Déroulement du TP/TD « Détection de CNV »

Jeudi 17/01/2013, 15h15-16h45 & 17h15-18h15

Description du dataset

- La source des données est un WGS Illumina sur 2.5 pistes ayant généré 900M reads avant mapping (96% mappés).
- Les qualités globales de la lib et des séquences sont excellentes.
- L'échantillon est une lignée de leucémie tétraploïde : YT1.
- Le set sélectionné correspond à un grand fragment du chr17 (49Mb), choisi pour se nombreux remaniements, plus différents fragments de 2Mb chacun sur d'autres autosomes pour assurer la centralisation du profil (79Mb au total).
- L'ensemble totalise ~15M reads (1.5Mds nt).
- L'alignement de cette quantité de données étant bien trop long pour le TP, et le fichier aligné au format SAM trop lourd (5.2Go), le fichier d'entrée sera un fichier BAM après alignement par bwa, avec les paramètres par défaut.
- *Rq : les temps d'exécution notés dans les diapos suivantes correspondent à une exécution locale (pas sous Galaxy), ils seront donc certainement un peu plus long dans la réalité.*

Description du dataset (coordonnées)

- chr1:40,000,000-42,000,000
- chr2:153,000,000-155,000,000
- chr3:48,000,000-50,000,000
- chr5:110,000,000-112,000,000
- chr7:20,000,000-22,000,000
- chr8:75,000,000-77,000,000
- chr10:75,000,000-77,000,000
- chr11:75,000,000-77,000,000
- chr12:75,000,000-77,000,000
- chr13:48,000,000-50,000,000
- chr14:75,000,000-77,000,000
- chr16:50,000,000-52,000,000
- chr17:0-49,000,000
- chr19:15,000,000-17,000,000
- chr21:22,000,000-24,000,000
- chr22:25,000,000-27,000,000

Description générale du processus

- L'ensemble tourne sous Galaxy.
- La 1ère partie consiste en un **préprocessing/formatage** des données avec un tout petit peu de **QC** (Picard & stat on bam) (bam -> rmdup -> pileup). L'intégralité est réalisé par samtools.
- La 2e partie consiste au **processing des données** : fenêtrage -> normalisation GC + centralisation + segmentation et fusion -> visualisation. L'ensemble est effectué via 2 scripts, l'un en Perl5 et l'autre en R2.10+.
- Une 3e partie consistera en une présentation des annotations génomiques réalisables sur ces données, et une présentation de comparaison entre ces résultats de copy-number WGS et ceux issus de données CGHarray du même échantillon. Cette dernière partie nécessitant des scripts maison et une structure que nous n'avons que sur notre serveur à l'IGR, ce ne sera qu'une présentation.

1. 15h15 – 15h20 : Présentation, introduction, et explication du déroulement de la séance

2. 15h20 – 16h20 : Cours théorique (support majoritairement CGH, mais transposition WGS constamment à l'esprit).

3. 16h20 – 16h45 : TP partie 1 : Pré-processing des données
    1. Stats sur bam brut (2')
    2. Filtrage des réplicats PCR par rmdup (3') *bam -> bam*
    3. Stats sur BAM filtré (2')
    4. Génération du pileup (2'20") *bam -> pileup*

4. *16h45 – 17h15 : Pause café.*

1. 17h15 – 17h55 : TP partie 2 : Processing.
    1. Fenêtrage (script Perl5 : pupwm) (20' peut être lancé avant la pause café). *pileup -> pwm*
    2. Normalisation GC + segmentation + fusion + centralisation (R+limma+HaarSeg+aCGH : arcseg.R) (6'). *pwm -> segments + images*.
    3. Interprétation des profils et tableau de segments.

2. 17h55 – 18h15 : Démo/bilan.
    1. Démo au projo : comparaison aux données obtenues en CGHarray et annotation des régions variantes (compositions en gènes, annotations cancer, polymorphismes, sno/miRNA, CpGi).
    2. Bilan (apports / limitations) et mise en perspective (affinage du niveau de lecture, apport des SNP pour la LOH).