TD Détection de CNV

But du TD

Ce TD vise à familiariser les utilisateurs avec l'utilisation de la technique WGS à des fins d'identification et de quantification de variations de nombre de copies génomiques (CNV, copy number variations).

Description du set de données

- La source des données est un WGS Illumina sur 2.5 pistes ayant généré 900M reads totaux avant mapping (95% mappés).
- Les qualités globales de la librairie 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 15 autres autosomes, afin assurer la centralisation du profil (79Mb au total).
- L'ensemble totalise ~15M reads (1.5Mds nt).
- L'alignement de cette quantité de données pour le génome complet est bien trop long et massif pour le TP, et le fichier pré-aligné au format SAM reste trop lourd (5.2Go). Le fichier d'entrée du pipeline sera donc le fichier BAM après alignement par bwa, (paramètres par défaut).

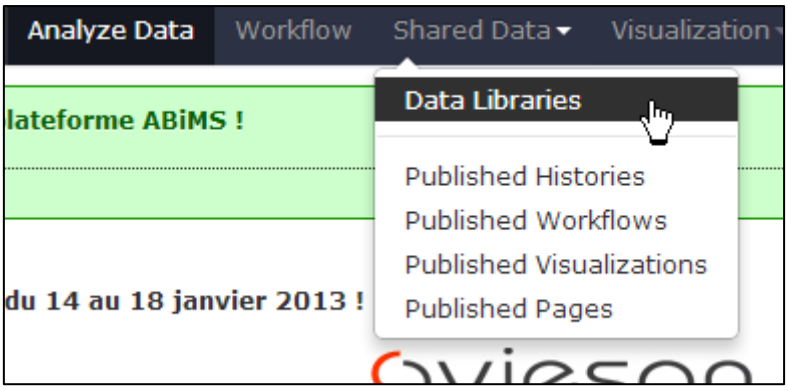
Coordonnées des régions sélectionnées.

Chromosome	Start	End	Width
chr1	40,000,000	42,000,000	2,000,000
chr2	153,000,000	155,000,000	2,000,000
chr3	48,000,000	50,000,000	2,000,000
chr5	110,000,000	112,000,000	2,000,000
chr7	20,000,000	22,000,000	2,000,000
chr8	75,000,000	77,000,000	2,000,000
chr10	75,000,000	77,000,000	2,000,000
chr11	75,000,000	77,000,000	2,000,000
chr12	75,000,000	77,000,000	2,000,000
chr13	48,000,000	50,000,000	2,000,000
chr14	75,000,000	77,000,000	2,000,000
chr16	50,000,000	52,000,000	2,000,000
chr17	0	49,000,000	49,000,000
chr19	15,000,000	17,000,000	2,000,000
chr21	22,000,000	24,000,000	2,000,000
chr22	25,000,000	27,000,000	2,000,000

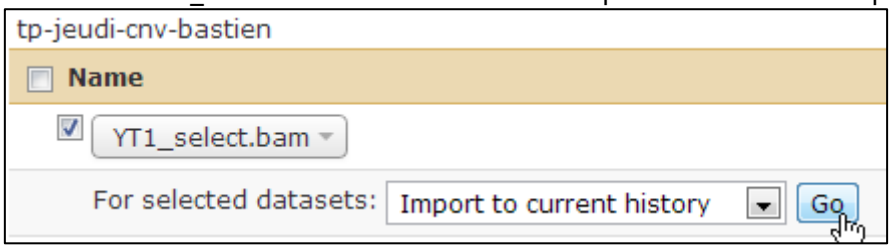
Partie I : Pré-processing des données

Importation du set de données

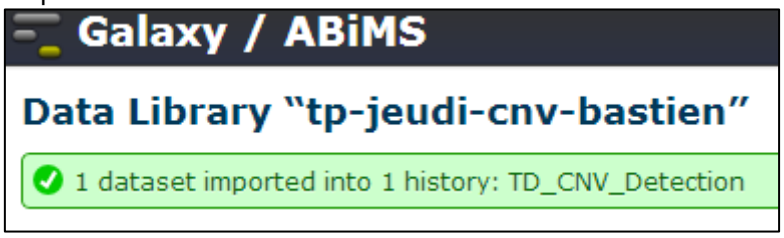
- Créer un nouvel historique, et le nommer (pour ce protocole, l'historique s'appellera « TD_CNV_Detection »).
- Récupérer le set de données : Shared Data => Data Libraries



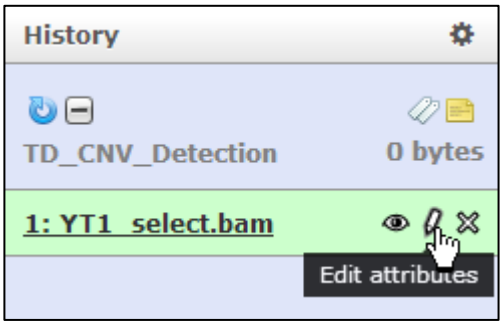
- Sélectionner : tp-jeudi-cnv-bastien.
- Sélectionner le fichier « YT1_select.bam » et choisir de l'importer dans l'historique en cours :

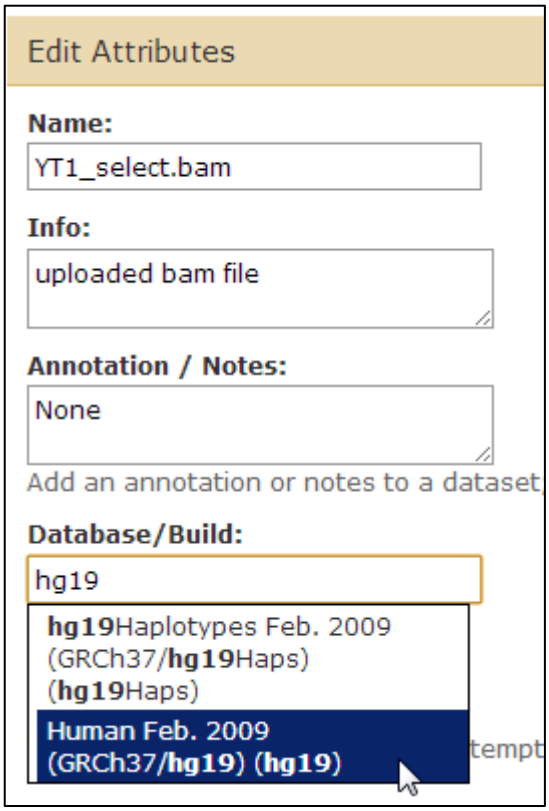


- Galaxy confirme l'import :



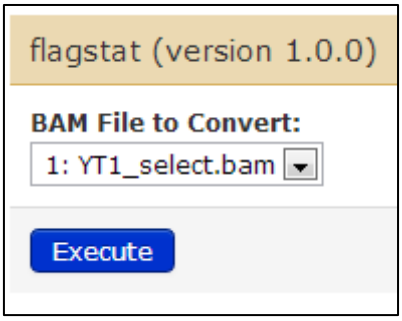
- Retourner à la zone d'analyse des données. Le nouveau set est disponible.
- Modifier les attributs du fichier et définir le génome de référence correspondant (GRCh37/hg19), et sauver :





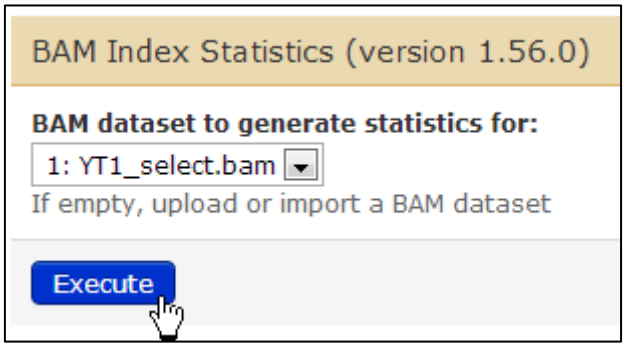
Statistiques sur le BAM source

- Faire tourner l'outil « **flagstat** » sur le BAM :



- Renommer le résultat en « flagstat on YT1_select ».

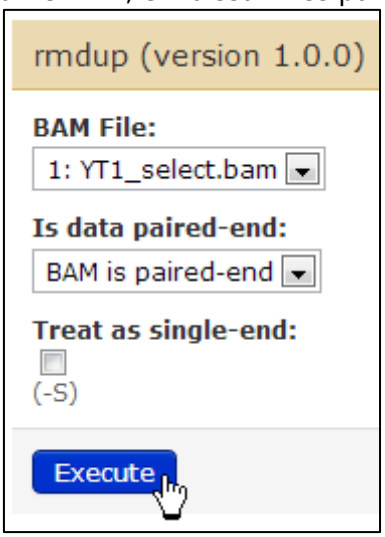
- Faire tourner l'outil « **BAM Index Statistics** » sur le BAM :



- Renommer le résultat en « BAM Index Statistics on YT1_select ».

Elimination des réplicats PCR

- Faire tourner l'outil « **rmdup** » sur le BAM, en laissant les paramètres par défaut) :



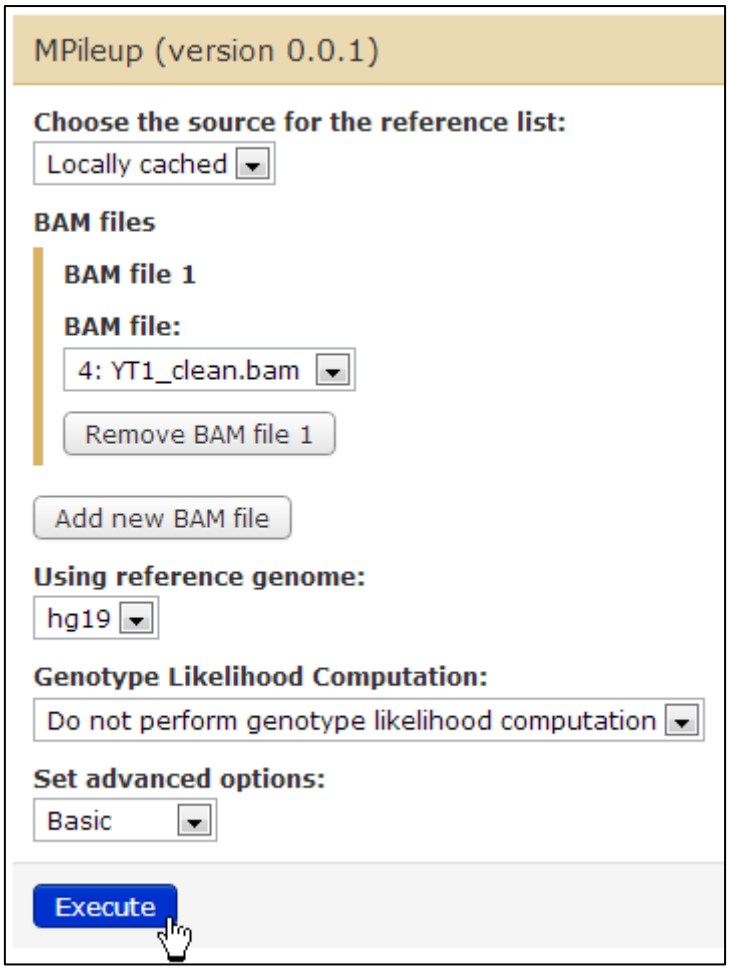
- Renommer le fichier BAM obtenu en « YT1_clean.bam ».

Statistiques sur le BAM filtré

- Reproduire les actions effectuées avec les outils « **flagstat** » et « **BAM Index Statistics** », mais sur le nouveau BAM filtré « YT1_clean.bam », et renommer les résultats en « flagstat on YT1_clean » et « BAM Index Statistics on YT1_clean », respectivement.
- Comparer les valeurs entre ces statistiques calculées sur le BAM d'origine et le BAM filtré.

Génération du fichier pileup

- Faire tourner l'outil « **Mpileup** » avec le BAM « YT1_clean.bam » et en laissant les paramètres par défaut.

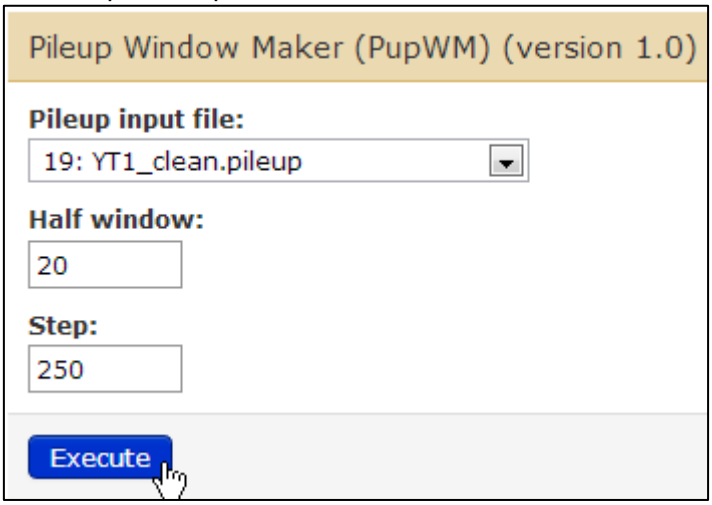


- *En cas d'erreur associée au génome de référence, récupérer le fichier « hg19.fa » depuis la library tp-jeudi-cnv-bastien vers votre historique. Relancer « **Mpileup** » en sélectionnant cette fois « History » pour le 1er paramètre, et choisir le fichier « hg19.fa » nouvellement récupéré.*
- Renommer le fichier pileup obtenu en « YT1_clean.pileup ».

Partie II : Fenêtrage, identification et quantification de CNV

Réduction des données par fenêtrage

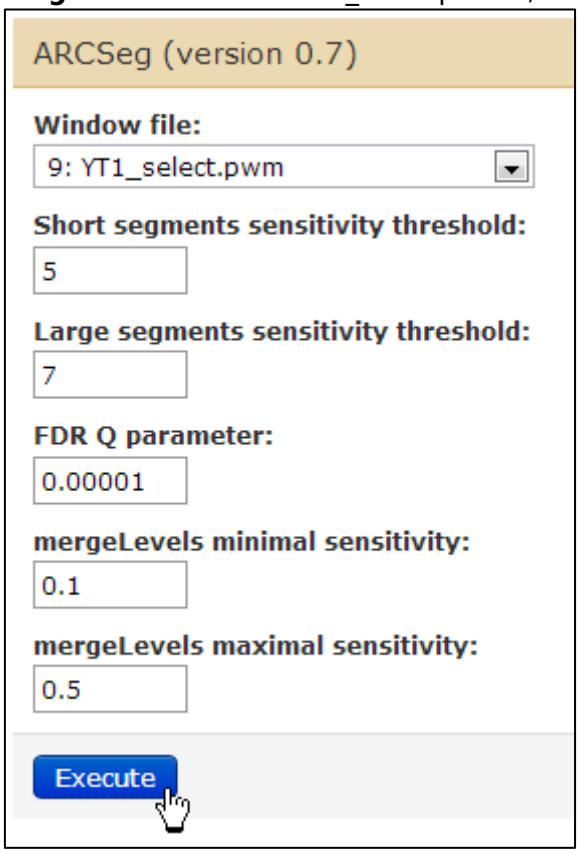
- Faire tourner l'outil « **Pileup Window Maker (PupWM)** ». Définir une demi-fenêtre (half-window) de **20** bases, et un pas (step) de **250** bases :



- Renommer le fichier obtenu en « YT1_clean.pwm ».

Normalisation, segmentation, centralisation et génération des fichiers de résultats

- Faire tourner l'outil « **ARCSeg** » sur le fichier « YT1_clean.pwm », avec ces paramètres :



- L'outil donne 5 fichiers de résultats : 5 graphes et 1 table.