Compte Rendu Visio Conférence *PepiAnnot*

12 Mars 2021 -10h00 / 11h30

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

Membres (33)	Unité	Mail
AMSELEM Joëlle	URGI, INRAE, Versailles	joelle.amselem@inrae.fr
BOISARD Julie	MNHN	julie.boisard@edu.mnhn.fr
BOUDET Nathalie	IPS2, MdC UEVE, Gif sur Yvette	nathalie.boudet@inrae.fr
BRIONNE Aurélien	INRAE, Tours	aurelien.brionne@inrae.fr
BRUNAUD Véronique	IPS2, Gif sur Yvette	veronique.brunaud@u-psud.fr
CANAGUIER Aurélie	INRAE, EPGV, Evry	aurelie.canaguier@inrae.fr
CHARLES Mathieu	INRAE, GABI,Jouy-en-Josas	Mathieu.Charles@inrae.fr
CHOULET Frédéric	INRAE, GDEC-UCA, Clermont-Ferrand	frederic.choulet@inrae.fr
CORRE Erwan	CNRS Roscoff	corre@sb-roscoff.fr
DA-ROCHA Martine	INRAE, Sophia Agrobiotech, Antibes	martine.da-rocha@inrae.fr
DERRIEN Thomas	IGDR - CNRS - UMR6290, Rennes	thomas.derrien@univ-rennes1.fr
DEVILLIERS Hugo	INRAE, Micalis, Jouy-en-Josas	hugo.devillers@inrae.fr
FAIVRE RAMPANT Patricia	INRAE, EPGV, Evry	patricia.faivre-rampant@inrae.fr
HILLIOU Frédérique	INRAE, Sophia Agrobiotech, Antibes	frederique.hilliou@inrae.fr
HINSINGER Damien	INRAE,EPGV,Evry	damien.hinsinger@inrae.fr
HUNEAU Cécile	INRAE, GDEC-UCA, Clermont-Ferrand	cecile.huneau@inrae.fr
JOETS Johann	INRAE, Moulon, Orsay	johanne.joets@inrae.fr
KORNOBIS Etienne	Pasteur,	etienne.kornobis@pasteur.fr
KREPLAK Jonathan	INRAE, Dijon	jonathan.kreplak@inrae.fr
LASSERRE-ZUBER Pauline	INRAE, GDEC-UCA, Clermont-Ferrand	pauline.lasserre-zuber@inrae.fr
LE DANTEC Loïc	INRA Bordeaux	loick.le-dantec@inrae.fr
LEGEAI Fabrice	INRA, BIPAA, Rennes	Fabrice.Legeai@inrae.fr
LEROY Philippe	INRA GDEC-UCA, Clermont-Ferrand	philippe.leroy.2@inrae.fr
MARDOC Emile	INRAE, GDEC-UCA, Clermont-Ferrand	emile.mardoc@inrae.fr
MOLLION Maeva	INRA, Moulon, Orsay	Maeva.Mollion@inrae.fr
NEUVEGLISE Cécile	INRAE, Micalis, Jouy-en-Josas	cecile.neuveglise@inrae.fr
ORJUELA Julie	IRD, Montpellier	julie.orjuela@ird.fr
PALLIER Vincent	INRAE GDEC-UCA, Clermont-Ferrand	vincent.pailler@inrae.fr
RIMBERT Hélène	INRAE GDEC-UCA, Clermont-Ferrand	helene.rimbert@inrae.fr
ROGIER Odile	INRA Orléans	odile.rogier@inrae.fr
SIMON Adeline	INRAE, Versailles	adeline.simon@inrae.fr
TOFFANO-NIOCHE Claire	I2BC, CNRS, Gif-sur-Yvette	claire.toffano-nioche@u-psud.fr
VELT Amandine	INRAE, Colmar	amandine.velt@inrae.fr

Si des personnes manquent dans la liste ne pas hésiter à contacter Véronique ou Philippe pour une mise à jour.
Si besoin compléter et/ou corriger le tableau ci-dessus. Merci par avance.

Lors de la visio PepiAnnot du **12 Mars 2021** :

- Il y avait 18 personnes connectées
- Visioconférence avec GoToMeeting (licence PEPI IBIS). Cette Visioconférence a été enregistrée.
- Une nouvelle personne dans le groupe c'est Damien Hinsinger le responsable de la plateforme en génomique (damien.hinsinger@inrae.fr) de l'EPGV (Evry) avec Patricia Faivre Rampant et Aurélie Canaguier.

Ordre du jour :

1. Prochains thèmes possibles

- **Kmer pour l'annotation** - *CHÂTEAU Annie (à contacter)*
- **R shiny & R markdown** - outils puissants pour l'analyse et la traçabilité en bioinformatique
- **SibeliaZ** - *JOETS Johann*
- Annotation des **TEs** - *CHOULET Frédéric*
- 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)*
- Intégration des données omiques - *ALAUX Michael (à contacter)*
- Réflexions autour de l'intégration statistique des données omiques - *MARTIN MAGNIETTE Marie-Laure (à contacter)*
- Pan Génomique
- Variants structuraux
- Réseaux de gènes
- Exposé didactique de la théorie des graphes
- Les infrastructures de calcul et de stockage - bonnes pratiques
- Plan de gestion des data avant tout projet !
- États de l'art sur tous les éléments constitutifs connus à ce jour (features) d'un génome
- Apport du « Deep Learning » sur l'analyse des données omiques

2. Organisation d'une réunion du PEPI IBIS, 2 demi-journées

PepiAnnot propose l'organisation d'une réunion plutôt en Novembre 2021 pour avoir une chance en présentiel !

Choix de conférence coté PepiAnnot :

- Philippe L. propose une conférence sur le développement d'une nouvelle séquence de référence du génome du blé tendre (variété Renan) dans le cadre du projet wheatOMICs avec le Génoscope et FranceGénomique (*Fred Choulet, Pauline, Hélène & Philippe*). Optimisation de TriAnnot pour annoter un génome plus rapidement (Collaboration avec le Mésocentre de Clermont-Ferrand - David Grimbichler)
- Véro B. propose une conférence sur les "*Stats intégration multi-omics*" via Guillem Rigaill ou ML Martin-Magniette.

3. Prochaine réunion PepiAnnot

- **Fin Mai - début Juin 2021 - 10h00 - 11h30**
- **Appel à candidature ! ☺ Veuillez vous manifester d'ici la mi-Mai que l'on puisse organiser la visio. Merci par avance ….**

4. Johann JOETS, INRAE le Moulon, Orsay – *Liftoff* - Hélène Rimbert, INRAE GDEC, Clermont-Ferrand – *Magatt*

Objectifs :

- Prendre l'annotation d'un génome de référence et la transférer sur un nouveau génome proche (même espèce avec une autre variété ou autre espèce proche)
- En effet, de plus en plus de génomes de plantes sont séquencés et assemblés, et de plus en plus rapidement. L'annotation structurale et fonctionnelle de ces génomes pose un problème méthodologique pour des raisons de temps d'analyse et de transfert d'annotation entre divers génomes et les séquences de références internationales publiées. Des alternatives à l'annotation *de novo* complète sont envisagées.

Liftoff - Johann JOETS

- https://www.biorxiv.org/content/10.1101/2020.06.24.169680v1
- Liftoff est développé dans le groupe de Salzberg
  - Cherche les exons équivalents
  - Cherche les gènes identiques, si 1 gène de référence = 1 gène du nouveau génome
  - Marche pour des espèces proches (genre homme et chimpanzé)

1. Alignement via *Minimap2* de l'annotation du génome de référence contre le nouveau génome
   - Liftoff prend en entrée un GTF et un Fasta, il crée une base de données Sqlite3
   - *Minimap2* (on peut modifier les paramètres) permet un mapping avec des erreurs/différences (utilisé par exemple pour les long reads qui contiennent 10 à 20% d'erreurs)
   - Alignement des gènes de la séquence de référence contre la nouvelle séquence
   - *Minimap2* peut produire plusieurs alignements : full alignement tout est correct, sinon liftoff fait un acyclic graph pour choisir meilleur alignement car différences entres des exons
   - Prend en compte le brin et les distances entre exons pour prendre les décisions
   - Bonne réponse : chemin le plus court dans le graphe
2. Néanmoins, problème avec les alignements de famille de gènes
   - Prend en compte le score d'alignement pour choisir
   - Recherche aussi des alignements supplémentaires (secondaires) : même copie mais aucun autre gène ne doit correspondre à cette région, donc propose le 2ème emplacement, dans le cas de paralogues, mais cette annotation n'était pas marquée dans la référence
3. Attention, avec des familles de gènes et des paralogues équilibrés, l'outil liftoff peut se mélanger, car impossible de choisir !
4. Liftoff génère le GFF/GTF de cette annotation sur le nouveau génome.
5. Liftoff a été testé sur un nouvel assemblage du blé tendre (*transfert des annotations précédentes sur la nouvelle version*) selon les auteurs les résultats sont satisfaisants et Liftoff a trouvé une bonne quantité de nouveaux gène.
6. Exemple réalisé par Johann : transfert de l'annotation de la version 5 vers la version 4 du génome de référence de maïs B73.
   - La majorité des gènes sont transférés
   - La majorité de liens sont uniques (1 to1)
   - Seulement 8 440 "one to many" et 475 "unmapped"
   - Vérification si le mapping est au bon endroit: majorité bon et 3,8% des gènes non retrouvés au bon endroit 1 928 sur 49 582 gènes étudiés
   - 475 nouveaux gènes dans la v5 du maïs par rapport à la v4 : nouveaux gènes vraiment ? en fait ces gènes sont écartés car déjà un autre gène associé au même endroit, donc on retrouve le problème des régions dupliquées ou avec plusieurs annotations qui se superposent.
7. La performance est le point fort de Liftoff (sur le maïs) - transfert d'annotation en 30 min avec une machine de 80 cœur et 1,5 To de RAM : donc très performant.
8. Liftoff est très facile à télécharger, à installer et à utiliser
9. Si même assemblage/séquence possibilité de comparer les GFF avec gffcompare

- Discussion
  - Application pour aller trouver les gènes additionnels dans différentes variétés, utilisation de Liftoff pour générer une 1ère annotation structurale
  - Dans l'article il y une échelle de différence entre distance entre génomes comparés

Magatt - Hélène RIMBER

- https://forgemia.inra.fr/umr-gdec/magatt
- Magatt a été créé pour transférer l'annotation structurale de la version v1.0 de la séquence de référence internationale IWGSC de la variété de blé tendre Chinese Spring publiée en 2018 dans Science sur la version v2.1 de cette même variété mais avec un nouvel assemblage en cours de publication dans Plant Journal
- La particularité de Magatt est que le programme s'appuie sur des marqueurs génomiques ISBP (pattern TEs - Transposable Elements) qui sont utilisés pour guider le transfert des gènes. En effet, l'idée est de se servir de markers spécifiques d'une région pour positionner cette nouvelle annotation, basée sur le fait que les patterns d'éléments transposables sont conservés entre variétés de blé
    1. Chercher 150 bp autour d'une jonction TE-TE unique dans le génome (ISBP) : vérifier l'unicité dans le génome
    2. Ensuite BWA est utiliser pour placer ces marqueurs dans le nouveau génome (autoriser des mismatches ou pas en fonction du génotype cible)
    3. Identifier les couples de markers ISBPs qui cadrent les gènes et rechercher ces couples de markers sur le génome cible
    4. BLAT de la « feature gene » dans la région délimité par les ISBPs flanquants
    5. Gmap est également utilisé pour ancrer les gènes pour lesquels la méthode d'ancrage avec ISBPs a échoué (pas de couple de markers retrouvés sur le génome cible, ou pas de hit BLAT dans la région théorique)
- Magatt a été amélioré depuis pour le transfert des gènes de la RefSeq IWGSC v1.0 sur de nouveaux génomes de blé
    - Programme 10+genome (http://www.10wheatgenomes.com/)
    - Une nouvelle séquence de référence développée dans le cadre du programme wheatOMICs (Génoscope / FranceGenomics)
    - D'autres génomes à venir et notamment des génomes tétraploïdes et ou diploïdes si possibles
- Exemple avec le transfert de l'annotation de la RefSeqIWGSCv1.0 sur la nouvelle RefSeqIWGSCv2.1
    - 99% des gènes v1.0 sont transférés sur la v2.1 (106 913 HC gene & 159 840 LC gene)
    - 1 258 gènes non positionnés dans la v1.0 (chromosome unknown) on été intégrés à la v2.1
    - 2 792 gènes (1%) restent non ancrés à ce jour …
    - En résumé, sur les 269 545 gènes V1.0 sont transférés sur la v2.1
        - 241 201 (100% id/cov)
        - 7 895 avec un mismatch
        - 15 422 (G1) + 2235 (G2) mappé finalement avec Gmap
        - 2 792 unmap
    - Hélène a observé que 2 552 CDS étaient modifiés après le transfert dû notamment au "gap filling" avec des contigs pacbio entre la v1.0 et la v2.0 (erreurs résiduelles dans les contigs intégrés au niveau des gaps comblés)
    - L'équipe BioInfo du GDEC n'utilise pas beaucoup Busco, quelqu'un fait remarquer que l'on peut utiliser Busco avec la possibilité de ne pas regarder que les gènes conservés, mais tous les gènes d'une espèce

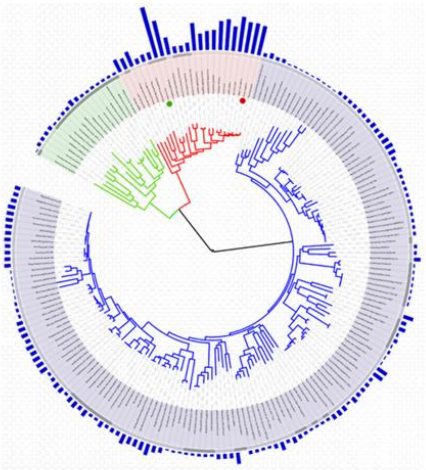
- Magatt a été réalisé en snakemake avec un fichier d'environnement via conda (yml). Un fichier de config de snakmake pour mettre les inputs. Pas mal de dépendance. Donc installation via conda
- Magatt a besoin en input (Génome de référence) : le Fasta, le GFF3, le bed des ISBP et les mRNA pour le génome cible: le bam de l'alignement des ISBP, le fasta de nouvel assemblage et un index Gmap. En sortie, Magatt donne un GFF et le tableau de mapping avec les marqueurs associés

- En terme de performance l'analyse prend 12hrs au lieu de 2hrs avec Liftoff
- Magatt est portable sur un cluster de calcul et installé actuellement sur le cluster (hpc2) du Mésocentre de Clermont-Ferrand
- Remarque : Magatt devrait en principe mieux gérer les familles de gènes sauf dans le cas où il y a des chevauchements car cela pause un problème puisque Magatt cherche des marqueurs ISBP délimitant une région génomique.
- La limite de Magatt est qu'il est conçu pour des génomes pour lesquels on peut annoter des ISBP (ce qui est le cas des Triticeae) à moins de trouver d'autres type de marqueurs conservés entre génotypes ?
- Magatt est actuellement utilisé au GDEC pour l'annotation de la séquence de référence de la variété Renan (wheatOMICs) en parallèle avec TriAnnot (pour les gènes spécifiques à Renan)

Quelques remarques en fin de visioconférence sur Singularity/Dockers vs nextflow/snakemake

- Nextflow prend le dessus en ce moment
- Plutôt singularity que dockers
- Hackaton Reproductibility inter-cati en Novembre 2020, donc accessible - Information Martine Da-Rocha : https://inter_cati_omics.pages.mia.inra.fr/reproductibility/

5. Photo du Jour



https://fr.wikipedia.org/wiki/Taille_du_g%C3%A9nome#/media/Fichier:Tree_of_life_with_genome_size.svg