Evaluer la sélection génomique au sein de croisements bi-parentaux d'espèces pérennes fruitières génotypés par séquençage : point d'étape du projet FruitSelGen

Coordinateur: Timothée Flutre

UMR AGAP (Montpellier), Biologie et Amélioration des Plantes, INRA

timothee.flutre@inra.fr

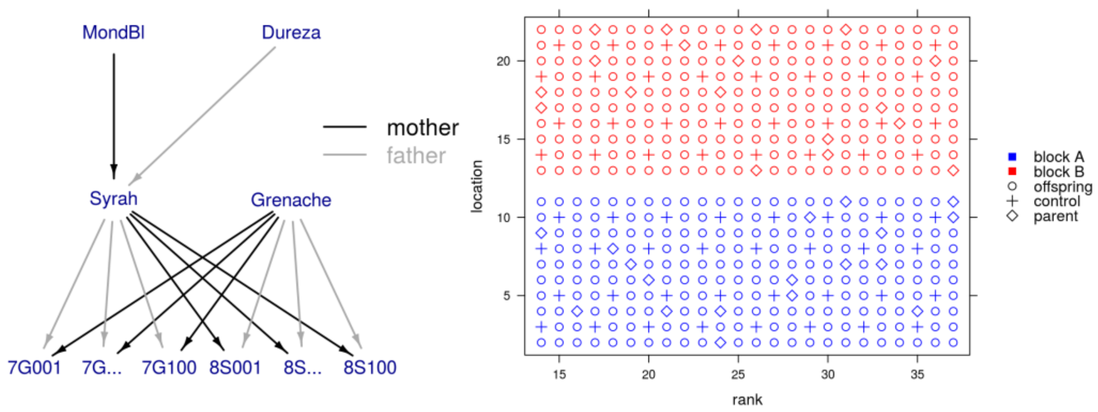
Co-auteurs : A. Launay (UMR AGAP), C. Denancé (UMR IRHS), H. Muranty (UMR IRHS), J-M. Audergon (UR GAFL), C. Confolent (UR GAFL), P. Lambert (UR GAFL), B. Quilot-Turion (UR GAFL), P-F. Bert (UMR EGFV), E. Marguerit (UMR EGFV), E. Dirlewanger (UMR BFP), J. Quero Garcia (UMR BFP), S. Decroocq (UMR BFP), V. Decroocq (UMR BFP), G. Butterlin (UMR SVQV), E. Duchêne (UMR SVQV), E. Costes (UMR AGAP), J-J. Kelner (UMR AGAP), B. Pallas (UMR AGAP), P. Mournet (UMR AGAP).

1. Contexte

Le projet FruitSelGen, financé par le méta-programme Selgen de 2015 à 2016, a pour but d'évaluer la sélection génomique au sein de croisements bi-parentaux d'espèces pérennes fruitières. Ces espèces sont caractérisées par leur difficulté à collecter des données phénotypiques sur les caractères d'intérêt, notamment qualitatifs, en raison de stades juvéniles souvent longs et une forte plasticité phénotypique, contraignant ainsi les programmes de sélection à travailler sur des tailles restreintes de population. De plus, la sélection actuellement pratiquée chez ces espèces n'utilise pas toujours d'information génétique selon l'architecture des caractères. L'innovation que représente la sélection génomique pourrait donc permettre de sélectionner des caractères coûteux à phénotyper, présentant une architecture polygénique à faible héritabilité. Enfin, le génotypage par puce, étant peu flexible et pouvant avoir un coût prohibitif, nous avons utilisé la technique de génotypage par séquençage après digestion par une enzyme de restriction.

2. Croisements et dispositifs au champ

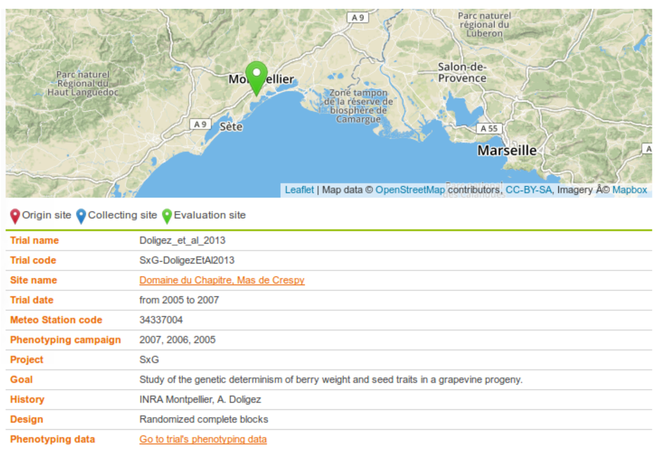
Le projet s'intéresse à 10 croisements, 8 F1 et 2 F2 (leurs parents F1 n'étant plus disponibles). Exemple du croisement de vigne, Syrah x Grenache, à Montpellier :



Deux dispositifs au champ ne sont plus disponibles. Parmi les autres, certains n'ont qu'un seul bloc et pas de témoins répartis.

3. Tâche 1 — Phénotypes

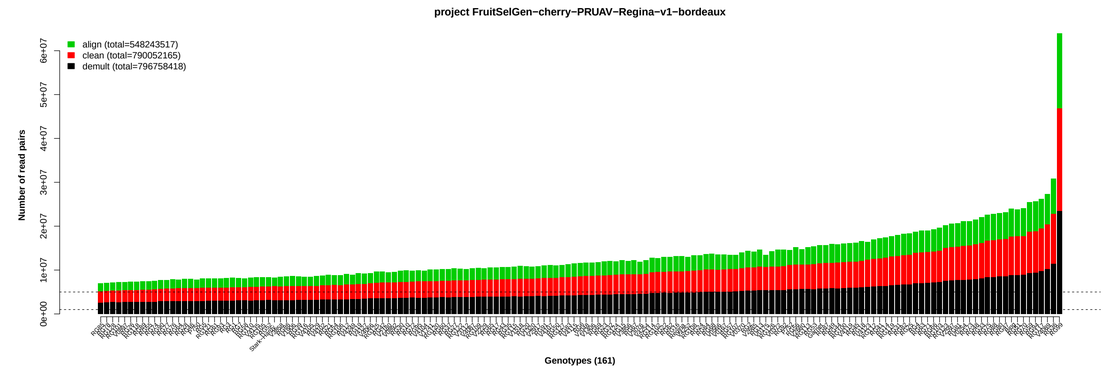
L'objectif est de déterminer si les données phénotypiques brutes peuvent être directement fournies aux logiciels de prédiction génomique, ou bien s'il faut au préalable tranformer les données et tenir compte d'hétérogénéité spatiale et de corrélations inter-annuelles. Cette étape est l'occasion de discuter autour de la modélisation statistique (modèles mixtes, sélection de modèles, diagnostiques des résidus) et a aussi pour but de former les participants aux bonnes pratiques d'analyse de données (traçabilité, reproductibilité).
Pour tous les croisements, les données phénotypiques brutes avaient déjà été acquises. Un dépôt git a donc été mis en place, hébergé sur la plateforme Mulcyber de l'INRA. Il rassemble tous les fichiers de données, ainsi que les scripts d'analyse. Le logiciel `R`, étant connu par la majorité des participants, a été choisi pour cela, avec le paquet `lme4` (Bates et coll., 2015). Toutes les équipes ont débuté l'analyse de leurs données, et certaines l'ont finie.
Par ailleurs, en accord avec les conseils de Heffner et coll. (2009) et afin de préparer la vie des données après le projet, les partenaires "vigne" ont comparé les bases de données VitPhe (UMR MISTEA) et GnpIS (Steinbach et coll., 2013). Cette dernière a été choisie pour commencer à insérer les données du croisement "vigne" de Montpellier. Les partenaires de Colmar ont commencé à faire de même, et les partenaires "vigne" de Bordeaux ont préféré s'investir dans l'amélioration de VitPhe. GnpIS est multi-espèces donc les autres partenaires pourraient y insérer leurs données.



4. Tâche 2 — Génotypage (biologie moléculaire)

Chaque équipe participante a eu la responsabilité d'extraire les ADN de ces échantillons. La qualité de plusieurs extractions n'ayant pas toujours été suffisante, il a été nécessaire de re-extraire certains ADN à Montpellier. Un protocole adéquat a donc été recommandé pour des analyses futures afin d'éviter ce type de problèmes.
Tous les ADN ont ensuite été digérés par l'enzyme ApeKI, et les banques construites à l'aide du protocole d'Elshire et coll. (2011). Une version détaillée de ce protocole a été fournie aux équipes et des formations ont été données sur la plateforme de l'UMR AGAP afin de rendre autonomes les équipes ayant missionné l'un de leurs membres permanents.

Le séquençage a été effectué à GenoToul. A l'automne 2015, 1728 échantillons multiplexés en 48 (pommier) ou 96 (autres espèces) ont été séquencés sur 3 flowcells complètes Illumina HiSeq3000 en 2x75. Le séquençage suite à ce protocole étant caractérisé par une non-uniformité substantielle de la profondeur de couverture, il a été nécessaire de repasser 384 échantillons (certains ayant dû être re-extraits) à l'automne 2016, cette fois-ci sur 4 lanes de HiSeq3000 en 1x150.



5. Tâche 3 — Génotypage (bioinformatique)

Le traitement des lectures brutes a été réalisé à l'aide d'un nouveau programme développé spécifiquement pour le projet, `gbs.py`, utilisant la structure des jeux de données pour les exécuter en parallèle. Il est flexible, documenté et versionné sur GitHub sous licence libre, et dispose d'un test fonctionnel. Ces différentes étapes sont : le contrôle qualité avec `FastQC` ; le démultiplexage ; le nettoyage avec `CutAdapt` ; l'alignement avec `BWA` suivi de `Samtools` et `Picard` ; les réalignements, *SNP/genotype calling* et filtres (profondeur, qualité et certaines erreurs mendéliennes) avec `GATK`. Pour pommier et cerisier, les nouveaux génomes de référence ont été utilisés.
Pour chaque croisement, la médiane de couverture du génome est entre 3,5 et 35% (résultat attendu étant donné la digestion) et dans les zones séquencées, la couverture médiane est entre 19 et 29x (a priori suffisante pour un *calling* de qualité).

espèce	équipe	croisement	descendants	nb.SNP
vigne	Montpellier	SxG	192	28624
vigne	Colmar	RIxGW	256	113353
vigne	Bordeaux	CSxRGM	120	45861
pommier	Angers	PxX6398	251	30797
pommier	Montpellier	X3263xB	301	22946
abricotier	Avignon	GoxMo	186	31334
pêcher	Avignon	PxR (F2)	186	en cours
cerisier	Bordeaux	RxG	120	8640
pêcher-amandier	Bordeaux	CxG	71	38390
pêcher-amandier	Bordeaux	HxC (F2)	48	en cours

A ce stade, d'autres filtres de ségrégation sont réalisés et des cartes génétiques parentales sont en train d'être construites, afin de vérifier que les blocs de recombinaison sur la majeure partie des chromosomes sont bien couverts, en comparant `CarthaGene` (de Givry et coll., 2005) avec `ASMap` (Taylor et Butler, 2017) et `JoinMap` (Van Ooijen, 2011). Une étape d'imputation a également débuté avec `FImpute` (Sargolzaei et coll., 2014).

6. Tâche 4 — Prédiction génomique (biostatistique)

Le choix du logiciel s'est porté sur `GS3` (Legarra et Misztal, 2008 ; Legarra, Ricard et Filangi, 2016) car il implémente efficacement (en Fortran) les modèles statistiques supposant une architecture polygénique ainsi que ceux pour une architecture oligogénique, tout en prenant en compte le pedigree et les marqueurs, ceux-ci pouvant être codés en additif et en dominant. Afin de faciliter son emploi par les participants, le paquet `R rgs3` a été développé ; il est documenté avec des tutoriels, versionné sur GitHub sous licence libre et dispose de tests unitaires. Il automatise notamment la validation croisée par partition permettant ainsi d'évaluer facilement la précision de prédiction.
Dans l'exemple du caractère "taille de la baie" du croisement SxG (Doligez et coll., 2013), les BLUP empiriques des valeurs génotypiques des descendants ont d'abord été obtenues à partir des phénotypes bruts par régression des effets du dispositif expérimental, avec une héritabilité au sens large (répétabilité inter-annuelle) de 0.68. Puis, à l'aide du *single-step* GBLUP (Legarra et coll., 2014) avec effets additifs aux 1600 marqueurs sans données manquantes, en prenant comme poids la variance des BLUP, la corrélation de Pearson vaut 0.78 +- 0.02 (moyenne et déviation standard après validation croisée à 5 partitions).

7. Perspectives

Certaines équipes doivent encore versionner leurs données phénotypiques, et la plupart sont encore en cours d'analyser leurs phénotypes. La majorité ont aussi commencé à se familiariser avec les données de génotypage notamment en construisant des cartes génétiques. La partie principale du travail restant consiste donc à évaluer la précision de prédiction sur leurs caractères, ce qui permettera d'entamer les comparaisons entre croisements en vue d'une soumission d'article début 2018.

8. Financement et Remerciements

Méta-programme Selgen de l'INRA, South Green et URGI.