Compte Rendu Visio Conférence *PepiAnnot*

15 février 2022 -14h00 - 15h30

https://pepi-ibis.inrae.fr/annotation-genomes

Lors de la visio PepiAnnot :

- Il y avait 24 personnes connectées

Ordre du jour :

---

1. Animation PEPIAnnot

Martine Da Rocha, Jacques Lagnel et Véronique Brunaud

Veut-on faire 1 journée présentielle sur Annotation à la fin de l'année ? Pas d'enthousiasme débordant donc à priori non

2. Utilisation de REPET pour la détection d'éléments transposables : 2 intervenants Nathalie Choisne et Djampa Koslowski

2.1/ Nathalie Choisne présente REPET V3 et ses nouvelles fonctionnalités

3 pipelines indépendants :

- TEdenovo : détection et classification de TE de novo dans le génome, recherche les éléments répétés en plusieurs étapes à l'échelle du génome : découpe le génome, recherche à les aligner, au moins 3 séquences à aligner, HSP…clusterisation avec plusieurs outils
- PASTEC : recherche à classer les TE dans les familles de TE connus. Utilise des bases connues comme repbase (payante !) et GypsyDb. Recherche de séquences particulières, répétées (comme PolyA), banque rDNA. Utilise la classification de Wicker et al. Nature 2007 avec un code associé en trigramme R=retrotransposan, D dnatransposon
- TEannot : annotation plus détaillée. Aligner la librairie de consensus passée en entrée via repatMasker/blaster/censor/matcher. Recherche de SSR, merge toutes les annotations au niveau du génome. Sortie sous forme de fichier GFF3. Double passage de TEannot (plus sûr)

Conseils :

- Regarder les stats d'assemblage, petits outils pour faire les stats et contrôler
- Généralement on fait TEdenovo avec une partie de génome seulement, par exemple 300Mo peut être suffisant
- On ne prend que les consensus > 95% du consensus de départ ou pleine longueur

Changements entre v2.5 et v3 dans PASTEC : des changements dans les noms de consensus qui contenaient le trigramme Wicker + les outils ou méthodes utilisées + minimum de séquences gardées (map20) → le code Wicker a été retiré. Nouveau fichier de classifications contenant une classification aussi pour les virus, et ajout de la super-famille.

REPET est :

- Distribué sur plusieurs PF Bioinfo
- Image Docker existe
- Une V3.1 devrait arriver
- Y a un groupe de travail sur snackmake-workflow pour mettre TEannot dans un workflow

2.2/ Djampa Kozlowki : développement d'un outil de post-processing de REPET, REP3T-pal

3 modules pour prédire les ET :

- Genome-stats pour avoir des stats sur son génome
- Genome-format créer une version modifiée du génome, en ne retenant qu'un % du genome significatif en entrée de REPET, génome moins fragmenté, oter les petits scaffolds.
- Module Annot : filtre les outputs de REPET « post-processing » qui sont nombreux pour traduire en stats et graphiques plus simples à examiner rapidement.

Application à M. incognita V3 de 183Mb : le but étant de rendre l'outil le plus automatique possible même sur les données finales à récupérer.

Génome encore un peu trop fragmenté donc on réduit la fragmentation et on retire les plus petits scaffolds, taille environ 5kb. Donc à la fin de ce nettoyage, on perd 10% du génome.

Partie Annot sur la filtration des données : 8 filtres modulaires en python 3, stats et graphes pour chaque filtre avec choix de ce qu'on veut garder

- TE plus sûr donc exclure les match partiels et les annotations inconnues
- L'information trigramme de la classification est gardée car permet d'avoir déjà la famille
- Filtrer les annotations avec une long min/max
- On peut filtrer sur le % d'identité du consensus et la proportion du consensus conservée
- Possibilité d'ôter toutes les annotations chevauchantes
- Récupérer la séquence ET annotée et aller rechercher par blast réciproque si les séquences consensus et source (parent consensus) correspondent

A la fin sont générés des stats et des graphes

- Conclusion pour M incognita : c'est mieux que si on utilise REPET mais de façon brut, automatique et avec les défauts. Génère des graphes et stats beaucoup plus facile à interpréter.
- Par contre il y des filtres assez strict donc voir si besoin d'ajuster
- Disponible sur github privé à Djampa pour l'instant, mais Djampa est prêt à distribué ses scripts pour que quelqu'un prenne le
- Très complémentaire de REPET
- Discussion sur l'importance du coté graphique pour résumer les filtres et informations

3. Thèmes à aborder

- Assemblage genome/transcriptome (Erwan Corre)
- Assemblage **Transcriptome *de novo* short et Long reads** - *BRUNAUD Véronique*
- **DGenies** - DotPlot de chromosomes entiers - *KLOPP Christophe (à contacter)*
- Annotation fonctionnelle - Mercator4/MapMan - *DELANNOY Etienne (à contacter)*
- Annotation fonctionnelle avec Eggnog (image singlarity)
- Pan Génomique - *JOETS Johann*
- Génomique comparée
- Épigénomique
-

Plus général concernant le PEPI-IBIS

- **R shiny & R markdown** - outils puissants pour l'analyse et la traçabilité en bioinformatique
- SnackMake vs NextFlow : comparaison et qui utilise quoi …
- Intégration des données omiques - *ALAUX Michael (à contacter)*
- Intégration statistique des données omiques - *MARTIN MAGNIETTE Marie-Laure (à contacter)*
-

4. Prochaine réunion PepiAnnot

- **Avril/mai 2022 à fixer**
- **Transcriptome de novo Long-read – short reads Véro et Jacques ?**

5. Photo du Jour

A quelle heure la répétition ?