

Analyse de données de ChIP-Seq

Denis Puthier

27 janvier 2011

Laboratoire INSERM TAGC/ERM206, Parc Scientifique de Luminy case 928,
13288 MARSEILLE cedex 09, FRANCE.

1 Les données

Les données sont issues de a base de données *SRA* hébergée au *NCBI*.

“The Sequence Read Archive (SRA) stores raw sequencing data from the "next" generation of sequencing platforms including Roche 454 GS System[®], Illumina Genome Analyzer[®], Applied Biosystems SOLiD[®] System, Helicos Heliscope[®], Complete Genomics[®], and others”

Il s’agit de données analysant la fixation du facteur de transcription NF-KB dans un modèle de lignée lymphoblastoïdes humaines. L’identifiant de cette étude dans SRA est *SRS025277* :

Summary : We examined genome-wide variation in transcription factor binding in different individuals and a chimpanzee using chromatin immunoprecipitation followed by massively-parallel sequencing (ChIP-Seq). The binding sites of RNA Polymerase II (Pol II) as well as a key regulator of immune responses, NFkB, were mapped in ten HapMap lymphoblastoid cell lines derived from individuals of African, European, and Asian ancestry, including a parent-offspring trio. We also mapped gene expression in all ten human cell lines for two treatment conditions : a) no treatment and b) following induction by TNF-alpha. Overall Design : Genome-wide comparison of Pol II and NF-KappaB binding in ten individuals. ChIP-seq with NF-KappaB.

Nous analyserons deux échantillons : SRR038464 et SRR038466. Quelles sont les caractéristiques de ces échantillons. Quel(s) traitement(s) ont-ils subi.

4 Alignement des étiquettes (reads) sur le génome

De nombreux logiciels sont disponibles pour cette opération. Nous utiliserons le programme *bowtie*. Celui-ci est disponible sous la forme d'un utilitaire en ligne de commandes fonctionnant sous *unix*.

```
[user@machine] cd ~/NGS_DATA
[user@machine] wget      http://garr.dl.sourceforge.net/project/bowtie-bio/bowtie/0.12.7/bowtie-0.12.7-src.zip
[user@machine] unzip    bowtie-0.12.7-src.zip
[user@machine] cd      bowtie-0.12.7
[user@machine] make
```

4.1 Indexation de la référence

Notre référence sera le chromosome 22 (version hg19). On le télécharge depuis l'UCSC et on l'indexe avec *bowtie-build*. Notez que des fichiers avec des génomes complets indexés sont disponible sur le site de *bowtie*.

Télécharger un chromosome et l'indexer :

```
[user@machine] cd ~/NGS_DATA/bowtie-0.12.7/indexes
[user@machine] wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/chr22.fa.gz
[user@machine] gunzip chr22.fa.gz
[user@machine] ../bowtie-build chr22.fa chr22.hs
```

4.2 Alignement

Les paramètres utilisés dans l'alignement peuvent influencer de manière très notable sur le résultat notamment si l'on ne tient pas compte de l'existence de séquence répétées dans le génome. Ici nous positionnons l'argument *-m* sur la valeur 1 pour indiquer à *bowtie* de ne pas considérer les étiquettes se positionnant sur plus d'une localisation (on peut sans doute, être un peu moins sévère sur ce critère en deuxième intention). Nous positionnons l'argument *-n* sur 1 pour indiquer à *bowtie* que nous acceptons un mismatch dans l'alignement. Nous utilisons l'argument *-best* pour toujours récupérer la meilleure solution d'alignements. Les reads non alignés sont stockés dans un fichier "*unmapped".

```
[user@machine] cd ~/NGS_DATA/SRA
[user@machine] ../bowtie-0.12.7/bowtie -p 7 --best -m 1 -n 0 --sam -l 28 --un SRR038466.chr22.unmapped chr22.hs
SRR038466.chr22.fastq > SRR038466.chr22.sam
[user@machine] ../bowtie-0.12.7/bowtie -p 7 --best -m 1 -n 0 --sam -l 28 --un SRR038464.chr22.unmapped chr22.hs
SRR038464.chr22.fastq > SRR038464.chr22.sam
```

On obtient en sortie un fichier SAM qui contient les informations sur l'alignement des étiquettes (chromosome, position, qualité de l'alignement au format *CIGAR*). La valeur 28M au format *CIGAR* indique 28 "matches" dans l'alignement.

```
[user@machine] less SRR038464.chr22.sam
```

4.3 Conversion SAM -> BAM

Le fichier SAM est ensuite compressé sous la forme d'un fichier BAM. Pour ce faire, nous devons installer les utilitaires de la suite *samtools*.

```
[user@machine] cd ~/NGS_DATA/  
[user@machine] wget http://switch.dl.sourceforge.net/project/samtools/samtools/0.1.11/samtools-0.1.11.tar.bz2  
[user@machine] bunzip2 samtools-0.1.11.tar.bz2  
[user@machine] tar xvf samtools-0.1.11.tar  
[user@machine] cd samtools-0.1.11  
[user@machine] make
```

Ensuite, on lance la conversion.

```
[user@machine] cd ~/NGS_DATA/SRA  
[user@machine] ../samtools-0.1.11/samtools view -bS -o SRR038466.chr22.bam SRR038466.chr22.sam  
[user@machine] ../samtools-0.1.11/samtools view -bS -o SRR038464.chr22.bam SRR038464.chr22.sam
```

Il y a ensuite nécessité de trier et indexer les fichiers.

```
[user@machine] ../samtools-0.1.11/samtools sort SRR038464.chr22.bam SRR038464.chr22.sorted  
[user@machine] ../samtools-0.1.11/samtools sort SRR038466.chr22.bam SRR038466.chr22.sorted  
[user@machine] ../samtools-0.1.11/samtools index SRR038466.chr22.sorted.bam  
[user@machine] ../samtools-0.1.11/samtools index SRR038464.chr22.sorted.bam
```

5 Visualisation des alignements dans IGV

Utilisez l'application *IGV* pour charger les alignements (chargez les fichiers BAM). Utilisez deux pistes, l'une pour le contrôle, l'autre pour le ChIP NF-KB.

Chargez la piste contenant les positions des éléments répétés au format *BED*.

```
[user@machine] cd ~/NGS_DATA/  
[user@machine] wget http://biologie.univ-mrs.fr/upload/p245/Repeat.tar.gz  
[user@machine] tar xvfz Repeat.tar.gz  
[user@machine] less RepeatMasker_wholeGenome_hg19.chr22.bed
```

6 Recherche des pics avec macs

Téléchargez la dernière version stable de *macs*. Installez la :

```
[user@machine] cd ~/NGS_DATA/  
[user@machine] tar xvfz MACS-1.3.7.1.tar.gz  
[user@machine] cd MACS-1.3.7.1  
[user@machine] chmod 777 setup.py  
[user@machine] python setup.py install --user  
[user@machine] chmod 777 bin/*
```

Maintenant on lance la recherche de pics dans le jeu de données :

```
[user@machine] cd ~/NGS_DATA/SRA/  
[user@machine] ../MACS-1.3.7.1/bin/macs --format=BAM -c SRR038464.chr22.bam -t SRR038466.chr22.bam --name NFKB.chr22 --g  
--tsize 28 --mfold 10  
[user@machine] sort -nrk5 NFKB.chr22_peaks.bed
```

7 Recherche du motif NFkB dans un pic

A partir du site de l'*UCSC*, sélectionnez dans *Galaxy* dans le menu. Dans *Galaxy*, récupérez la séquence correspondant au meilleur pic trouvé par *macs*. Sur le site Jaspar, scan cette séquence avec la matrice poids-position correspondant à NF-KB.