Compte Rendu Visio Conférence *PepiAnnot*

17 Septembre 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
BRETAUDEAU Anthony	PF GenOuest, Inria/IRISA, Rennes	anthony.bretaudeau@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
LAPALU Nicolas	INRAE, BIOGER	nicolas.lapalu@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
ROBIN Stéphanie	PF GenOuest, Inria/IRISA, Rennes	stephanie.robin@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 **17 Septembre 2021** :

- Il y avait 18 personnes connectées
- Visioconférence avec GoToMeeting (licence PEPI IBIS). Cette Visioconférence a été enregistrée.
- Deux nouvelles personnes dans le Groupe PepiAnnot : **Stéphanie ROBIN** & **Anthony BRETAUDEAU** qui travaillent avec **Fabrice Legeais** à Rennes.

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

Le PEPI IBIS organise ses journées à la cité de l'espace du 16 au 18 novembre 2021 !
Lien : https://pepi-ibis.inrae.fr/journees-pepi-ibis-2021

Le PepiAnnot propose 3 interventions :

- Philippe Leroy - GDEC : « *Nouveau paradigme pour l'annotation des génomes* » - Introduction
- **Hélène Rimbert** - GDEC : « *Identification et annotation des ARNnc : focus sur quelques familles d'ARNnc* »
- **Christine Gaspin** - MIAT : « *MAGATT un outil de transfert d'annotation.* »

3. Prochaine réunion PepiAnnot

- **Décembre** - réunion informelle portant sur le remplacement de Philippe (qui part à la retraite !) pour la coordination du PepiAnnot avec Véronique
- **Janvier 2022 - 10h00 - 11h30**

4. Alexandre CORMIER, IFREMER IRSI-SeBiMER, Finistère

GigaStore2 - Assemblage de novo résolu au niveau des haplotype du génome de Crassostrea gigas (huitre creuse)

GigaStore 1&2 dans Ifremer
Assemblage *de novo* de l'huitre creuse (C. gigas) qui est le modèle phare pour l'ifremer

Problématique

- Espèce indo-pacifique introduite dans les années 70s
- Problème actuel ➔ infections virales + réchauffement climatique
- Génome très complexe à assembler (550-650 Mb)
- Complexe de déterminer la taille du génome
- Déjà plusieurs génomes de publiés, mais problème de « phasage » des génomes assemblées car espèce hétérozygote

GigaStore 1

- ONT-Illumina v0.1 ➔ problème read courts
- 1er individu choisi était autotriploïde
- De plus une infection a obligé de détruire toute la culture

Donc, en résumé sur GigaStore 1 : 3 génomes sont sortis avec du PacBio Sequel + Illumina mais unphased chromosomes c'est à dire qu'on ne sait pas si le chromosome assemblé est maternel ou paternel. La plupart des assembleurs ne tiennent pas compte des allèles distincts donc mélange les parents. Chez les mollusques 3 ou 4% d'hétérozygotie. Le but est d'obtenir le phasage de SNPs, afin de savoir à quel parent l'allèle appartient.

Des long-reads PacBio/Nanopore peuvent permettre un assemblage ou on sait si chromosome maternel et/ou paternel mais on switch d'un haplotype à l'autre sans savoir lequel est lequel

GigaStore 2 : Objectif : assemlage « phasé »

Dans ce nouveau projet les autotriploïdes ont été éliminés !

PacBio Hifi : moins d'une erreur pour 1000 bases, donc reads de haute qualité (qscore min de 20) et de longueur moyenne 15kb avec le but obtenir des génomes assemblés et phasés, donc un obtenir un haplotype spécifique.

De nombreux papier ont été publiés pour obtenir des haplotypes « phasés » dans la majorité des cas des informations sur des parents sont nécessaires.

Alexandre a testé 2 outils

Canu : basé sur des kmer par haplotype. Les limites c'est que ce sont 2 assemblages séparés et il faut réconcilier les assemblages ensuite…pas toujours simple …

HiFiasm : basé sur des string-graph et utilise les extrémités de graphes mais avantage car fait un seul assemblage et le phasing est fait directement à partir du graph.

⇨ L'objectif serait de comparer les deux

Donc, l'objectif de GigaStore 2 a été de séquencer en Illumina (Génotoul) les 2 parents (120x) et la descendance en PacBio HiFi (50x). Il a également été réalisé du RNA-seq (annotation), HiC (scaffolding) et l'Iso-seq (isoformes) est prévu. Toujours le même problème avec cet organisme, le taux d'hétérozygotie autour de 3% pose beaucoup de problèmes car détection de kmer correspondant à cette hétérozygotie, non homogène. Par exemple, en PE, le R1 est bien avec Qscore > 20 voire 30, mais en R2 très moche.

Une profondeur de 100x est effectivement très contraignante mais nécessaire, a priori, pour couvrir des kmer de 16 à 21 mais Problème de couverture au sein des individus entre distribution des kmers homozygotes ou hétérozygotes.

- Assemblage Hifiasm

836 contigs mais 1Gb au lieu de 650Mb car bcp de duplications dues à l'hétérozygotie. Que ce soit avec String graph ou graphe de De Bruijn, on a des boucles correspondantes aux 2 haplotype même si avec String graph on est censé avoir moins de boucles !

- Assemblage via Canu : ne marche pas mieux ….

Néanmoins, la technologie PacBio Hifi permet quand même un très bon assemblage sur des espèces complexes.

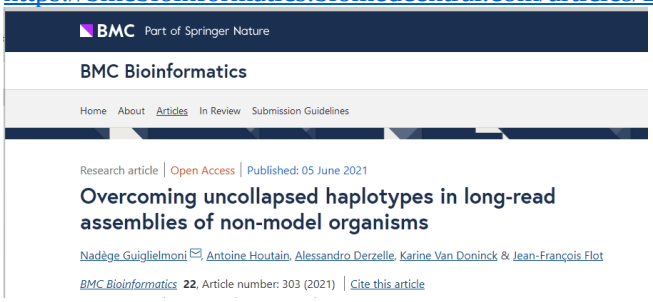
Essai de HiC sur ce génome : marche et a donné 13 ou 10 chromosomes suivant les haplotypes, attendus 10 chromosomes.

Annotations des haplotypes : 2 annotations avec en plus l'annotation des isoformes, donc en plus des librairies Illumina a été utilisé de l'Iso-seq (PacBio) ➔ en cours

- Outil d'annotation *BRAKER2* avec RNA-seq et Iso-seq
- Pour l'annotation d'isoformes il y a l'outil de PacBio *SQanti3*

A l'Ifremer pas mal d'assemblage *de novo* côté génome et transcriptome, et le but est de faire une annotation collaborative via Jbrowse et Web-appolo pour une annotation collaborative en ajoutant les différentes couches.

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04118-3



Remarque : cette démarche marche mieux pour l'humain car 0.1% d'hétérozygotie !

Discussion :

- Nicolas Lapalu mais en garde sur l'utilisation de data isoseq pour l'annotation des isoformes, il faut également valider avec du RNAseq.
- Pour le mapping attention à minipap2 si trop petit exon, mieux gmap

5. Photo du Jour