Compte Rendu Visio Conférence *PepiAnnot*

2 Octobre 2020 -10h00 / 11h30

https://rendez-vous.renater.fr

Membres (28)	Unité	Mail
AMSELEM Joëlle	URGI, INRAE, Versailles	joelle.amselem@inrae.fr
BOISARD Julie		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	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
DIOT Thomas		thomas69.diot@laposte.net
FAIVRE RAMPANT Patricia	INRAE,	patricia.faivre-rampant@inrae.fr
HILLIOU Frédérique	INRAE,	frederique.hilliou@inrae.fr
JOETS Johann	INRAE, Fermes du 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
MOLLION Maeva	INRA, Fermes du Moulon, Orsay	Maeva.Mollion@inrae.fr
MONAT Cécile	INRAE, GDEC-UCA, Clermont-Ferrand	cecile.monat@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 **2 Octobre 2020** :

- Il y avait entre 18 et 15 personnes connectées
- Pauline, Cécile et Vincent de INRAE GDEC (Clermont-Ferrand) ont été rajouté.es au Groupe PepiAnnot (mailing list)
- Visio avec rendez-vous - parfait.

Ordre du jour :

1. Discussion sur les séminaires à prévoir et fait-on quelque chose en semi-présentiel et plus long ?....2

2. Prochains thèmes possibles à aborder....2

3. Média/outils Visio....3

4. Aurélie CANAGUIER INRAE – *Long DNA technologies evaluation for plant structural variations detection*....3

5. Photo du Jour....5

1. Discussion sur les séminaires à prévoir et fait-on quelque chose en semi-présentiel et plus long ?

- Vu qu'on ne peut pas faire de pauses café et qu'un repas ensemble sera impossible, aucun intérêt de faire du présentiel même par Centre, car obligé aussi de porter les masques !
- Tout le monde s'accorde à penser qu'une seule présentation, sur un thème à la fois, est préférable car cela laisse plus de temps pour les questions et les échanges. Néanmoins, on peut tout à fait regrouper 2 présentations sur un même thème (1h30 à 2h max si 2 présentations)

⇨ Donc on reste sur le même système que les visio PepiAnnot et il est possible d'augmenter la fréquence : une réunion tous les 2 mois par exemple.

⇨ Attention : les PEPIs vont changer de structure mais cela ne devrait pas impacter PepiAnnot ?

Remarque : Concernant l'ouverture des visios PepiAnnot à tout bioinfo de INRAE.
Si nous sommes trop nombreux en visio, cela va compliquer les questions et les échanges, et on veut que ça reste des groupes de travail avec beaucoup d'échanges. De plus, certaines présentations sont des retours d'expériences non publiés, donc pas toujours à diffuser à toute la communauté. Cela n'empêche aucunement d'inviter d'autres collègues et de faire chacun.e de notre côté la promotion de PepiAnnot. Le Groupe peut donc s'agrandir ….

2. Prochains thèmes possibles à aborder

- **Rfam** - TOFFANO-NIOCHE Claire
- **SibeliaZ** - *JOETS Johann*
- **liftoff** (?) - *JOETS Johann* & **TransfertAnnot** (?) *RIMBERT Hélène* sur le transfert d'annotation à partir d'une séquence de référence sur un nouveau génome assemblé
- Annotation des **TEs** - *CHOULET Frédéric*
- Assemblage **Transcriptome *de novo* short et Long reads** - *BRUNAUD Véronique*
- **DGenies** - DotPlot de chromosomes entiers - *KLOPP Christophe (à demander)*
- Annotation fonctionnelle - Mercator4/MapMan - *DELANNOY Etienne (à demander)*
- Intégration des données omiques - *ALAUX Michael (à demander)*
- Réflexions autour de l'intégration statistique des données omiques - *MARTIN MAGNIETTE Marie-Laure (à demander)*
- 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

➔ Prochaine visio PepiAnnot début Novembre de 10h00 à 11h30. **Qui / Quoi ?** ☺

3. Média/outils Visio

- Nous utilisons jusqu'à présent « Rendez-Vous »
- De plus en plus d'outils voient le jour …
    - À noter **GoToMeeting** qui semble très bien et stable. Véronique voit avec la Dipso la possibilité d'acquérir une licence pour le PEPI IBIS que l'on pourra exploiter pour le Groupe PepiAnnot
    - Il y a également **BigBlueButton** de plus en plus utilisés par certaines communautés (MIAT ; SupAgro) - aucun lag/freeze avec 112 connexion (source Hélène Rimbert). Tests au MIAT (Vidéos/Micros sur plus de 30 connexions avec une qualité exceptionnelle). Ne pas utiliser, par contre, la version gratuite en ligne. Il faut adosser une instance de BigBlueButton sur un serveur (8cpu/32G RAM). Attention investissement important d'informatique (à vérifier).

4. Aurélie CANAGUIER INRAE – *Long DNA technologies evaluation for plant structural variations detection*

Objectifs :

- Intérêt des variations structurales (SV - Structural Variation) pour construire un pan-génome - impacte des SV sur le génotypage et sur l'évolution en général
- Comparaison de stratégies : assemblage ONT (Oxford Nanopore) vs Bionano (Cartes Optiques)

Éléments de contexte

- 2 écotypes d'*Arabidopsis thaliana* landsberg ***Ler-1*** et Colombia ***Col-0***
- Couvertures dans les 2 technos, trimming via CANU ou outils bionano, SMARTdenovo pour l'assemblage
- MuMmer/Bionano pour trouver les SV > 1kb
- bedtools pour comparer les chevauchements entre SV
- Génome de référence ***Col-0*** TAIR10 et une version Landsberg de 2016 (Zapata *et al*. 2016)

Comparaison des 2 technos chez *Ler-1*

- INS et DEL sont majoritaires avec ONT et avec Bionano
- Détection de 43% de SVs < 1kb avec ONT et 25% avec Bionano
- 2 fois moins de SVs<1kb en Bionano par rapport à ONT
- ~30% des SVs ONT ont une correspondance 1:1 avec un SV Bionano, ce qui correspond à ~65% des SVs Bionano (408 SVs)

Caractériser les différences entre les 2 techno chez Ler-1

- 20% des SVs vues en ONT ne le sont pas en Bionano (même à 1 base prêt)
- 5% des SVs vues en Bionano ne le sont pas en ONT (même à 1 base prêt)
- 529 SVs ONTs (~50% des SVs ONT) sont chevauchants avec 151 SVs Bionano (~30% des SVs Bionano)
- La détection à partir des assemblages ONT donne une image plus fragmentées des variations structurales.

Différence avec publication de Zapata *et al* 2016

- Prise en compte que des SV > 10 bases dans la publication de Zapata donc 2 656 SVs
- 67% des 1 185 SVs ONT, (85% de la taille cumulée) chevauchent 52% des 2 656 SVs de Zapata (72% de la taille cumulée)
    - ⇨ Donc la différence vient surtout des petites variations

Quelques remarques / conclusions

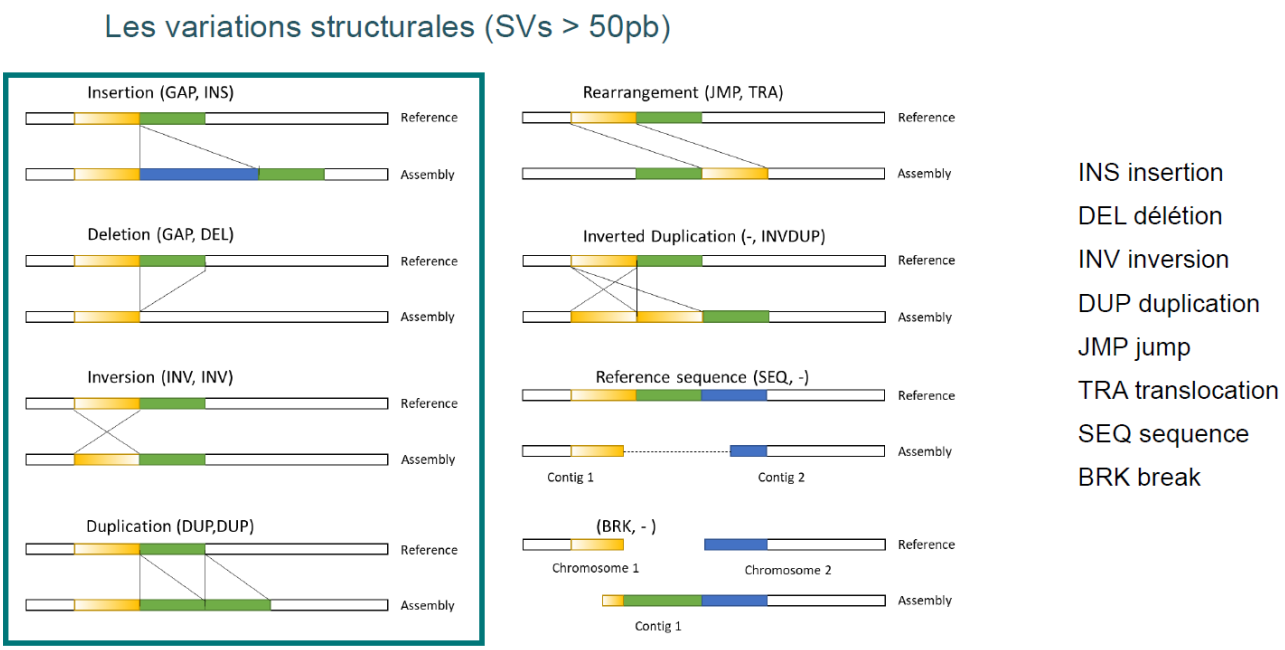
- Les séquences « ONT » ont un taux d'erreurs important (~10%) donc cela pose un problème, peut-être, pour ensuite la détection. Il y a beaucoup de faux positifs (SV) avec ONT vs Bionano
- Mesures ***Col-0*** ONT/***Col-0*** ➔ 52 variations, ***Col-0*** Bionano/***Col-0*** : 36SVs, dont une grande : une translocation au niveau du chromosome 2 connue dans ***Col0*** (ADN mito dans chr2). Ler-1 ONT/ ***Ler-1*** Zapata (118 SVs), ***Ler-1*** Bionano /***Ler-1*** Zapata (91 SVs). Donc les assemblages sont assez complets
- Comparaison ***Ler-1*** ONT/bionano versus ONT ***Col-0*** : mêmes métriques, méthodologie semble efficace pour détecter des SV
- L'avantage en ONT : détection des SVs plus efficace, par contre Bionano permettra de consolider des variants plus courts entre 1 et 10kb car basé sur une absence de marquage commun. En ONT, il suffit d'une rupture pour détecter une variation, les positions sont précises à la base près mais pas forcément justes. Les séquences ONT sont beaucoup plus fragmentées vs les cartes optiques de BioNano.

Questions / Réponses

- Donc ONT plus sensible, par contre attention à la qualité de la référence, puisque basé sur l'alignement, alors que bionano moins dépendant
- MuMmer= favorise les petits alignements, favorise l'alignement local, arbre des suffixes
- MuMmer= aligner des génomes les uns contre les autres, construire les plus grands alignements et ensuite des clusters ➔ actuellement outil le plus efficace pour comparer les assemblages entre 2 génomes. Mais MuMmer est inutilisable sur des grands génomes (maïs, blé)
- Avantages des « Long Reads »
    - Les « Long Reads » vont aider pour l'assemblage des génomes qui contiennent beaucoup de répétitions. Probablement pas si c'est le cas du génome du blé et/ou du maïs. Mais probablement pour des génomes plus classiques.
    - Mais actuellement la comparaison de SVs ne peut pas se faire à grande échelle même avec les « Long Reads » à cause des régions répétées ! Si c'est sur tout génome, cela est vrai pour le maïs et le blé actuellement, mais pas pour Arabidopsis pour lequel il y a des analyses SVs sur tout le génome à partir d'un assemblage.
- Sans passer par une annotation des TEs avant, donc par une expertise par espèce, il est presque impossible de comparer les génomes. Donc il faut annoter les séquences avant de comparer les variations. Même pour Arabidopsis pour valider les SVs la simple comparaison de génomes assemblés ne suffit pas. Dans le cas d'Arabidopsis, une analyse sur des populations en ségrégation est nécessaire pour valider les SVs.
- La vraie question biologique reste de trouver ces variations et pour l'instant les algorithmes de comparaison génomique ne le permettent pas pour tous les génomes!

La conclusion générale semble pointer le fait que l'utilisation des « Long Reads » revient à une problématique d'assemblage de génomes et donc, dans ce contexte, la recherche de variants structuraux à une comparaison génomique et dans ce cas l'annotation experte est très importante (TEs/ncRNA/gènes), voire incontournable.
L'utilisation des short-reads (illumina) reste à ce jour une bonne approche pour la validation de variants structuraux (illumina) dans des régions identifiées par une différence d'annotation ou une rupture d'alignement, par exemple. L'étude de famille en est une autre. Il faut absolument aligner les « short-reads » sur son propre génome, quand cela est possible, pour mettre en évidence les artefacts …
BioNano, finalement, en étant très sensible à la colinéarité pour détecter les SVs, se rapproche peut-être plus d'un comparaison d'annotations que d'assemblages….
Les SV traduisent des stratégies évolutive et adaptatives, sans pour autant ignorer les SV neutres sur un pas de temps plus court (à définir).

5. Photo du Jour



Aurélie Canaguier