

Table ronde « SI génomique »

Enjeux majeurs, Besoins, Solutions?

Contexte/constat

- Nombreuses données dispersées et spécifiques à des expérimentations (RNAseq, ChIPseq, ...) disponible comme des fichiers fastq, bam, vcf, gff, wig, ...
- Doivent pouvoir être comparée
 - entre elles,
 - avec des données utilisateurs privées,
 - à des données de références (disponible sur des sites UCSC genome browser, Ensembl, NCBI, ...)

Visualisation

- Données doivent pouvoir être visualisées "localement" --> IGV ou Jbrowse.
- Trouver les données pertinentes et y accéder.
 - Découverte et accès aux données
 - Solution: Indexation et recherche distribuée (SolR, Elastic search, ...)
 - Solution: Utilisation de TrackHub pour distribuer ces données qui peuvent difficilement être centralisées.
 - Problématique des métadonnées: condition expérimentales, outils et paramètres d'analyse
 - Solution: A régler à la soumission des données dans des repository. Des standards à préciser sans doute.
 - Problématique de mise à jour avec l'évolution des séquences génomiques et des annotations de références.
 - Solution: Prise en charge par des plateformes? Par les utilisateurs?

Calculs

- Données doivent pouvoir être comparée sur leurs recouvrement → mapper, bedtools, s-mart,
 - Problématique de disponibilité des données dans des infrastructure de calcul.
 - Données dispersées à rapatrier.
 - Solution: iRODS pour transfert des données.
 - Problématique d'espaces temporaires pour le calcul
 - Solution: Espaces partagés temporaire avec ménage régulier. Infrastructure de Cloud.