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

Du 17 mars 2025

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

→ Sylvain Foissac propose un retour d'expérience sur l'outil TAGADA (voir les dates pour juin)

Thème : Annotations structurales des génomes eucaryotes, « Peut mieux faire ! »

Veronique Brunnaud : Annotation Structurale d'Arabidopsis

Véronique présente un retour d'expérience sur l'annotation structurelle d'Arabidopsis thaliana, la première plante séquencée en 2000. Elle explique que son équipe travaille sur la régulation de la transcription, et la detection de motifs cis-regulateurs via une méthode in silico appelée PLMdetect. Cette méthode permet de détecter des DNA motifs enrichis dans les régions proximales des gènes (500 bases autour des gènes). Véronique souligne l'importance des régions proximales pour les sites de fixation des facteurs de transcription (TFBS) et détaille le fonctionnement de PLMdetect, qui permet d'identifier des motifs préférentiellement localisés par rapport au début ou à la fin des gènes. Ainsi un PLM pour Preferentially Located Motif est caractérisé par sa séquence et sa fenêtre préférentielle correspondant à la zone d'enrichissement du motif.

Différences dans L'annotation.

Véronique présente les différences entre les annotations TAIR10 et Araport11 pour Arabidopsis thaliana, en soulignant les problèmes liés à la définition des extrémités des gènes dans Araport11. Elle explique que TAIR10 semble plus précis pour les extrémités des gènes, ceci ayant été prouvé par des données expérimentales (2 articles), ce qui est crucial pour leur travail sur les motifs PLM. Pour résoudre ce problème, Margot a combiné les gènes de TAIR10 avec les nouveaux gènes d'Araport11. Véronique mentionne également des efforts en cours pour améliorer l'annotation, via des données RNA-seq (short-reads et long-reads) et la possibilité d'utiliser l'outil Helixer basé sur l'IA. Elle conclut en soulignant l'importance de choisir l'annotation appropriée selon la problématique de recherche.

Discussion sur la présentation

Sylvain soulève la question des transcrits alternatifs et des TSS multiples par gène, suggérant de considérer tous les TSS sans se limiter à un seul par gène. Véronique reconnaît que l'approche pourrait être étendue à d'autres organismes avec plus d'épissage alternatif. La discussion se poursuit sur différents outils pour fusionner et analyser les annotations de gènes et de transcrits, notamment TAMA, Cufflinks, TAGADA et Tmerge. Les participants partagent leurs expériences avec ces outils et discutent de leurs avantages et limitations.

Sophie Lemoine : Annotation Des Génomes Non Modèles

Sophie présente les défis rencontrés dans l'annotation structurelle des génomes d'espèces non modèles, en particulier pour les projets de séquençage d'ARN en cellule unique. Elle compare différents outils d'annotation comme StringTie2, PASA, IsoQuant et RNABloom, soulignant leurs forces et faiblesses. StringTie2 a tendance à reproduire l'annotation de référence quand elle est en donnée d'entrée, il a tendance à fusionner les gènes ou à étendre les UTRs sans se baser sur les données de séquençage quand les annotations sont absentes. PASA est très lourd a faire tourner et les résultats sont décevants. Il reproduit l'annotation en entrée, ne trouve pas les gènes pourtant présents dans les données expérimentales. IsoQuant est incomplet mais si on l'utilise en avec le paramétrage ONT par défaut, il est plutôt bon même s'il loupe des transcrits pourtant

Commenté [1]: a enrichir

bien représentés dans les données de séquençage. RNABloom annote presque tout mais renvoie beaucoup d’isoformes qui peuvent être compliqués à gérer pour une analyse au niveau des transcrits. La difficulté est de combiner les résultats de différents outils et de les comparer à une annotation de référence. Elle n’a pas trouvé de solution correcte dans des outils classiques qui ont tendance à raccourcir les nouvelles annotations dès que l’ancienne rentre en jeu. Elle propose une approche combinant IsoQuant et RNABloom, en n’introduisant l'annotation de référence qu'à la fin du processus pour éviter les biais. Sophie conclut en présentant un pipeline en développement appelé Egzotek, dédié à l’annotation des espèces non-modèles dites exotiques, et souligne l'importance d'utiliser les lectures longues et de minimiser l'influence de l'annotation de référence à corriger dans le processus d'annotation.

Annotation Génomique et Contrôle Qualité.

Sophie présente les perspectives de son équipe concernant l'annotation génomique. Ils souhaitent tester Helixer sur les insectes et ré-entraîner le modèle invertébré qui semble moins performant selon les discussions de la dernière réunion MERIT. Ils prévoient également d'ajouter des outils de contrôle qualité sur Egzotek, notamment pour la validation des GFF et l'annotation fonctionnelle. La discussion s'oriente ensuite sur la problématique du partage et de la découvrabilité des annotations améliorées, avec des suggestions pour les rendre plus accessibles via des data papers ou des entrepôts de données (data.gouv).