RESEARCH

Pipeline d'analyse de données sRNAseq

Alexandra Mancheno Ferris[*†] and Soukaïna Timouma[†]

[*]Correspondence: manchenoferrisalexandra@gmail.com
Master 2 Bioinformatique et Biologie des Systèmes, Université Paul Sabatier, Toulouse, FR
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

First part title: Pipeline d'analyse de données sRNAseq. Ce pipeline permet de nettoyer les données, puis d'aligner les *reads* sur un génome de référence. Après une étape de normalisation des données (dans ce wrapper : RPKM), l'analyse des *reads* s'alignant sur le génome est effectuée. Nous avons intégré les différentes étapes dans un *wrapper* sur Galaxy.

Keywords: pipeline d'analyse; sRNAseq; Galaxy; wrapper

Introduction

Les données de transcriptomiques sont essentielles dans l'étude des gènes et de leur régulation. Le nombre de ces données est en croissance exponentielle depuis l'arrivée des technologies de séquençage.

Le transcriptome est l'ensemble des transcrits présents dans une population de cellules dans des conditions données.

Il existe plusieurs type d'ARNs, les codants (ARNm) et les non-codants.

Parmi les non codants, nous pouvons classifier selon leur taille les grands ARNs (rRNA, tRNA,...), les petits ARNs (snoRNA, snRNA,...) et les micros ARNs (miRNA, piRNA,...).

Les petits ARN non codants et les microRNA (miRNA) sont des régulateurs majeurs de l'expression des gènes, de la formation de l'hétérochromatine et de l'organisation du noyau. Leur action est dépendante du tissu, de l'environnement, et par des mécanismes variés, a une influence capitale dans les cancers et de nombreuses maladies.

Dans ce projet, nous avons choisi de travailler avec des données miRNAs. Du fait de l'absence de queue polyA et de leur taille, les techniques de RNAseq ne peuvent pas être utilisées pour l'étude des miRNAs. La technique utilisée est le small RNAseq.

La figure 1 présente le méchanisme d'action des miRNAs. Les gènes codant pour des miRNA sont transcrits en pri-miRNA par des ARN polymérases II/III, dans le noyau. Ces pri-miRNA portent un motif structural de type tige-boucle (*stem-loop motif*), une coiffe en 5' et une queue polyA en 3'. L'endoribonucléase Drosha, associée à d'autres protéines (DGCR8, PACT), supprime les régions 5' et 3' de part et d'autres de la tige boucle, ce qui forme le pré-miRNA. Les pré-miRNAs sont ensuite exportés activement dans le cytoplasme. La ribonucléase DICER, avec l'aide de TRBP, clive la région correspondant à la boucle, ce qui génère un complexe pré-miRNA/miRNA, partiellement complémentaire. Ce complexe est ensuite chargé par le complexe multi-protéique RISC, où l'un des deux brins ARN est dégradé, et

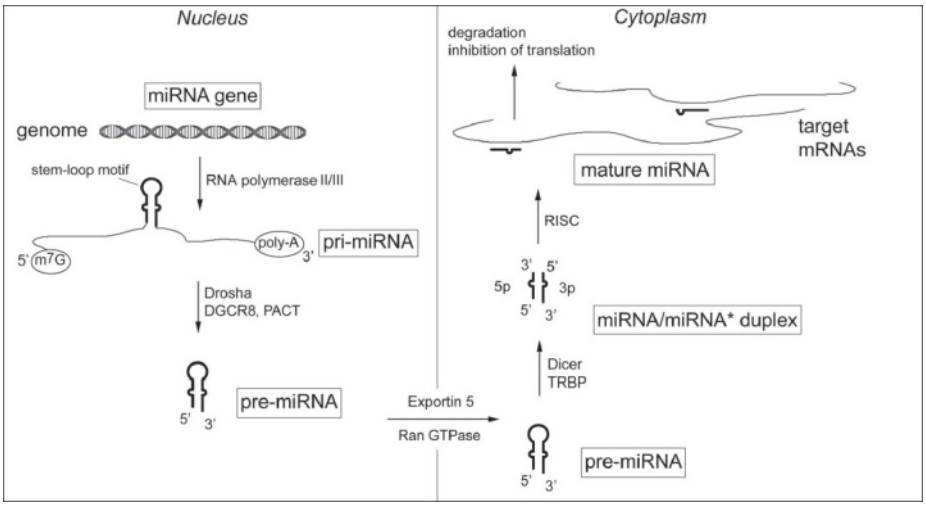

Figure 1 Méchanisme d'action des miRNA (DOI: 10.4103/1477-3163.67074)

l'autre, le miRNA mature, est guidé jusqu'à la cible ARN messager. L'effet est la répression de la transcription ou la dégradation des ARN messagers.

Les miRNA ont la particularité d'être conservés du point de vue de la structure.

Les plateformes qui font du sRNAseq sont les suivantes :

- Illumina
- HiSeq
- 454

A partir des ARN totaux, les petits ARNs sont purifiés sur gel d'acrylamide et soumis à la préparation des librairies en vue de leur séquençage. Cette étape rajoute un adaptateur spécifique à chaque extrémité des molécules d'ARN, convertit les ARN en cDNA double brin et réalise une étape d'amplification. Les librairies sont ensuite contrôlées et quantifiées.

Dans l'étude des miRNAs, plusieurs questions peuvent se poser : les miRNA étudiés sont-ils connus ou nouveaux, quelle est leur cible, en quelle quantité les trouve t-on et dans quelle condition, pouvons nous mettre en évidence une expression différentielle...

Les premières choses que nous devons vérifier avec les données que nous traitons est s'il existe un génome de référence et quel type de données avons-nous, *single-end* ou *paired-end*. En effet, S'il n'y a pas de référence, nous devons faire de l'annotation à travers les bases de données miRbase, Rfam, Silva et GtRNAdb. Mais s'il y a une référence, nous alignons les *reads* dessus (étape de *mapping*) avec Bowtie ou BWA (pour ce projet Bowtie2 a été retenu). Avec mirTrap et MirDeep2 nous pouvons faire de la prédition de nouveaux miRNA.

En ce qui concerne le trancriptome de référence, pour les étapes d'analyses, nous devons sélectionner un miRNA qui n'est pas affecté par les conditions expérimentales étudiées, et qui ne présente pas de variabilité dans son modèle

d'expression. Il est préférable de prendre un miRNA de longueur comparable (en raison des divers effets sur l'efficacité des sondes utilisées sur la PCR).

Présentation du pipeline

Ci-dessous les étapes suivies :

1 Vérification de la qualité des données

Nous utilisons pour cela l'outil FASTQC Cette étape est nécessaire dans la mesure où les échantillons traités sont des échantillons d'ARN totaux issus de cultures de cellules, de tissus, de biopsies, de fluides mais également de miRNA préamplifiés issus de capture laser ou de tri cellulaire, et aussi de tissus FFPE.

2 Nettoyage des séquences

Suppression des adaptateurs avec Cutadapt

3 Alignement des séquences

avec *Bowtie2*. Suppression des séquences redondantes, nous ne conservons que les lectures qui s'alignent sur un seul locus.

4 Comptage brut en utilisant l'outil *HTSeqCount* sur plusieurs jeux de données

5 Normalisation RPKM

L'étape de normalisation est nécessaire. En effet, un faible changement dans l'expression des miRNA peut être biologiquement très significatif. La normalisation se fait sur la taille de la librairie, la taille des *reads* et/ou le nombre de *reads*.

6 Obtention d'une matrice de quantification prête à être analysée

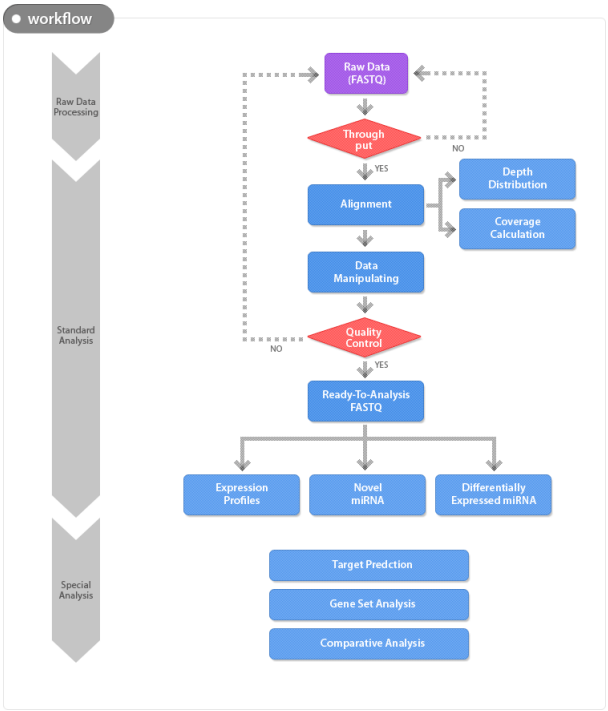

Figure 2 Exemple de pipeline de small RNA Sequencing, compagnie Macrogen

Pipeline avec un génome de référence présent

Chargement du fichier et détermination de la nature de l'input

Nous chargeons l'outil *Upload File* ou *GetData* déjà présent dans le workbench. Quand nos données sont chargés dans Galaxy, le pipeline à proprement parlé commence. Nous allons travailler à partir de données brutes issues directement du séquenceur. Le fichier est donc au format *Fastq*.

Contrôle qualité

Cette partie du pipeline a pour but de vérifier si les séquences (*reads*) obtenues après séquençage sont conformes au niveau des prestations attendues (taille, nombre, qualité ...). Pour cela, nous vérifions que les *reads* peuvent répondre aux critères concernant le biais technique et le biais technique. Tout cela permet de déterminer les paramètre de nettoyages des données. Nous présentons à ce stade une sortie HTML que l'utilisateur pourra analyser.

Il est important de tenir compte de l'existence des biais spécifiques dûs à la technique elle-même. Ces biais peuvent être dûs au mode de préparation de la banque : en effet, les banques sont générées par amplification héxamérique aléatoire (Random hexamer priming), ce qui induit un fort biais de composition des 13 premiers nucléotides. Le séquençage peut être également influencé par la composition des séquences (contenu en GC), par la longueur des transcrits. La capacité à observer un transcrit comme étant différentiellement exprimés est directement reliée à sa longueur.

La *mapabilité* du transcriptome apporte également un biais dans l'étude des données issues d'expérience de sRNAseq. La qualité de la référence, du fait de son assemblage, de sa finition, de sa composition (zones répétées) et de la qualité de son annotation, influe sur les étapes bioinformatiques de l'analyse de données de sRNAseq. La mise en place d'un pipeline est donc importante pour faciliter le traitement de ces données. Le but de ce projet est donc de développer un pipeline qui réalise le comptage des *reads* (brut et RPKM) s'alignant sur le génome puis l'intégration de celui-ci dans un ou plusieurs wrappers Galaxy.

Nettoyage des données

Après l'étape de contrôle qualité, nous préparons les données pour leur analyse. Cette étape de préparation des données se fait en deux parties:

- Pré-nettoyages des reads: suppression des adaptateurs de séquençages, suppression des adaptateurs de multiplexage
- Nettoyage des reads: tronquer les extrémités de mauvaise qualité, suppression des lectures contaminantes.

Nous allons utiliser **Cutadapt**, logiciel qui permet d'enlever l'adaptateur de séquençage pour n'avoir que la séquence des miRNA dans notre fichier.

Alignement des *reads* à la référence

Il y a diverses façons de choisir le génome référence où vont s'aligner les *reads*. Nous pouvons le sélectionner dans une liste de génomes de référence prédéfinis. S'il n'est pas présent, nous pouvons également le télécharger sur Ensembl ou NCBI. En ce qui concerne l'alignement des *reads*, nous souhaitons conserver uniquement les

reads qui s'alignent à un seul locus. Les *matchs* uniques et les *matchs* ambigus sont retirés de l'input à chaque étape. Le nombre de copies d'un *reads* est proportionnel à son niveau d'expression. La difficulté dans cette étape est de tenir compte de l'épissage, en particulier de l'épissage alternatif. Cette étape peut être réalisée en prenant en compte les listes d'exons-exons connus, mais cela doit être fait dans un temps raisonnable.

Il existe divers outils pour réaliser cette étape d'alignement contre la référence. STAR (Spliced Transcripts Alignment to a Reference) est un logiciel d'alignement rapide, performant qui utilise son propre logiciel d'indexation du génome. TopHat est une suite de programme qui permet d'aligner les *reads* au génome. Pour l'utiliser, il faut au préalable indexer le génome de référence. Nous réalisons cette étape grâce à Bowtie2 (*bowtie2-build*. Il existe 2 versions de Tophat, TopHat1 et Tophat2. Cependant Tophat1 présente des problèmes lors du *mapping* des *reads* sur le génome quand il y a des pseudo-gènes. Tophat2 est le logiciel le plus utilisé. A chaque étape, il utilise des heuristiques pour aligner les *reads* sur le génome, ce qui permet de résoudre des cas difficiles. Le logiciel Bowtie peut également permettre d'aligner les reads. Nous allons utiliser **Bowtie**, qui utilise le logiciel **Bowtie2** pour aligner les reads à notre référence. Nous allons travailler avec la version présente dans l'ordinateur de l'utilisateur.

Quantification des transcrits

Après avoir aligné les *reads* aux séquences, et après les avoir filtrés, nous pouvons les quantifier. Plus le gène est transcrit, plus il y a de miRNA, et plus il y a de *reads* alignés sur ce gène.

Nous avons utilisé **HTSeqCount**, qui permet de quantifier les *reads* pour chaque *feature*. Il fournit en sortie un résumé du traitement effectué. Il prend en entrée un fichier BAM trié ou non, le fichier *GTF* de la référence et compte le nombre de *reads* s'alignant sur chaque gène. Pour effectuer cela, nous faisons appel au logiciel **HTSeqCount**, présent dans l'ordinateur de l'utilisateur.

Statistiques : calcul du RPKM

Après avoir quantifié le nombre de *reads* pour chaque gène de l'organisme étudié, nous passons à l'étape de normalisation des données afin de pouvoir les comparer. Pour cela, nous avons utilisé la normalisation RPKM (*Read per kilo per million mapped read*). Ce calcul tient compte de la longueur des gènes. Cependant il a été montré que la correction de la longueur des gènes a pour effet d'introduire un biais dans la variance par gène, en particulier pour les gènes faiblement exprimés.

$$RPKM = \frac{nombre de Reads pour le gene}{\frac{longueur du gene}{1000} * \frac{nombre total de Reads de l'echantillon}{1000000}}$$

Implémentation du pipeline sur Galaxy

Nous avons fait un *wrapper* qui effectue les différentes étapes du pipeline. Nous avons implémenté ce pipeline sur Galaxy. Nous avons travaillé en local. Notre wrapper se présente sous forme d'un script *Python* qui effectue séquentiellement les différentes étapes du pipeline. Il fait appel aux logiciels déjà présents dans la machine de l'utilisateur. La seule exigence demandée est que l'utilisateur ait installé

les différents logiciels sur sa machine et qu'un lien symbolique pour Fastqc ait été fait. Les explications pour faire ce lien symbolique sont présentées dans le fichier *README*. Nous avons choisi de mettre à disposition de l'utilisateur les différents logiciels utilisé dans notre wrapper, dans le cas où celui-ci voudrait réaliser l'étude pas à pas, et avoir accès aux différents fichiers intermédiaires.

Ajout des outils que nous avons utilisé

```
<section id="mytools" name="My sRNAseq analysis">
  <tool file="my_tools/mRNA_wrapper/RNA_wrapper.xml" />
  <tool file="my_tools/fastqc/rgFastQC.xml" />
  <tool file="my_tools/tophat2/tophat2_wrapper.xml" />
  <tool file="my_tools/htseq_count/htseq-count.xml" />
  <tool file="my_tools/cutadapt/cutadapt.xml" />
  <tool file="my_tools/mirdeep2/mirdeep2.xml" />
  <tool file="my_tools/fastq_to_fasta/fastq_to_fasta.xml" />
```

Versions des logiciels utilisés

- FastQC:v0.11.5
- Cutadapt: v1.12
- HTSeqCount : v0.6.0
- Bowtie2: v2.2.6

Données de test

Nous avons testé notre *wrapper* avec les données des expériences de la publication *In vivo oncomiR screen identifies miR-21* independent of miR-21 as a driver in skin cancer (house mouse)* du Bioproject PRJNA281170 disponible sur https://www.ncbi.nlm.nih.gov/bioproject/281170 avec le sample téléchargé depuis ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR197/008/SRR1974288/SRR1974288.fastq.gz. Il s'agit de données issues de la souris. Nous avons téléchargé les données de référence sur ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz et le GFF des miRNA de la souris connus sur la base de données miRBase ftp://mirbase.org/pub/mirbase/CURRENT/genomes/mmu.gff3. Nous avosn également testé notre *wrapper* avec comme référence le génome de la souris téléchargé sur le site du NCBI à l'adresse ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/635/GCF_000001635.25_GRCm38.p5/GCF_000001635.25_GRCm38.p5_genomic.fna.gz. Nous avons également testé notre *wrapper* de la publication *miRNASeq of musculoskeletal ageing using human mesenchymal stem cells and their tissue constructs* avec le numéro d'accession PRJEB8905 , disponible à l'adresse https://www.ncbi.nlm.nih.gov/bioproject/352186. Nous utilisons l'échantillon ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR792/ERR792997/ERR792997.fastq.gz Il s'agit de données issues d'une étude sur l'Homme. Nous avons utilisé le fichier de référence précédemment téléchargé car il contient des informations sur diverses espèces et nous avons téléchargé les données GFF3 sur l'Homme disponibles à l'adresse ftp://mirbase.org/pub/mirbase/CURRENT/genomes/hsa.gff3.

Conclusion

Ce pipeline a pour but d'être utilisé dans le cadre d'une analyse de données issus de small RNAseq. Le sRNAseq permet de quantifier des petits et micro ARNs (eucaryotes et procaryotes), de caractériser des polymorphismes, de découvrir de nouveaux sRNA et de faire du *profiling* à partir de microquantités d'ARN. Cependant des études récentes montrent que Tophat ou même Cufflinks sont plus adaptés pour ce genre de recherche que Bowtie2.

References

Additional Files

Additional file 1 — mRNAwrapper.py
Script exécutable du wrapper en Python.

Additional file 2 — mRNAwrapper.xml
Fichier descriptif du wrapper

Additional file 3 — README
Fichier expliquant le fonctionnement du wrapper pour l'utilisateur

Additional file 4 – tool_conf.xml.sample
Fichier de configuration des outils de Galaxy

Additional file 5 – my_tools

Dossier contenant tout nos outils