ways. Adapter sequences can contain IUPAC wildcard characters. Also, paired-end reads and even colorspace data is supported. If you want, you can also just demultiplex your input data, without removing adapter sequences at all.
STAR	**STAR** is a Spliced Transcripts Alignment to a Reference. https://github.com/alexdobin/STAR
Cufflinks	**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/
SAMtools	**SAM** (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. http://samtools.sourceforge.net/
RSEM	RSEM: accurate quantification of gene and isoform expression from RNA-Seq data
Integrative	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 (username) : **anemone arome aster bleuet camelia capucine chardon clematite cobee coquelicot cosmos cyclamen dahlia digitale geranium gerbera**

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

Exercice n°1: using basic unix commands

1) Récupérer MobaXterm (http://mobaxterm.mobatek.net/download-home-edition.html) et le lancer (clic sur “Portable edition”), enregistrer le zip sur votre ordinateur, accepter la décompression du zip, puis double cliquer sur le fichier executable (.exe) pour lancer MobaXTerm. Connectez-vous sur le frontal : ssh - XY username@genotoul.toulouse.inra.fr
2) Afin de retrouver rapidement MobaXterm, créer un raccourci du fichier MobaXterm_Personal_10.2.exe sur votre bureau.
3) Connectez vous sur un noeud du cluster pour commencer à travailler (qrsh ou qlogin).
4) Listez les répertoires existants et déplacez-vous dans votre répertoire de travail (work).

5) Créer, dans votre répertoire work, un répertoire de travail : **tp_rnaseq** et un répertoire **reads/** puis positionnez vous dans ce répertoire nouvellement créé.

6) Récupérer, dans votre répertoire reads/, un par un les fichiers fastq contenant les lectures qui seront utilisées pour l'analyse des données sur le chromosome 6 de la Tomate. Ces données sont localisées dans le répertoire http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/reads/
Vous pouvez télécharger les fichiers fastq directement sur votre compte « genotoul » en utilisant la commande « wget » depuis genotoul (en copiant l'adresse du lien et coller), penser à vous placer dans le répertoire correspondant sur genotoul. Combien de fichiers récupérez-vous ? A quoi correspondent ces fichiers ?

7) Visualisez le contenu d'un de ces fichiers (zmore pour des fichiers compressés).

Exercice n°2: Format manipulation

1) Après avoir décompressé vos fastq.gz (`gunzip -c )`, transformez ces fichiers au format fasta (fastq_to_fasta_fast)
2) Combien de séquences contiennent ces fichiers (wc) ? Expliquez.
3) Quelle est la longueur des séquences (fastalength) ?

Exercice n°3: Analyse de la qualité des données avec fastqc

1) Quelle est la longueur des lectures ? Est-ce la même que celle que vous avez obtenue à l'exercice précédent ?
2) La qualité du séquençage vous parait-elle correcte ?
3) Regarder les résultats concernant les biais décrits lors du cours, lesquels retrouve-t-on ?

Exercice n°4: Nettoyage des données avec cutadapt et Sickle

1) Créer un répertoire cutadapt dans le répertoire tp_rnaseq et lancer la ligne de commande dans ce répertoire.
2) Supprimer les adaptateurs à l'aide de cutadapt, en spécifiant que la taille minimale des reads trimmées doit être de 36 pbs.
Adaptateurs : `-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT`
Penser à appeler les fichiers avec un nom explicite, par exemple : WT_rep1_1_Ch6.clean.fastq. Penser aussi à aller voir les log du logiciel.
3) Utiliser Sickle : tapez sickle pe -h. Comme précédemment enlevez les reads trimmées lorsqu'elles sont plus courtes que 36 pbs.
4) A quelle valeur de qualité sickle trimme-t-il par défaut ?
5) Relancer fastQC sur les résultats.
6) Quelle est la taille des fichiers fastq nettoyés ?
7) Compresser les fichiers fastq nettoyés avec la commande gzip. Quelle taille

font-ils maintenant ?

Exercice n°5: alignement/visualisation

A/ Générer l'index STAR à partir du fichier fasta et du gtf :

Se connecter à un nœud du cluster en réservant 4 cpu (`-pe parallel_smp 4`)
Créer un répertoire star-index dans le répertoire tp_rnaseq et aller dedans.
Depuis la page de données, récupérer la séquence du chromosome 6 (ITAG2.3_genomic_Ch6.fasta).
Rechercher le manuel de « rnaSTAR » sur internet.
Indexer le genome avec le fichier de transcriptome (`--sjdbGTFfile`) sur 4 threads (`--runThreadN 4`)

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

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

B/ Réaliser les alignements épissés

Quelle version de STAR est utilisé par défaut ? (`STAR --version`)
Quelle est la version la plus récente disponible sur genotoul ? (lister le répertoire `/usr/local/bioinfo/src/STAR/`)
Quelle est la dernière version de STAR disponible sur internet ?
Utiliser via le path absolu la dernière version de STAR.

Dans un nouveau répertoire star-mapping dans tp_rnaseq, créer un fichier contenant une ligne de commande STAR comme vu pendant le cours, en fournissant :

- le fichier de transcriptome (`--sjdbGTFfile`)
- les fichiers fastq nettoyés d'un des deux échantillons (`--readFilesCommand zcat --readFilesIn MT_rep1_1_Ch6.clean.fastq.gz MT_rep1_2_Ch6.clean.fastq.gz`)
- les tailles min et max des introns (`--alignIntronMin`, `--alignIntronMax`)
- le nombre max de mismatches (`--outFilterMismatchNmax`)
- comme les données ne sont pas brin spécifique, pour utiliser le fichier avec cufflinks il faut ajouter `--outSAMstrandField intronMotif` pour cufflinks encore il faut les options suivantes `--outFilterType BySJout --outFilterIntronMotif RemoveNoncanonical --outSAMattrIHstart 0`
- sortie souhaitée: BAM trié (`--outSAMtype BAM SortedByCoordinate`)
- pour la quantification `--quantMode TranscriptomeSAM` si on traite deux échantillons dans le même répertoire : `--outFileNamePrefix MT`
- nombre de threads

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`

Vérifier que votre job tourne sur le cluster et est lancé sur 4 CPU (qstat).
Quels sont les fichiers de sortie ?
Combien de read sont alignées de façon unique et de façon multiple ? (voir Log.final.out)

C/ Préparation pour la visualisation

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 STAR (bam et SJ.out.tab) et le fichier d'indexation (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 et les .bai

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

Regardez les régions suivantes :
SL2.40ch06:34,298,666-34,306,292
SL2.40ch06:34,209,900-34,260,000
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

NB. Pour des jeux de données de taille plus conséquente, pensez à trier et indexer

vos fichiers GTF avec les commandes `igvtools sort` et `igvtools index`.

Exercice n°6 : mesure d'expression brute au niveau gènes/transcripts

Manipulation du GTF, se familiariser avec sa référence :
À 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 uniq)?

Préparation du fichier de référence RSEM
Pour estimer l'abondance avec RSEM, il faut un fichier de référence.
Créer un répertoire `rsemGenomeDir` dans tp_rnaseq
Préparer la référence à l'aide du programme `rsem-prepare-reference`

Création de la matrice de comptage.
Créer un répertoire rsem_out et dedans : utiliser `rsem-calculate-expression` pour estimer le nombre de reads sur les gènes annotés dans chacune des conditions.
Utiliser le script suivant pour créer la matrice de comptage :

```
/usr/local/bioinfo/Scripts/bin/merge_cols.py -f
MT_Quant.genes.results,WT_Quant.genes.results -n MT,WT -c 4 -o matrice.txt
```

Exercice 7 : Recherche de nouveaux transcrits et comparaison des gtf

Créer un répertoire 'cufflinks' dans tp_rnaseq pour l'analyse par cufflinks de l'ensemble du jeu de données.

Que signifie RABT ? A quoi sert l'option -g de cufflinks?
Quelle version de cufflinks est disponible sur genotoul ? Et sur internet ?

Lancer cufflinks sur chaque condition avec les options suivantes :
- -g pour faire un assemblage RABT
- library-type : fr-unstranded
- max-intron-length : 5000
- si vous souhaitez paralléliser utiliser l'option -p

Utiliser cuffmerge pour fusionner les 2 nouveaux gtf.

Combien de transcrits obtenez vous ? Comparer ce résultat au comptage de l'exercice 6.

L'outil cuffcompare permet d'obtenir une comparaison entre deux fichiers d'annotation.
Syntaxe : cuffcompare -r reference.gtf merged.gtf ...

Extrayez du fichier tmap, les lignes dont la troisième colonne n'est pas '=' et allez voir pour chaque type de transfrag un exemple issu du nouveau gtf (merged.gtf) dans IGV, puis

retournez voir les zones citées dans l'exercice 5.

Pour cela pensez à étendre la piste du gtf.

Exercice 8 : réalignement et recomptage (facultatif)

Lancer a nouveau RSEM (et donc STAR) avec ce nouveau transcriptome de référence.

Pour cela, il vous faudra commencer par modifier un peu le merged.gtf. En effet certains des transcrits ne sont assignés à aucun des deux brins. Pour les garder il faut les remplacer par deux transcrits : un dans un sens et un dans l'autre.

```
cat merged.gtf | perl -laF"\t" -ne 'if ($F[6] eq "."){$F[6]="+";$F[8]=~s/transcript_id
"TCONS_0/transcript_id "TCONS_1/;print join("\t",@F);$F[6]="-";$F[8]=~s/transcript_id
"TCONS_1/transcript_id "TCONS_2/;}print join("\t",@F)' > compliant.gtf
```

Inspirez-vous ensuite des exercices 5 et 6 pour les lignes de commandes star et RSEM en utilisant de nouveaux répertoires de sortie.

Alignez-vous plus de reads en guidant l'alignement avec le nouveau gtf ?
Assignez-vous plus de reads sur ces nouvelles annotations ?

Il existe un autre outil pour comparer un/des fichier(s) gtf à un gtf de référence :

Lisez bien l'aide de l'outil /home/sigenae/bin/gffcompare2highcharts.pl et lancer les /home/sigenae/bin/gffcompare nécessaires au préalable.

Lancer le et regarder le résultat à l'aide d'un navigateur web.