CR réunion PEPI-Annot PEPI-IBIS et GT-Annot de MERIT

19 juin 2025

Questions sur les personnes présentes

- 42 personnes présentes
- 35 personnes ont répondu /42
- Origine bio (2/35 4%) bioinfo (33/35 94%)
- Instituts : CNRS (14/35 40%) INRAE (13/35 37%) INSERM (1/35 3%) universités (6/35 17%)

Prochaines présentations possibles :

Ne pas hésitez à proposer des thèmes
- Comparaison d'outils d'annotation entre Helixer, et l'outil
- Annotation des gènes non-codants
- Annotations fonctionnelles

→ Sandrine Caburet (université Paris Cité) propose de parler de l'annotation chez l'humain (voir si date possible en sept)

Présentation de Sylvain Foissac (INRAE GenPhySE,Toulouse) : retour d'expérience utilisation de TAGADA : Annotation génomes animaux

Complexité de l'annotation des génomes animaux et d'avoir un génome de référence

Complexité à plusieurs niveaux pour les définitions de gènes, de transcrits et une grande différence peut exister entre les sources d'annotations.

Par exemple sur la proportion de non-codant entre différentes sources et différentes versions du génome. De plus suivant les sources, les ID des gènes ne sont pas toujours les mêmes (NCBI/Ensembl)

Évaluation de la qualité des annotations

Sylvain présente un outil appelé Dragibus (https://github.com/cguyomar/dragibus) pour évaluer la qualité des annotations de référence, en soulignant les variations importantes entre différentes versions et espèces. Il discute des critères de qualité tels que la proportion de transcrits mono-exoniques, la présence de signaux de polyadénylation et la conservation des sites d'épissage. Erwan apporte des nuances sur la conservation des sites d'épissage chez certaines espèces non vertébrées. Sylvain conclut en présentant un pipeline d'annotation qui utilise des données RNA-seq pour améliorer les annotations existantes et effectuer une quantification des gènes et des transcrits.

TAGADA Pipeline

Outil open source pour prédire les structures de gènes et transcrits potentiels (https://github.com/FAANG/analysis-TAGADA). TAGADA produit un assemblage de transcrits par tissu pour éviter la création de transcrits chimériques. TAGADA génère de nombreux rapports pour évaluer la qualité d'annotation ainsi que les options de commande pour la quantification et l'assemblage.

Améliorations envisagées

- Utilisation de l'outil Tmerge (https://github.com/guigolab/tmerge) pour traiter la fusion des transcrits
- Tenir compte du nombre de reads couvrant le transcrit mais aussi trouvé dans plusieurs échantillons
- Gestion à prévoir des transcrits circulaires.
- Importance de considérer la structure intronique lors de la fusion des transcrits.

Conclusion

Le pipeline TAGADA intègre des annotations de référence tout en permettant l'extension des gènes connus, et des comparaisons avec les annotations historiques.

Référence/Lien :

Kurylo C, Guyomar C, Foissac S, Djebali S. TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data. NAR Genomics and Bioinformatics. 2023

https://github.com/FAANG/analysis-TAGADA