Utilisation du nouvel assemblage du génome bovin ARS-UCD1.2 à partir de février 2020 - Bilan du projet NOVA

1. Qu'est-ce que l'assemblage du génome

A ce jour, le génome bovin (support des évaluations génomiques) n'est pas connu avec exactitude. La représentation qu'on en a actuellement est dite assemblage. Sa construction résulte directement des méthodes de séquençage qui fournissent un très grand nombre de petites séquences d'environ 150 bases. Ces séquences courtes doivent être positionnées les unes par rapport aux autres, c'est-à-dire assemblées comme un gigantesque puzzle d'environ 2,8 milliards de bases.

« *Le problème de l'assemblage peut être comparé à celui de la reconstruction du texte d'un livre à partir de plusieurs copies de celui-ci, préalablement déchiquetées en petits morceaux.* » - Wikipédia

C'est sur la base de cet assemblage qu'est construite la carte génétique utilisée dans les évaluations, qui correspond au positionnement des marqueurs les uns par rapport aux autres, sur chacun des chromosomes.

2. Pourquoi le mettre à jour ?

La qualité de l'assemblage n'est pas parfaite puisque des régions restent non assemblées (et donc non connues) et d'autres, plus ou moins grandes, peuvent être mal placées ou inversées. Or un assemblage imparfait conduit à des interprétations erronées sur l'organisation et le fonctionnement des gènes. Pour les applications de sélection, un assemblage erroné se traduit par des marqueurs à la position fausse et/ou l'absence de marqueurs dans des régions mal connues.

Des efforts récurrents de la communauté scientifique internationale conduisent à des améliorations progressives de l'assemblage. Ces améliorations résultent du progrès des méthodes de séquençage (les fragments séquencés sont plus longs et donc plus faciles à assembler), des méthodes d'ancrage des séquences, et des algorithmes de calcul. Un nouvel assemblage peut ainsi remettre en cause les positions antérieures de marqueurs. Par exemple, le passage de l'assemblage UMD2 (avril 2009) à UMD3.1 (décembre 2009) avait permis de positionner 17,6 millions de paires de bases supplémentaires.

Toutefois, bien que l'assemblage UMD3 de 2009 soit connu pour être perfectible, aucune évolution significative de l'assemblage bovin n'avait été publiée depuis. Le nouvel assemblage, dénoté ARS-UCD1.2, publié en avril 2018, était donc très attendu. Il comporte 66,2 millions de nouvelles paires de bases positionnées, et corrige un grand nombre d'erreurs de l'assemblage précédent (plus de marqueurs positionnés, et de façon plus précise, donc une carte génétique plus fiable).

Cette bascule vers le nouvel assemblage ARS-UCD1.2 est progressivement mise en œuvre par tous les centres de calcul d'évaluations génétiques.

L'utilisation de l'assemblage ARS-UCD1.2 va donc impacter qualitativement l'imputation des génotypages (la position relative des marqueurs les uns par rapport aux autres ayant pu évoluer, certains allant jusqu'à changer de chromosome), apportant plus de précision et de fiabilité dans la constitution des phases, donc des index. La mise à jour de l'assemblage permet aussi une meilleure

définition des haplotypes liés à des gènes d'intérêts ou d'anomalies, et donc une meilleure identification des animaux porteurs de ces haplotypes.

3. Les travaux réalisés par l'UMT eBis, l'INRA-CTIG et GenEval

Depuis un an, les équipes de l'UMT eBis (INRA, Allice, Idele), le centre INRA-CTIG et GenEval collaborent sur le projet de mise à jour de l'assemblage utilisé dans les outils d'évaluation. Valogene a également été associé sur les étapes traitant des tests sur haplotypes.

3.1 Etablissement de la nouvelle carte génétique

Cette étape consiste à réviser les paramètres des marqueurs connus et utilisés, en fonction de leur positionnement sur le nouvel assemblage.

3.2 Etudes d'impacts sur les haplotypes

Il s'agit là d'évaluer les changements sur les haplotypes utilisés pour les évaluations et pour l'identification d'individus porteurs de gènes d'intérêt ou d'anomalies. La définition des haplotypes et le paramétrage des tests concernés devront être corrigés.

3.3 Gestion des typages et imputations

Ce projet a été une opportunité pour retravailler et optimiser l'échange de données entre l'UMT, l'INRA-CTIG et GenEval :

- paramètres des marqueurs (notamment lors de la mise en production d'une nouvelle puce ou version de puce),
- transferts des génotypages via la base de données nationale zootechnique hébergée par le CTIG.

La stratégie d'imputation des marqueurs a été modifiée pour en imputer un plus grand nombre (seuls les marqueurs utilisés en indexation étaient imputés jusque-là).

3.4 Paramétrages des chaines

Le contrôle de qualité des génotypages a été amélioré (optimisation du calcul du call rate par chromosome).

Les chaines de calculs des IPVGénos, des index et des indicateurs ORI ont été ajustées pour intégrer les nouveaux paramètres, en maintenant au maximum l'isofonctionnalité.

3.5 Etudes d'impacts sur les résultats

Suite aux différents changement réalisés, des comparaisons des résultats sur les IPVGénos, index, indicateurs ORI et tests sur haplotypes ont été effectuées pour estimer les variations à attendre suite à l'implémentation de l'assemblage ARS-UCD1.2.

4. Les résultats

Différents tests ont été réalisés afin de mesurer les impacts de ce nouvel assemblage sur l'ensemble des applications susceptibles d'utiliser les nouvelles phases à savoir l'imputation, les évaluations génomiques, les IPVGénos et indicateurs ORI et les tests sur haplotypes.

4.1 Nouvelle carte et imputation

Depuis sa mise en place en 2011 (à l'arrivée des puces basses densité « LD »), le processus d'imputation ne concerne que les marqueurs utilisés dans les évaluations, soit environ 44 000 marqueurs présents

sur les puces Illumina exploitées à l'époque. Parmi ceux-ci, 99.4% sont présents dans le nouvel assemblage (les marqueurs manquants n'ont pas pu être positionnés de façon satisfaisante). L'UMT eBis, GenEval et Valogène ont validé l'ajout d'environ 10 000 marqueurs supplémentaires à imputer (présents sur les puces actuelles Illumina ou sur la puce EuroGMD), soit parce qu'ils sont utiles pour des tests sur haplotypes, soit parce qu'ils contribuent à l'amélioration de la qualité/fiabilité des phases résultant de l'imputation, soit parce qu'ils seront sans doute intégrés dans les évaluations futures. Enfin, la carte du chromosome X a été densifiée, pour préparer sa prise en compte à l'avenir (en indexation, pour des haplotypes, éventuellement des mutations causales). Aussi, ce sont dorénavant environ 54 000 marqueurs qui seront imputés et phasés chaque week-end.

Les phases générées avec le nouvel assemblage ARS-UCD1.2 peuvent, selon les races et le nombre de typages 50K ou HD disponibles, présenter des différences avec les phases actuelles :

- pour les races laitières bénéficiant actuellement d'une évaluation génomique, les résultats sont très proches, avec en moyenne des résultats présentant 0.06 à 0.6 marqueur imputé différent par animal selon les races ;
- pour les autres races laitières sans indexation génomique en France et où les effectifs d'animaux typés sont réduits ou réalisés principalement sur des puces LD (Jersiaise, Bretonne Pie-Noir, Bleue du Nord et Rouge Flamande), les résultats entre les phases imputés avec les 2 assemblages peuvent être sensiblement différents ;
- pour les 3 races allaitantes bénéficiant actuellement d'une évaluation génomique, les résultats entre les 2 types de phases sont proches, avec en moyenne des résultats présentant 0.2 à 46 marqueurs imputés différents par animal selon les races (selon le taux de typages sur puce 50k ou HD) ;
- pour les races allaitantes rustiques, les résultats sont très variables en fonction des effectifs et du taux d'animaux typés sur puce 50k (de 0.1 différence en moyenne à plus de 1000 différences).

4.2 Impact sur les index

Des comparaisons ont été réalisées entre une évaluation de routine, et une évaluation pilote à partir du nouvel assemblage (et nouveaux paramètres), toutes données égales par ailleurs.

4.2.1 Index laitiers

L'évaluation de référence est le traitement 1920 (juin 2019). La comparaison avec les mêmes données utilisées avec l'assemblage ARS-UCD1.2 montre une corrélation moyenne des index de 0.984 à 0.999 selon la race, et une corrélation moyenne des CD de 0.99. Autrement dit, les variations à attendre sont très faibles.

4.2.2 Index allaitants

L'évaluation de référence est le traitement 2020_01 (septembre 2019). La comparaison avec les mêmes données utilisées avec l'assemblage ARS-UCD1.2 montre une corrélation moyenne des index de 0.967 à 0.996 selon la race pour les évaluations au sevrage, et de 0.97 à 0.99 pour les évaluations carcasses. Les CD ne sont pas impactés (corrélation de 1). Quelques variations pourront donc être observées, mais comparables aux variations suite à 2 traitements consécutifs.

4.3 Impact sur les IPVGénos laitiers et allaitants

Les corrélations des IPVGénos obtenus en utilisant le nouvel assemblage avec les index calculés à partir de l'assemblage actuel sont de l'ordre de 0.97 à 0.99 selon la race, aussi bien en filière laitière qu'allaitante.

4.4 Impact sur les indicateurs ORI

Les indicateurs d'originalité alléliques (ORI) s'appuient directement sur la comparaison des phases des individus de chaque population à comparer. Aussi c'est sur le calcul de ces indicateurs (et sur les tests sur haplotypes, cf 4.5) que l'impact le plus fort est attendu.
Sur le même principe que pour les IPVGénos, les corrélations moyennes des différents indicateurs d'originalité allélique ORIT, ORIC, ORIF sont respectivement de 0.977, 0.972 et 0.989.

4.5 Impact sur les tests sur haplotypes
Les tests sur haplotypes sont potentiellement les résultats les plus impactés par le changement d'assemblage. Mais ce changement a pour but d'améliorer leur précision.
95% des tests présentent une stabilité supérieure à 99% (moins de 1% des individus changent de statut porteur ou non porteur).
Parmi les 4 tests à moindre concordance, 3 d'entre-eux (épilepsie en race Parthenaise, ataxie en race Charolaise et cdh en race Holstein) étaient considérés comme des tests moins précis et gagnent donc en fiabilité. Pour le test HMTCP en race Montbéliarde, des travaux sont encore en cours afin notamment de ne pas prendre en compte certains typages « anciens » dont le clustering des marqueurs pourrait être remis en cause. Pour ces 4 anomalies, un test sur mutation étant disponible par ailleurs, l'impact sur le choix des reproducteurs devrait être limité.
NB : Les résultats des tests sur mutation ne sont pas concernés par cette évolution car ils n'utilisent pas les cartes génétiques (mais s'appuient sur les génotypages vrais).

5. Bilan-conclusion
L'implémentation dans les outils de routine de l'assemblage ARS-UCD1.2 impactera peu les résultats d'évaluation, mais apportera plus de stabilité dans les phases produites chaque week-end, et surtout apportera plus de précision dans les tests sur haplotypes.
Ce projet aura aussi été une opportunité pour améliorer et sécuriser les échanges entre l'UMT eBis, le CTIG et GenEval. Les outils développés et les ajustements de programmes réalisés à cette occasion permettront plus de souplesse et de réactivité lors des futures mises à jour de l'assemblage bovin (mises à jour qui devraient se produire à une fréquence plus élevée compte-tenu de l'avancée des connaissances), et donnent la possibilité d'introduire beaucoup plus facilement de nouveaux marqueurs dans l'imputation (pré-requis à leur utilisation en indexation).

Au vu des études menées, le nouvel assemblage sera implémenté dans la semaine du 18 au 22 février 2020.

Les premiers résultats d'IPVGénos, indicateurs ORI et tests haplotypes sont donc à attendre pour le 24 février 2020.

Nos équipes seront mobilisées sur ce week-end particulier, afin de vous assurer un maximum de transparence dans cette évolution.
Les **index laitiers du traitement 2010** seront également calculés à partir de ce nouvel assemblage, puis dorénavant tous les autres index pour les deux filières.