Design d'une puce à SNP basse densité pour une lignée de poules pondeuses NOVOGEN

Florian Herry

AGROCAMPUS OUEST

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2015 - 2016

Spécialité :

Sciences de l'Animal pour l'élevage de demain

Spécialisation (et option éventuelle) :

……

Mémoire de Fin d'Études

☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☑ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ d'un autre établissement (étudiant arrivé en M2)

Design d'une puce à SNP basse densité pour une lignée de poules pondeuses NOVOGEN

Par : Florian HERRY

Soutenu à Rennes le Vendredi 9 Septembre 2016

Devant le jury composé de :

Maître de stage : Sophie Allais

Enseignant référent : Vanessa Lollivier

Autres membres du jury (Nom, Qualité) :

Rapporteur : Anne Laperche

Président : Thierry Bailhache

Fiche de confidentialité et de diffusion du mémoire

Confidentialité

☐ Non ☑ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible [(1)].

Date et signature du maître de stage [(2)] :

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur[(3)] **Nom Prénom** ----------

autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur :

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.

(2) Signature et cachet de l'organisme

(3).Auteur = étudiant qui réalise son mémoire de fin d'études

(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Remerciements

Je tiens tout d'abord à remercier ma maître de stage, Sophie Allais, pour m'avoir accepté comme stagiaire et pour m'avoir guidé tout au long du stage, aussi bien dans mes divers travaux que dans la rédaction de mon mémoire. Merci pour sa patience et sa pédagogie pour transmettre ses connaissances. Je garderai aussi de très bons souvenirs des nombreux moments passés à réfléchir sur les quelques points surprenants du stage (« Mais pour-quoi ??! »).

Je remercie ensuite Pascale Le Roy pour m'avoir aussi suivi tout au long du stage, pour avoir répondu à mes questions et pour m'avoir expliqué le fonctionnement des évaluations génomiques, que j'aurai la joie d'approfondir par la suite !

Merci à Frédéric Hérault pour sa bonne humeur et pour avoir, lui aussi, répondu à mes questions ! Merci pour les observations faites tout au long du stage et qui ont permis d'améliorer les travaux réalisés.

Merci à vous trois pour avoir fait des points réguliers avec moi, permettant ainsi de discuter sur les résultats obtenus, sur les éventuelles pistes de travail, tout ceci ayant permis de mener à bien le stage.

Je remercie Thierry Burlot et Amandine Varenne de Novogen, pour m'avoir accepté comme stagiaire, pour avoir fait des points réguliers avec moi, pour avoir répondu à mes nombreuses questions et pour m'avoir fait confiance tout au long de ce stage. Je vous remercie aussi de m'avoir fait découvrir l'entreprise !

Merci à Frédéric Lecerf pour m'avoir débloqué des situations de nombreuses fois lorsque j'avais des problèmes informatiques (vive les retours chariots !) et pour toutes les discussions multiples et variées que nous avons pu avoir tout au long de ces six mois de stage.

Merci à tout le labo de génétique animale, Sandrine, Colette, Morgane, Jean-Marc et Kévin, ainsi que les deux stagiaires Nicolas et Marie, pour m'avoir plus que bien accueilli ainsi que pour tous les bons moments passés lors des pauses cafés entre autres !

Liste des abréviations

A : Adénine (base azotée de l'ADN)
ADN : Acide Désoxyribo-Nucléique
ANR : Agence Nationale de la Recherche
BD : Basse Densité
BLUP : Best Linear Unbiaised Prediction
C : Cytosine (base azotée de l'ADN)
DL : Déséquilibre de Liaison
EBV : Estimated Breeding Values
G : Guanine (base azotée de l'ADN)
GEBV : Genomic Estimated Breeding Values
HD : Haute Densité
IFIP : Institut de la Filière Porcine
INRA : Institut National de la Recherche Agronomique
IP : Intensité de Ponte
LAB : Couleur des œufs
MAF : Minor Allele Frequency
PO : Poids d'Œufs
QTL : Quantitative Trait Locus
SNP : Single Nucleotide Polymorphism
SYSAAF : Syndicat des Sélectionneurs Avicoles et Aquacoles Français
T : Thymine (base azotée de l'ADN)
UtOpIGe : vers une Utilisation Optimale de l'Information Génomique dans les schémas pyramidaux

Liste des figures

Listes des Tableaux

Introduction

Dès 1970, l'entreprise Grimaud Frères Sélection, à l'origine du Groupe Grimaud, s'est lancée dans la sélection génétique en canard en collaboration avec l'INRA. Le Groupe Grimaud s'est progressivement développé et diversifié avec des élevages de lapins, de pigeons et l'acquisition du pool génétique ponte et chair en 1997. À partir de là, les programmes de sélection ponte et chair au sein du Groupe Grimaud se sont développés avec notamment l'acquisition de la société Hubbard (poulets de chair) en 2005 puis la création de la société Novogen (poules pondeuses) en 2008. Les objectifs de Novogen sont multiples : obtenir une excellente qualité d'œufs avec une productivité et un rendement optimaux, tout en s'adaptant aux divers systèmes de production. Novogen concentre donc aujourd'hui ses efforts sur la sélection, et notamment la sélection génomique, ainsi que le développement de produits répondant aux demandes des différents marchés.

En parallèle, l'émergence des marqueurs moléculaires tels que les SNP dans les années 90 a permis le développement de biotechnologies qui ont abouti en 2011 et en 2013 à la création, en filière volaille, de puces à SNP (Single Nucleotide Polymorphisms) haut débit (60 000 et 600 000 SNP) par les entreprises Illumina et Affymetrix (Groenen et al., 2011 ; Kranis et al., 2013). La sélection génomique s'est ainsi développée à partir de 2011 grâce à ces nouvelles technologies et utilise les informations des SNP, en complément des mesures de performances, afin de choisir les futurs reproducteurs parmi un grand nombre de candidats à la sélection.

Dans cette optique, sous l'impulsion de quatre laboratoires INRA (Rennes, Toulouse, Jouy-En-Josas et le Rheu), d'institutionnels (IFIP et SYSAAF) et des sélectionneurs Bioporc et Novogen, le projet ANR UtOpIGe (vers une Utilisation Optimale de l'Information Génomique dans les schémas pyramidaux) s'est mis en place en 2010 afin de fournir les informations nécessaires pour la mise en place d'une sélection génomique dans les filières ponte et porcine, caractérisées par des schémas de sélection pyramidaux. Les poules pondeuses de Novogen ont notamment servi de population d'étude du projet avec trois générations de 500 coqs et 40 000 poules pondeuses descendantes de la première génération de coqs. Les trois générations de coqs ont été génotypés sur des puces à 600 000 SNP. Cependant, le coût des puces à SNP haut débit est encore conséquent et le génotypage des coqs candidats à la sélection ne peut pas être utilisé en routine. Un des enjeux actuels de la sélection génomique est donc le développement de puces à SNP basse densité coûtant moins cher. À partir des génotypages haute densité des parents et des génotypages basse densité des candidats à la sélection, une imputation (prédiction) des génotypages manquants sur la puce à SNP haute densité des candidats à la sélection est ensuite réalisée pour obtenir leurs génotypages haute densité.

Au cours du stage, après avoir réalisé des imputations sur différentes puces basse densité à partir de plusieurs scénarii populations, les objectifs ont été de mesurer l'efficacité de l'imputation selon plusieurs critères en fonction des stratégies étudiées. L'influence de l'efficacité de l'imputation sur les évaluations génomiques a ensuite été étudiée. Une première partie bibliographique illustre l'organisation de la filière poules pondeuses ainsi que les points nécessaires pour la compréhension du fonctionnement de l'imputation et les mesures de son efficacité, puis dresse un panorama des différents travaux d'imputations qui ont été menés jusqu'à aujourd'hui. Les matériels et méthodes utilisés au cours du stage sont ensuite développés. Enfin, les résultats concernant les mesures d'efficacité de l'imputation en fonction des différentes stratégies étudiées et l'impact des imputations sur les évaluations génomiques sont présentés et discutés.

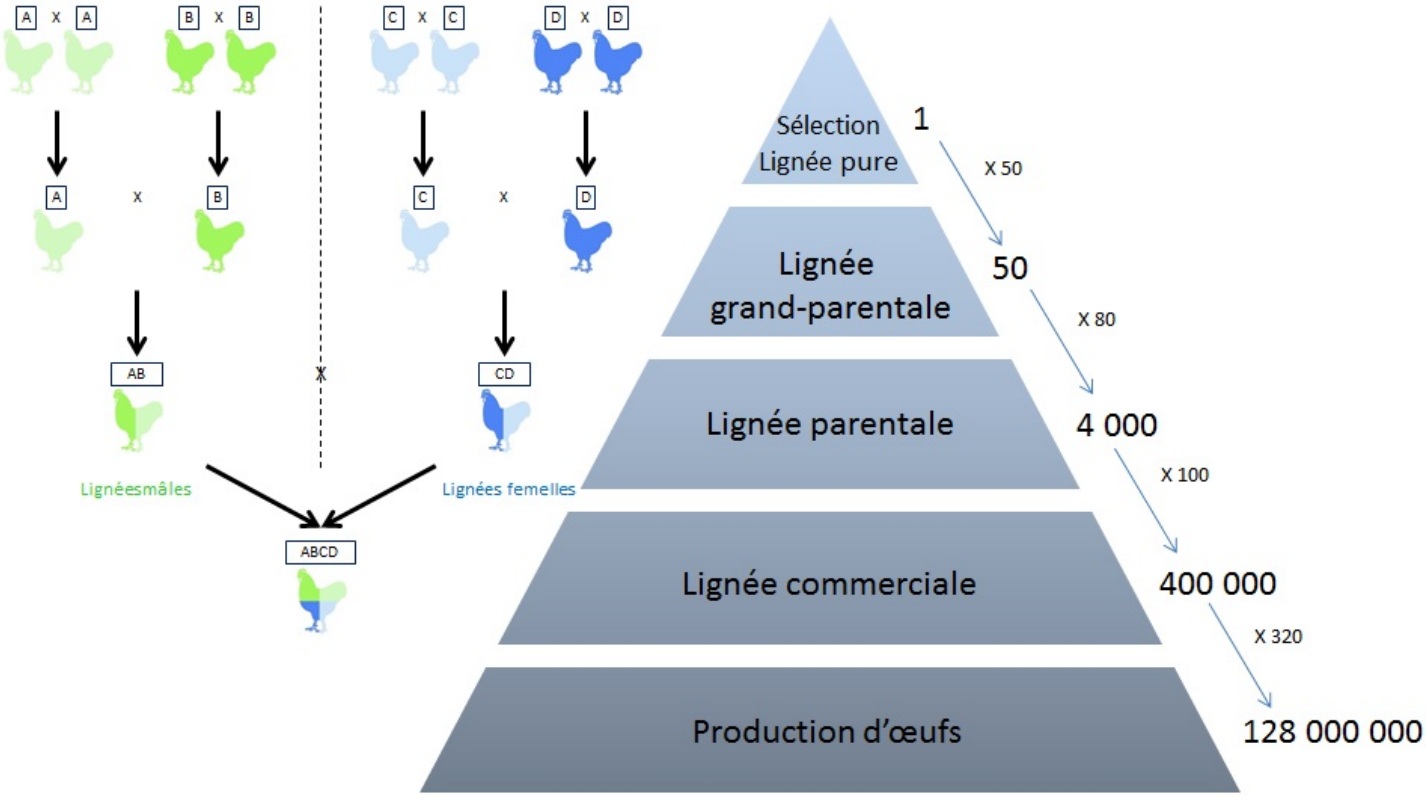


Figure 1 : Organisation pyramidale de la filière ponte - Exemple de NOVOGEN

Les lignées A et B correspondent aux lignées mâles, les lignées C et D aux lignées femelles.

Les lignées grand-parentales sont issues des croisements intra-lignées et les lignées parentales des croisements entre lignées mâles (A et B) et entre lignées femelles (C et D).

La lignée commerciale est obtenue par croisement des individus des lignées parentales mâles AB et femelles CD.

I) Partie bibliographique

A) L'organisation de la filière poules pondeuses

1) Une organisation pyramidale

La filière avicole repose sur une organisation pyramidale, avec une séparation entre les étages de sélection, de grand-parentales et parentales (multiplication), et de production (Figure 1).

Le haut de la pyramide de la filière volaille de ponte représente les sélectionneurs qui sont des entreprises privées et qui disposent le plus souvent de trois ou quatre lignées pures. Après avoir sélectionné les animaux qui sont à croiser, les sélectionneurs fournissent aux multiplicateurs les animaux des lignées pures afin que des croisements intra-lignées soient réalisés pour obtenir trois ou quatre lignées grand-parentales. Ces lignées sont ensuite croisées afin d'obtenir deux lignées parentales. Les poussins issus du croisement des deux lignées parentales et qui forment la lignée commerciale sont ensuite envoyés dans des élevages de production. La lignée commerciale sert pour la production d'œufs de consommation (Figure 1).

À chaque échelon de la pyramide on multiplie le nombre d'individus et on démultiplie aussi le progrès génétique réalisé à l'étage de sélection (Guéméné et al., 2011), progrès génétique rendu possible par le développement de plusieurs méthodes de sélection.

La sélection de poules pondeuses est gérée par de grands groupes internationaux (le groupe Wesjohann en Allemagne – sociétés Lohmann et Hylines, le groupe Hendrix Genetics aux Pays-Bas et le groupe Grimaud en France) et la compétition est très intense entre ces différents groupes. Il en résulte une très grande difficulté pour accéder à des données ou à des résultats.

2) Les méthodes de sélection en poules pondeuses

La sélection en poule pondeuse se fait selon plusieurs objectifs traduits en critères de sélection propres à chaque entreprise de sélection (nombre d'œufs par poule et par an, poids des œufs, indice de consommation, etc.). Les sélectionneurs estiment les valeurs génétiques des candidats à la sélection pour ces différents critères et, sur la base de ces estimations (appelées index), choisissent les meilleurs reproducteurs qui créeront la génération suivante. L'objectif est de produire du progrès génétique pour les différents critères de génération en génération.

Depuis les années 90, la méthode du Best Linear Unbiased Prediction (BLUP) développée par Henderson (Henderson, 1973 ; Henderson, 1975 ; Wolc et al., 2014) est utilisée par les sélectionneurs pour calculer les index. Cette méthode permet de calculer les valeurs génétiques de tous les individus ayant ou non des phénotypes et des descendances, en les corrigeant simultanément pour les effets fixes du milieu, tout en tenant compte des individus apparentés (Henderson, 1973 ; Robert-Granié et al., 2011). Aujourd'hui cette méthode est encore très utilisée en l'absence d'information moléculaire.

En parallèle, les années 90 ont vu l'émergence des marqueurs moléculaires. Ce sont des fragments d'ADN polymorphes qui peuvent servir de repères pour suivre la transmission d'une portion de chromosome d'une génération à l'autre (Boichard et al., 1998 ; Allais et al., 2015). Ces marqueurs sont de deux types : les microsatellites, correspondant à des répétitions de motifs très courts de 2 à 5 bases, très polymorphes mais peu fréquents, et les SNP (Single Nucleotide Polymorphism) correspondant à des changements d'une seule base (A, T, G, C) à un locus donné, très fréquents et apparaissant de façon assez régulière le long de l'ADN. De plus, de nombreuses études (Shrimpton and Robertson, 1998 ; Hayes and Goddard, 2001) ont montré que les caractères quantitatifs étaient sous l'influence d'un très grand nombre de gènes, chaque allèle de ces gènes ayant un petit effet sur les caractères, avec des effets cumulatifs. Ainsi, dès la fin des années 90, Haley et Visscher ont proposé d'utiliser plusieurs milliers de marqueurs moléculaires pour estimer la valeur génétique des individus, ces marqueurs devant permettre de suivre la transmission, d'une génération à l'autre, de tous les gènes intervenant sur un phénotype d'intérêt. Dans les faits, la sélection génomique s'est développée dans la filière avicole à partir de

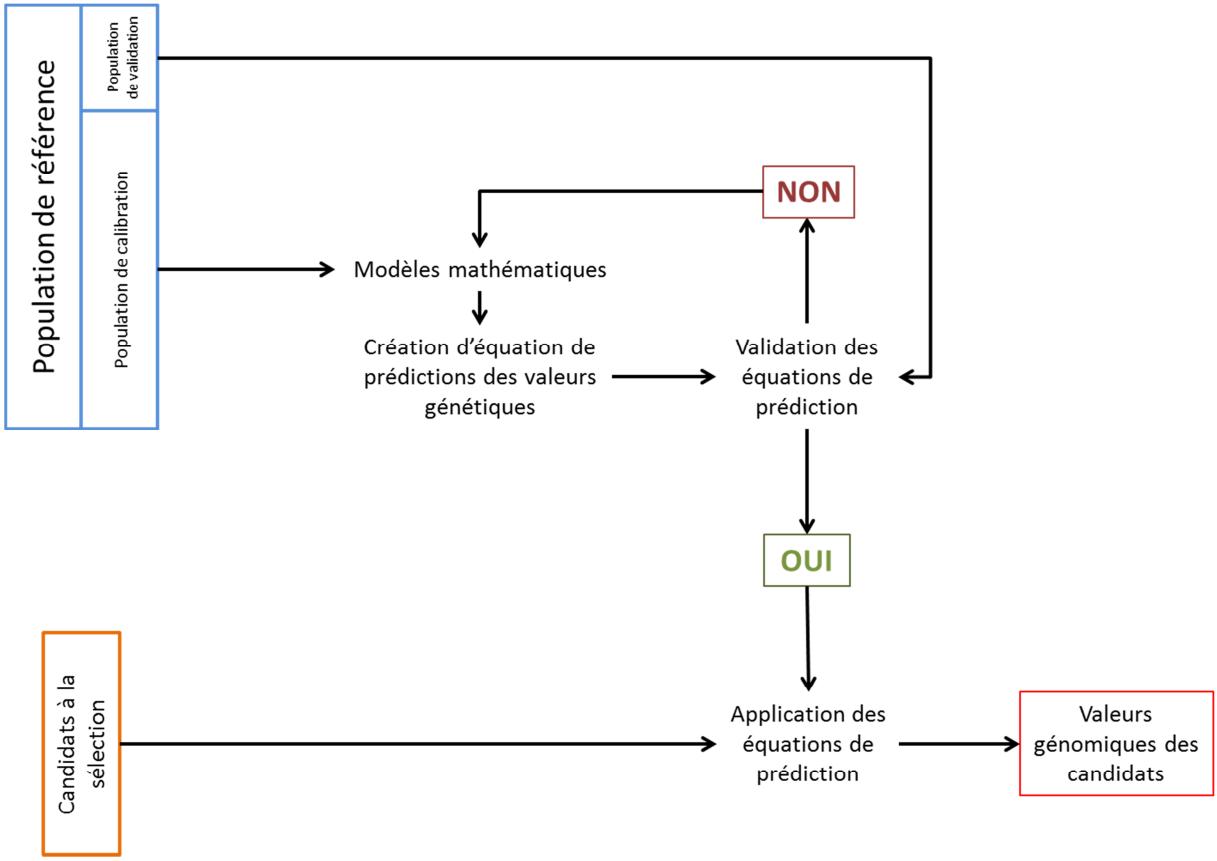


Figure 2 : Démarche de la sélection génomique

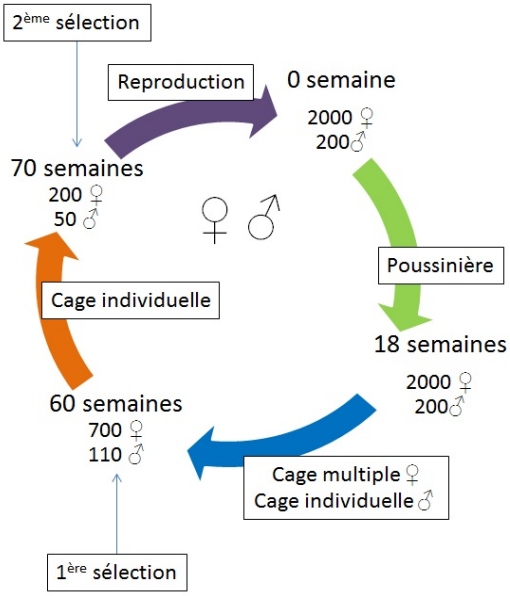


Figure 3 : Schéma de sélection classique - Exemple de NOVOGEN (adapté de Yoannah François, communication personnelle 2015)

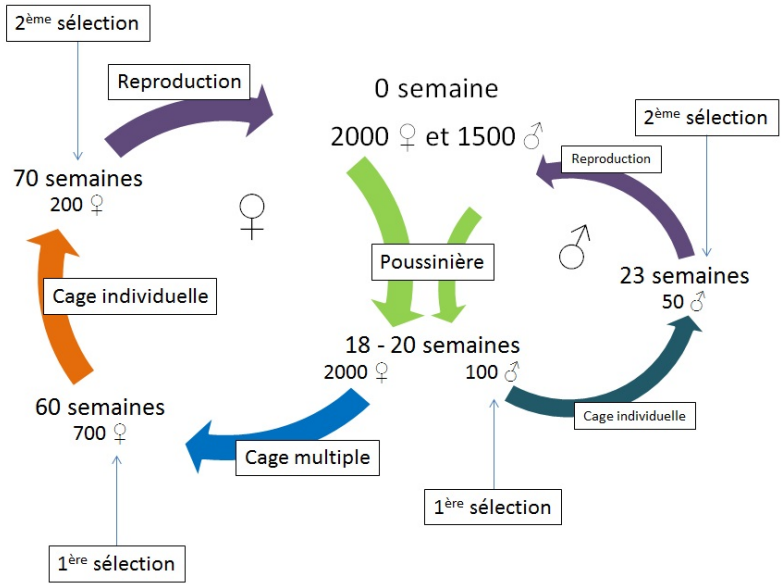


Figure 4 : Modification du schéma de sélection avec l'apport de la sélection génomique - Exemple de NOVOGEN (adapté de Yoannah François, communication personnelle 2015)

2011 grâce au développement de puces SNP haut débit par des entreprises de biotechnologies telles qu'Illumina et Affymetrix (Groenen et al., 2011 ; Kranis et al., 2013 ; Le Roy et al., 2014 ; Wolc et al., 2015), la poule ayant été en 2004 la première espèce d'élevage à avoir son génome entièrement séquencé (International Chicken Genome Sequencing Consortium, 2004a).
La sélection génomique se décompose en deux étapes (Robert-Granié et al., 2011 ; Le Roy et al., 2014). Dans une première étape, de nombreux animaux (le plus souvent des mâles pour des raisons de coûts) sont génotypés sur des puces à SNP haute-densité et phénotypés. Des équations de prédictions sont ensuite mises en place afin de faire le lien entre les génotypes aux marqueurs SNP et les phénotypes. Cette population constitue la population de calibration qui doit être aussi grande que possible. Dans un deuxième temps, grâce aux équations obtenues et validées sur une population de validation génotypée et phénotypée, on peut prédire les valeurs génétiques de jeunes animaux candidats à la sélection pour lesquels on ne dispose pas de phénotype, et éventuellement pas d'information de parenté, en se concentrant uniquement sur les génotypes aux marqueurs SNP (Figure 2).
Aujourd'hui, le prix des puces à SNP haute-densité (HD) est encore élevé (environ 150€ pour une puce 600 000 SNP). Il est cependant possible à partir de puce à SNP basse-densité (BD) d'obtenir les génotypages BD correspondant et de déduire les génotypages HD grâce à la technique de l'imputation (partie II).

3) Schémas de sélection en filière ponte

Les schémas de sélection sont spécifiques à chaque entreprise de sélection. Par exemple, la société NOVOGEN du Groupe Grimaud a opté pour une sélection en deux étapes.
Dans le schéma de sélection classique basé sur la méthode du BLUP (Figure 3), les femelles et mâles sont sélectionnés aux mêmes âges. La première période d'élevage des 2000 femelles et 200 mâles candidats dure 18 semaines et se fait en poussinière. À la fin de cette période, la quasi-totalité des poules pondeuses est transférée dans des cages multiples (cages collectives de pleines-sœurs) et les coqs en cages individuelles. Au bout de 60 semaines d'âge la première sélection intervient, notamment sur la base de performances de poids et de qualité d'œufs. On passe ainsi de 2000 à 700 femelles et de 200 à 110 mâles. Les animaux qui passent la première étape de sélection sont ensuite transférés pendant 25 semaines en cages individuelles. Après 10 semaines, une deuxième étape de sélection sur la base des performances individuelles des animaux intervient. Les sélectionneurs choisissent alors les 200 meilleures femelles et 50 meilleurs coqs afin de les mettre en reproduction pour obtenir la génération suivante, qui sera à nouveau constituée de 2000 femelles et 200 mâles.

4) Intégration de la sélection génomique dans les schémas de sélection

Avec l'apport de la sélection génomique, les schémas de sélection avicole évoluent comme ce fut le cas pour les grandes races bovines laitières avec l'arrêt du contrôle sur descendance permettant ainsi la diminution des intervalles de génération et l'utilisation de jeunes taureaux en sélection (Fritz et al., 2011). L'évolution du schéma de sélection Novogen est présenté en figure 4.
Le prix des puces HD étant élevé tout comme le nombre de candidats à la sélection, un génotypage de l'ensemble des candidats sur puces HD entraîne des coûts importants car on ne garde qu'une petite partie des candidats à la sélection. Une solution à ce problème est le génotypage des candidats à la sélection en utilisant des puces BD moins chères (environ 40€ pour des puces entre 3 000 et 10 000 SNP). Grâce à la méthode de l'imputation on peut déduire les génotypages HD des candidats à la sélection à partir des génotypages BD des candidats et des génotypages HD des fondateurs (Dassonneville, 2012 ; Wolc et al., 2015 ; Wolc et al., 2016). Avec la sélection génomique, de nombreux sélectionneurs augmentent la taille de la base de sélection (c'est-à-dire le nombre de candidats) et la sélection intervient plus tôt dans la vie des animaux. Par conséquent, on augmente l'intensité de sélection et on réduit l'intervalle de génération ce qui

	Nombre d'haplotypes	
	Cas 1	Cas 2
	50	140
	50	0
	50	0
	50	60
Fréquences alléliques	$F(A) = F(a) = 0{,}5$ $F(B) = F(b) = 0{,}5$	$F(A) = 0{,}7\,;F(a) = 0{,}3$ $F(B) = 0{,}7\,;F(b) = 0{,}3$
Fréquences haplotypiques théoriques	$F(AB) = F(A)F(B) = 0{,}5 * 0{,}5 = 0{,}25$ $F(Ab) = F(A)F(b) = 0{,}5 * 0{,}5 = 0{,}25$ $F(aB) = F(a)F(B) = 0{,}5 * 0{,}5 = 0{,}25$ $F(ab) = F(a)F(b) = 0{,}5 * 0{,}5 = 0{,}25$	$F(AB) = F(A)F(B) = 0{,}7 * 0{,}7 = 0{,}49$ $F(Ab) = F(A)F(b) = 0{,}7 * 0{,}3 = 0{,}21$ $F(aB) = F(a)F(B) = 0{,}7 * 0{,}3 = 0{,}21$ $F(ab) = F(a)F(b) = 0{,}3 * 0{,}3 = 0{,}09$
Fréquences haplotypiques observés	$F(AB) = \frac{50}{200} = 0{,}25 = F(Ab) = F(aB) = F(ab)$	$F(AB) = \frac{140}{200} = 0{,}7$ $F(Ab) = F(aB) = \frac{0}{200} = 0$ $F(ab) = \frac{60}{200} = 0{,}3$
	Équilibre de liaison gamétique	Déséquilibre de liaison gamétique

Tableau 1 : Exemple d'une situation d'équilibre et de déséquilibre de liaison gamétique

permet d'augmenter le progrès génétique. On peut ainsi réduire les pertes économiques liées aux génotypages des candidats qui ne seront pas sélectionnés.

B) L'imputation : du fonctionnement aux mesures d'efficacité

L'imputation consiste à déduire les génotypages HD des candidats à la sélection à partir de leurs génotypages BD et des génotypages HD de la population parentale (Dassonneville, 2012 ; Wolc et al., 2015 ; Wolc et al., 2016). L'imputation s'appuie sur le déséquilibre de liaison entre marqueurs.

1) Liaison génétique et déséquilibre de liaison

La liaison génétique correspond à la position de deux loci sur le même chromosome. Les deux loci sont dit complètement liés si un parent doublement hétérozygote ne produit que des gamètes non recombinants. Au contraire, s'il produit des gamètes recombinants et non recombinants dans les mêmes proportions, les deux loci sont dits non-liés. Le déséquilibre de liaison (DL) se définit comme une association non aléatoire d'allèles au niveau de deux loci dans les gamètes. On considère deux locus A et B sur un même chromosome, avec A et a les allèles du marqueur A, et B et b les allèles du marqueur B (Tableau 1). Un haplotype caractérisant une combinaison d'allèles de différents loci sur un même chromosome, quatre haplotypes sont possibles : A-B, A-b, a-B et a-b. Si les fréquences des différents allèles sont de 0.5, on peut alors s'attendre à ce que la fréquence des différents haplotypes dans la population soit de 0.25. Si la fréquence est différente de 0.25, on fait face à un déséquilibre de liaison (DL). Ce DL caractérise donc une association préférentielle des allèles aux 2 marqueurs, les fréquences gamétiques observées étant différentes du produit des fréquences alléliques. Enfin, il est à noter que plus la liaison entre loci est forte, plus le taux de recombinaison sera faible et plus le déséquilibre de liaison se maintient au fil des générations. La liaison génétique ne créée pas de DL mais elle l'entretient. En conséquence, si des SNP proches (liés) sont en DL dans la population de référence, ils le seront aussi dans la population cible.

Le DL est le plus souvent mesuré par le paramètre r^2 défini par l'équation suivante pour les 2 loci A et B : $r^2 = \frac{(x_{AB} - p_A p_B)^2}{p_A * p_B * p_a * p_b}$ avec x_{AB} la fréquence observée de l'haplotype AB et p_A, p_B, p_a et p_b les fréquences théoriques respectives des allèles A, B, a et b.

Quatre forces évolutives principales sont à l'origine du DL :

- La mutation qui n'apparait que ponctuellement dans un haplotype donné et qui est donc associée uniquement à l'haplotype et aux allèles correspondant. Le DL est alors total entre la mutation et l'haplotype. Au cours du temps, ce nouvel haplotype peut se transmettre aux descendants et devenir plus fréquent.
- La migration et le mélange de populations initialement en équilibre de liaison et qui se retrouvent en déséquilibre de liaison dès lors que les fréquences alléliques sont différentes entre populations.
- La dérive génétique qui décrit au fil des générations le changement de fréquences alléliques et haplotypiques d'une population lié à l'échantillonnage d'un nombre fini de reproducteurs.
- La sélection qui diminue le nombre de reproducteurs et donc réduit le nombre d'haplotypes présents dans la population, qui augmente l'apparentement entre les reproducteurs et la fréquence des allèles favorables, et qui induit donc un déséquilibre au locus soumis à la sélection.

2) Principe global et fonctionnement de l'imputation

L'imputation s'appuie le plus souvent sur deux méthodes assez similaires. La première, celle que nous utiliserons principalement par la suite, est la méthode de la fenêtre glissante chevauchante. La deuxième méthode est celle s'appuyant sur les modèles de Markov cachés. Dans chaque cas, les parents sont génotypés en HD et les descendants sont génotypés en BD.

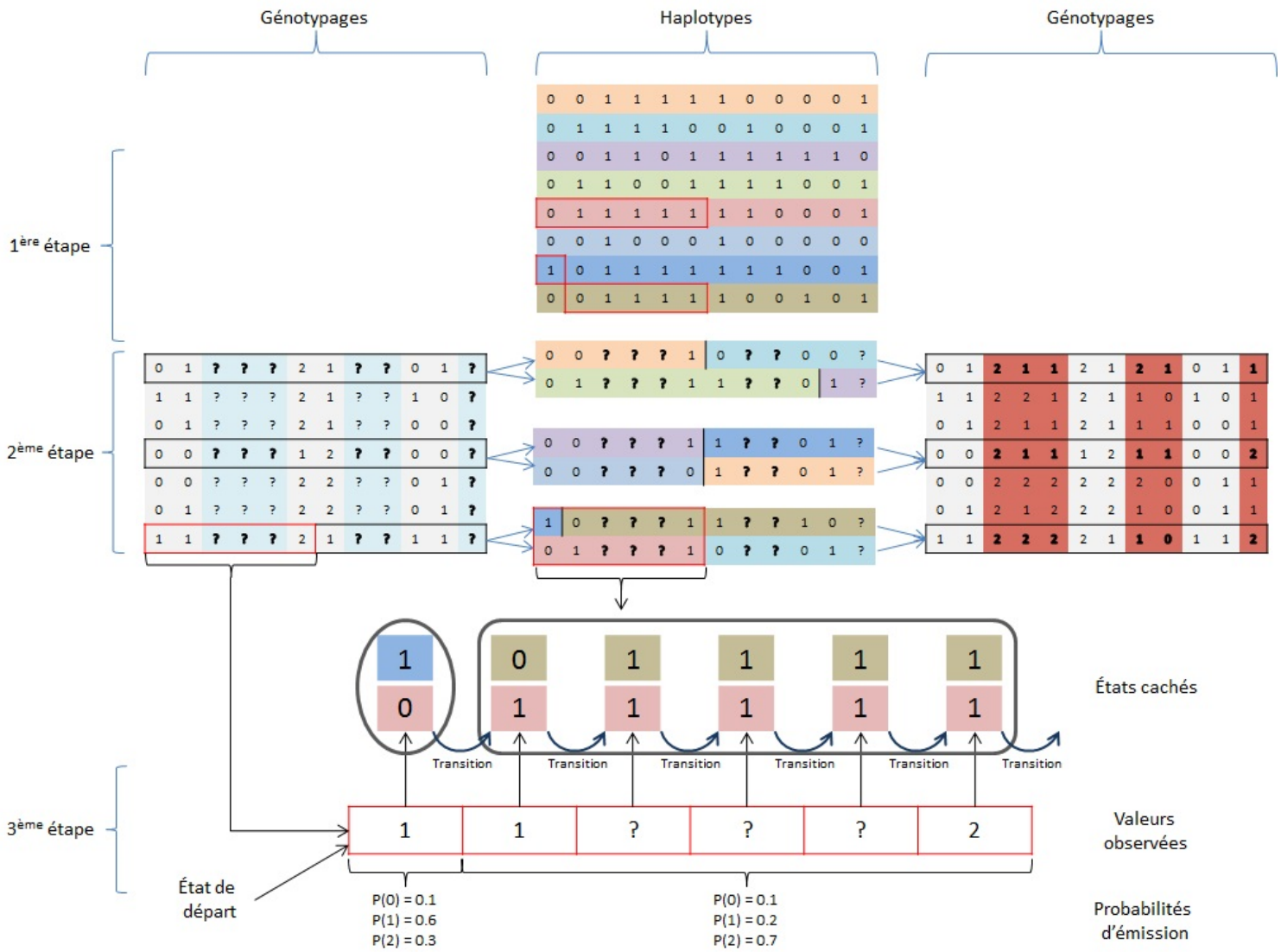


Figure 6 : Exemple de phasage et d'imputation selon la méthode des modèles de Markov cachés

1ère étape : Phasage de la population de référence et création de la librairie d'haplotypes ; 2ème étape : Phasage de la population selon leurs haplotypes (états cachés) correspondants à des fragments d'haplotype de la population référence ; 3ème étape : Calcul des probabilités associées à chaque phasage possible en fonction de l'état de départ du génotypage observé, des probabilités d'émission et des probabilités de transition ; 4ème étape : Conservation du phasage avec la meilleure probabilité associée ; 5ème étape : Imputation des génotypages manquants en fonction du phasage des fragments d'haplotype de référence

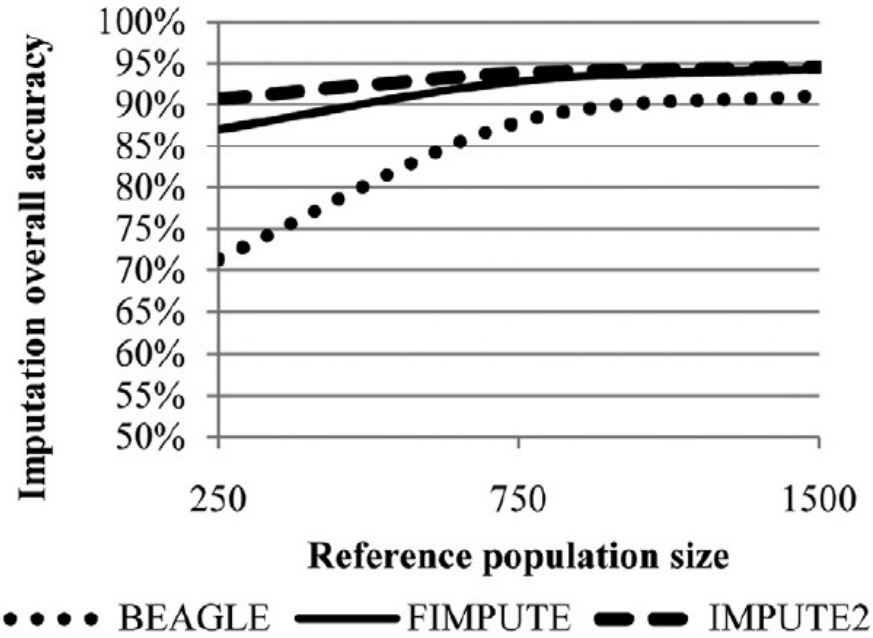


Figure 5 : Influence de la taille de la population sur l'efficacité de l'imputation en utilisant les logiciels Beagle, FImpute et Impute2 (Ventura et al., 2014)

a) Méthode de la fenêtre glissante chevauchante - FImpute

La méthode de la fenêtre glissante chevauchante a été développée en 2014 par Sargolzaei et al. avec le logiciel FImpute. Cette méthode se déroule en 6 étapes :

- 1) La première étape consiste à repérer des paires d'haplotypes parents-descendants avec une fenêtre recouvrant tout le génome.
- 2) On travaille ensuite chromosome par chromosome. On balaye le chromosome avec une fenêtre de 1000 SNP pour construire une librairie d'haplotype basée sur les génotypages HD des parents.
- 3) On essaie de repérer dans la population cible des haplotypes similaires (> 99%) avec ceux de la librairie. On peut ensuite imputer les génotypages manquants.
- 4) Si on n'observe pas d'haplotype similaire, on réduit progressivement la taille de la fenêtre jusqu'à trouver des haplotypes similaires. La taille minimale de la fenêtre est de 2 SNP.
- 5) Une fois que les haplotypes similaires ont été trouvés et que les génotypages ont été imputés, on passe à la fenêtre suivante de 1000 SNP, tout en chevauchant de 750 SNP la fenêtre précédente, et on recommence le processus.
- 6) Si après tous les balayages il reste des génotypages manquants, ils sont imputés en fonction de fréquences alléliques calculées sur le groupe de référence génotypés en HD.

b) Méthode des modèles de Markov cachés - Beagle

Les modèles de Markov cachés sont des modèles stochastiques qui supposent que la distribution des probabilités conditionnelles des états futurs ne dépend que de l'état présent (Dassonneville, 2012). Dans notre cas (Figure 5), ce sont des modèles utilisés pour mettre en relation des observations faites sur les génotypages parfois manquants d'une population cible, avec des « états cachés » correspondants aux haplotypes de la population cible qui sont des « fragments » d'haplotypes de référence, à savoir les haplotypes parentaux (Marchini and Howie, 2010). Le plus souvent, les logiciels utilisant les modèles de Markov cachés réalisent une meilleure imputation en utilisant une population de référence préalablement phasée (*ie* pour chaque haplotype on sait quel allèle vient du père ou de la mère) et génotypée pour tous les SNP. Les logiciels phasent la population cible et considèrent alors que les haplotypes de la population cible correspondent à une « mosaïque » d'haplotypes de la population de référence du fait de cross-over entre haplotypes, mais aussi de quelques rares mutations (Hayes, 2011).

L'imputation selon les modèles de Markov cachés se fait en 5 étapes (Eddy, 2004 ; Marchini et al., 2007 ; Marchini and Howie, 2010 ; Hayes, 2011 ; Dassonneville, 2012) :

- 1) Phasage de la population de référence et création d'une librairie d'haplotype.
- 2) Phasage de la population cible pour laquelle on dispose des génotypages manquants. Cette population est phasée selon leurs haplotypes qui correspondent à des fragments d'haplotypes de la population de référence. Il y a plusieurs possibilités de phasage de la population cible.
- 3) On calcule la probabilité associée à chaque séquence possible.
- 4) On retient le phasage pour lequel on obtient la meilleure probabilité associée.
- 5) Grâce au phasage retenu, on connaît les haplotypes parentaux associés aux génotypages manquants. On peut alors les imputer.

c) Inclusion d'informations de parenté

Dans les deux méthodes citées précédemment il est possible d'inclure des informations de parenté entre population de référence et population cible. Ces informations permettent d'augmenter la précision du phasage et de l'imputation. Ces règles sont basées sur les principes de ségrégation mendélienne. Par exemple, en ayant connaissance des haplotypes parentaux, si l'on suppose que le mâle est homozygote AA et que la femelle est homozygote GG, le descendant sera forcément hétérozygote AG. Ces marqueurs homozygotes servent de marqueurs ancrants pour ensuite phaser les haplotypes et imputer les génotypages manquants. Plus les individus sont proches en terme de parenté, plus ils ont en commun des haplotypes longs. À l'inverse plus les individus sont

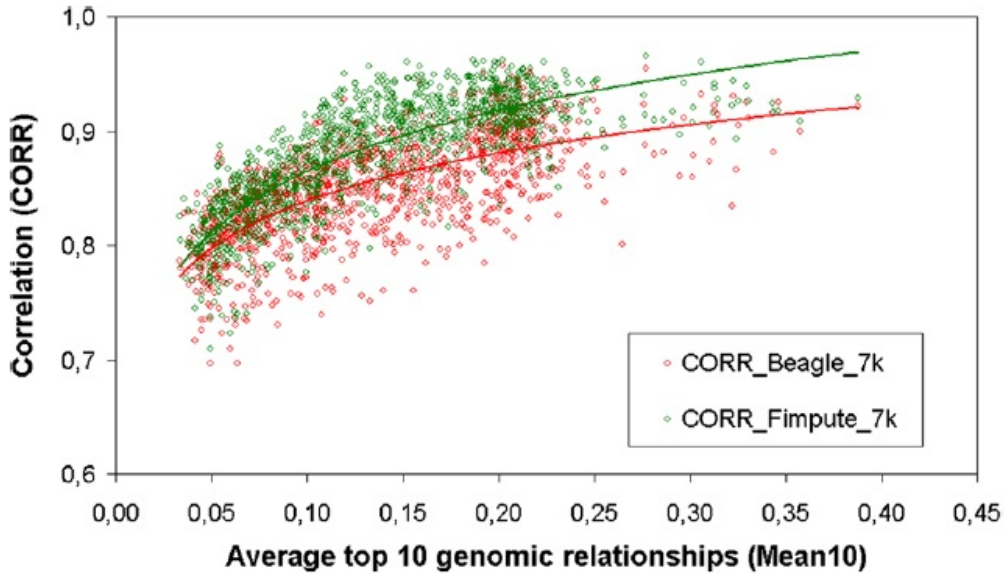


Figure 7 : Précision de l'imputation (corrélation) en fonction de la relation de parenté (Mean10) entre population de référence et population cible, en utilisant Beagle et FImpute (Carvalheiro et al., 2014)

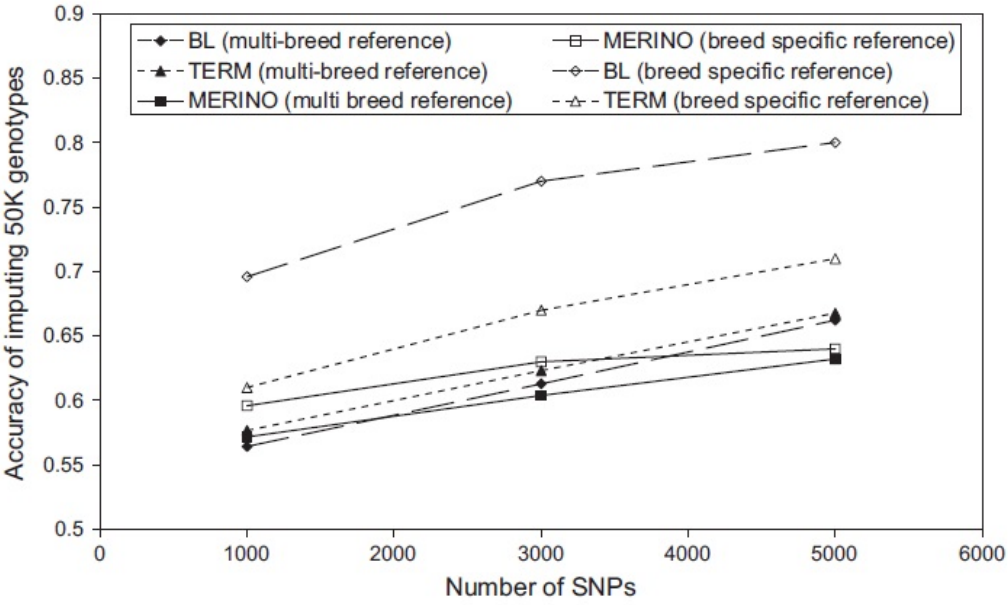


Figure 8 : Précision de l'imputation (corrélation) de la puce Ovine 50K Illumina BeadChip® à partir de puces 1K, 3K et 5K en utilisant une population de référence race pure (breed specific) ou races multiples (multi-breed) (Hayes et al., 2011)

BL : race ovine Border Leicester ; MERINO : race ovine Merinos ; TERM : races ovines White Faced Suffolk et Poll Dorset (races proches génétiquement)

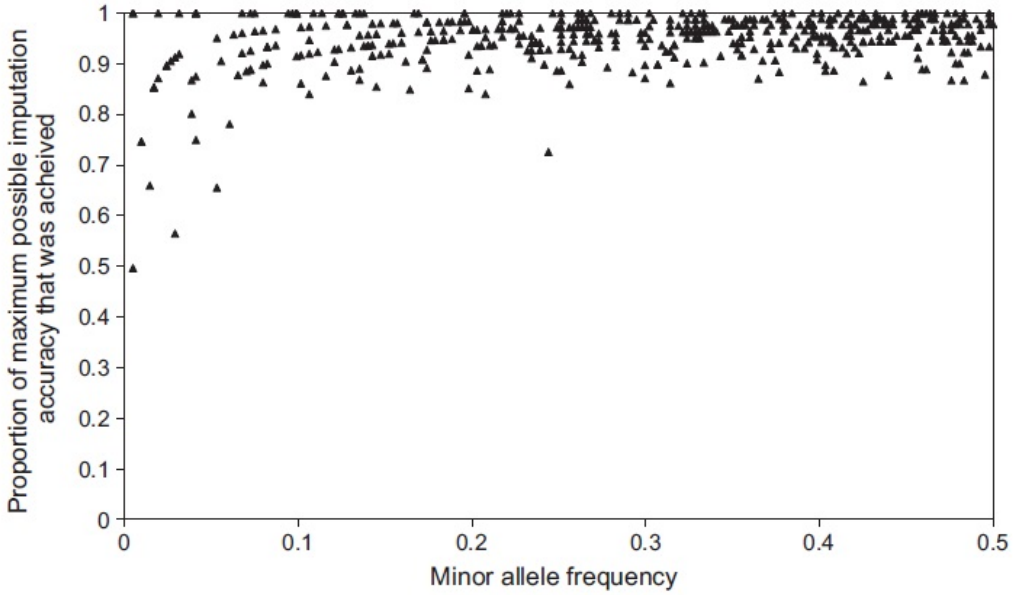


Figure 9 : Précision de l'imputation (corrélation) de la puce Ovine 50K Illumina BeadChip vers la séquence en fonction des différentes MAF (Hayes et al., 2011)

éloignés, plus la longueur des haplotypes en commun se réduit (Sargolzaei et al., 2014).

3) Les divers facteurs pouvant influencer l'imputation

a) Taille de la population de référence et de la population cible

Un des principaux facteurs pouvant influencer l'efficacité de la population est la taille de la population de référence (Figure 6). Plus la taille de la population de référence sera grande, plus la taille de la librairie d'haplotypes sera grande et donc plus on aura de chance de retrouver des fragments d'haplotypes de la population cible dans la librairie préalablement construite. Si des haplotypes de la population cible ne sont pas présents dans la librairie, il est alors très probable que les génotypages correspondant soient mal imputés (Browning and Browning, 2009 ; Hayes, 2011 ; Heidaritabar et al., 2014). Hozé et al. (2013) et Ventura et al. (2014) montrent que chez les bovins il n'y a plus de gain sur l'efficacité de l'imputation au-delà d'un certain nombre d'individus dans la population de référence. L'efficacité de l'imputation dépend de la diversité génétique de la population. Cela indique qu'au-delà d'un certain nombre d'individus dans la population de référence, toute la diversité possible des haplotypes a été captée. La taille de la population cible va aussi jouer sur l'efficacité de l'imputation. Toutefois, l'effet sera plus faible que l'effet dû à la taille de la population de référence (Browning and Browning, 2009), voire même légèrement opposé (Hozé et al., 2013). En effet, plus il y a d'individus dans la population cible, plus on augmente le nombre d'haplotypes à retrouver dans la librairie. Il est alors plus probable de ne pas retrouver un fragment d'haplotype dans la librairie de référence, ce qui entraîne une mauvaise imputation.

b) Relations de parenté entre population de référence et population cible

La relation entre la population de référence et la population cible est un autre facteur pouvant influencer l'imputation (Figure 7) (Hayes, 2011 ; Hickey et al., 2012 ; Hozé et al., 2013 ; Carvalheiro et al., 2014 ; Heidaritabar et al., 2014). Plus les relations de parenté entre la population de référence et la population cible sont proches, plus les individus ont en commun des fragments d'haplotypes de grande taille. À l'inverse, à mesure que les relations de parenté s'éloignent, on augmente le nombre de recombinaisons possibles entre haplotypes. Les individus ont alors en commun avec leurs ancêtres des fragments d'haplotypes de plus petite taille. En conséquence, plus la population de référence et la population cible sont proches, plus elles ont en commun des fragments d'haplotypes de grande taille et plus l'imputation sera facilitée.

c) Densité de marqueurs

Un autre facteur influençant l'imputation est la densité de marqueurs sur la puce BD (Figure 8). De façon assez intuitive, plus il y a de marqueurs sur la puce BD, plus l'imputation vers une puce HD sera meilleure (Hayes, 2011 ; Hickey et al., 2012 ; Hozé et al., 2013 ; Carvalheiro et al., 2014). La densité des marqueurs sur la puce BD est à mettre en relation avec le déséquilibre de liaison entre marqueurs présents sur la puce BD et marqueurs à imputer. En effet, si le nombre de marqueurs présents sur la puce BD est faible, il est plus probable d'identifier par hasard des haplotypes en commun entre la population de référence et la population cible. En conséquence, l'imputation ne sera pas bonne. En augmentant le nombre de marqueurs sur la puce BD, on diminue la probabilité d'identifier par hasard des haplotypes en commun entre la population de référence et la population cible. L'augmentation du nombre de marqueurs permet donc de diminuer l'importance de la relation de parenté entre population de référence et population cible.

d) Marqueurs avec une faible fréquence allélique

Pour les marqueurs avec une faible MAF (Minor Allele Frequency) (Figure 9), la probabilité de retrouver un haplotype concordant dans la librairie d'haplotypes est plus faible que pour des marqueurs avec des fréquences alléliques équilibrées. Cela a pour conséquence que le marqueur avec une faible MAF peut être imputé de façon aléatoire ce qui augmente le taux d'erreur. On voit donc ici toute l'importance d'un grand panel de référence afin de capter toute la diversité des haplotypes possibles et de réaliser une bonne imputation par la suite. Ceci d'autant plus que les marqueurs avec une faible MAF sont susceptibles de jouer un rôle important dans le déterminisme

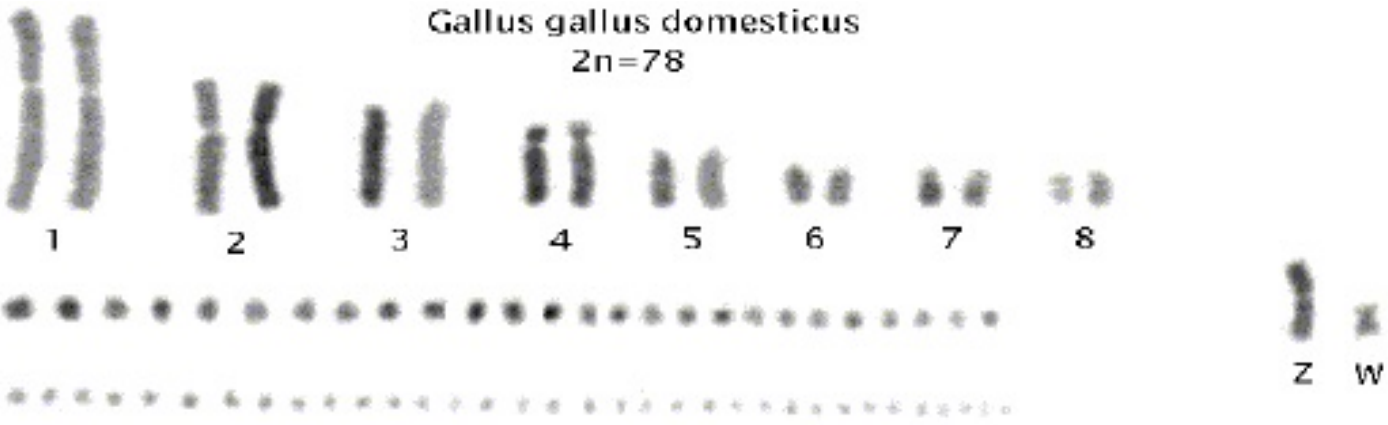


Figure 10 : Caryotype de la poule (Denjean et al., 1997)

Chromosome 1 à 8 : Macro-chromosomes
Chromosome 9 à 38 : Micro-chromosomes
Chromosome Z et W : Chromosomes sexuels (mâles homogamétiques ZZ et femelles hétérogamétiques ZW)

de caractères complexes et peuvent avoir plus d'effets que des marqueurs avec une MAF plus élevée (Hickey et al., 2012 ; Heidaritabar et al., 2014).

4) Les mesures d'efficacité de l'imputation

Il existe deux méthodes pour mesurer l'efficacité de l'imputation. La première consiste à disposer d'animaux génotypés à la fois sur une puce BD et sur une puce HD. L'imputation est ensuite réalisée à partir de la puce BD puis les résultats imputés sont comparés avec les génotypes obtenus par la puce HD. Les inconvénients majeurs de cette méthode sont qu'il faut génotyper deux fois les animaux, ce qui est particulièrement coûteux, et que ces opérations sont à réaliser à chaque fois que l'on veut tester une nouvelle puce BD.
La deuxième méthode est de disposer uniquement pour les animaux de génotypages HD. Puis on simule *in silico* des génotypages BD par création d'une puce BD en « effaçant » des marqueurs présents sur la puce HD. Une fois l'imputation réalisée, on compare les résultats avec les génotypages de la puce HD. On peut ainsi tester autant de puces BD que l'on souhaite. C'est cette méthode qui est majoritairement utilisée dans la littérature et qui est utilisée par la suite.
Quatre critères sont pris en compte pour mesurer l'efficacité de l'imputation (Dassonneville, 2012) :

- Le taux d'erreur génotypique : on regarde les différences entre les vrais génotypages et les génotypages imputés. Si on observe une différence entre les deux génotypages (par exemple un génotypage imputé AG au lieu d'être AA), on considère que l'imputation est fausse.
- Le taux d'erreur allélique : on regarde cette fois les deux allèles au niveau d'un locus. Si on observe une différence d'un seul allèle (le même exemple, un génotypage imputé AG au lieu d'être AA), on peut dire que l'imputation est « à moitié » bonne. Si on observe deux différences au niveau des allèles, l'imputation est totalement fausse.
- Les corrélations : on calcule la corrélation de Pearson entre génotypes imputés et génotypes HD.
- L'impact sur les évaluations génomiques : comparaisons des résultats des GEBV (Genomic Estimated Breeding Values, i.e. les index génomiques) calculées à partir des génotypages HD et des GEBV calculées à partir des génotypages imputés.

C) État des lieux des travaux d'imputation et particularité de l'espèce avicole

1) Les résultats d'imputation en filière bovine, porcine, ovine et avicole

De nombreux travaux d'imputation ont été menés jusqu'à aujourd'hui, aussi bien en filière bovine, porcine, ovine qu'avicole. Ces travaux se sont penchés sur différents logiciels dont FImpute et Beagle, ainsi que sur différents facteurs pouvant influencer l'efficacité de l'imputation.
Un tableau récapitulatif de quelques travaux d'imputation, avec la filière étudiée, le type d'imputation, les facteurs pouvant influencer les imputations et les principaux résultats des études est présenté en annexe 1.
Dans la filière volaille, de nombreux travaux ont été menés par Anna Wolc (Hy-Line) et Marzieh Heidaritabar (Hendrix Genetics). Anna Wolc s'est principalement concentrée sur des simulations d'imputations de diverses puces BD (de 400 à 42K SNP) vers la puce Chicken 600K Illumina BeadChip® et vers la séquence afin d'étudier les conséquences sur les évaluations génomiques. Marzieh Heidaritabar, quant à elle, a étudié les différentes stratégies d'imputations possibles (imputation de diverses puces BD vers la puce Chicken 60K Illumina BeadChip® et imputation de la puce Chicken 60K Illumina BeadChip® vers la séquence) et les divers facteurs pouvant influencer l'imputation.
Enfin, il est à noter que de nombreux travaux (essentiellement dans les espèces non avicoles) ont testé des puces BD avec différentes densités de SNP. Ceux-ci sont le plus souvent choisis à intervalles réguliers le long du génome. Ce choix de puces BD avec des SNP équidistants s'explique par une étendue du déséquilibre de liaison identique le long du génome. Or, en filière

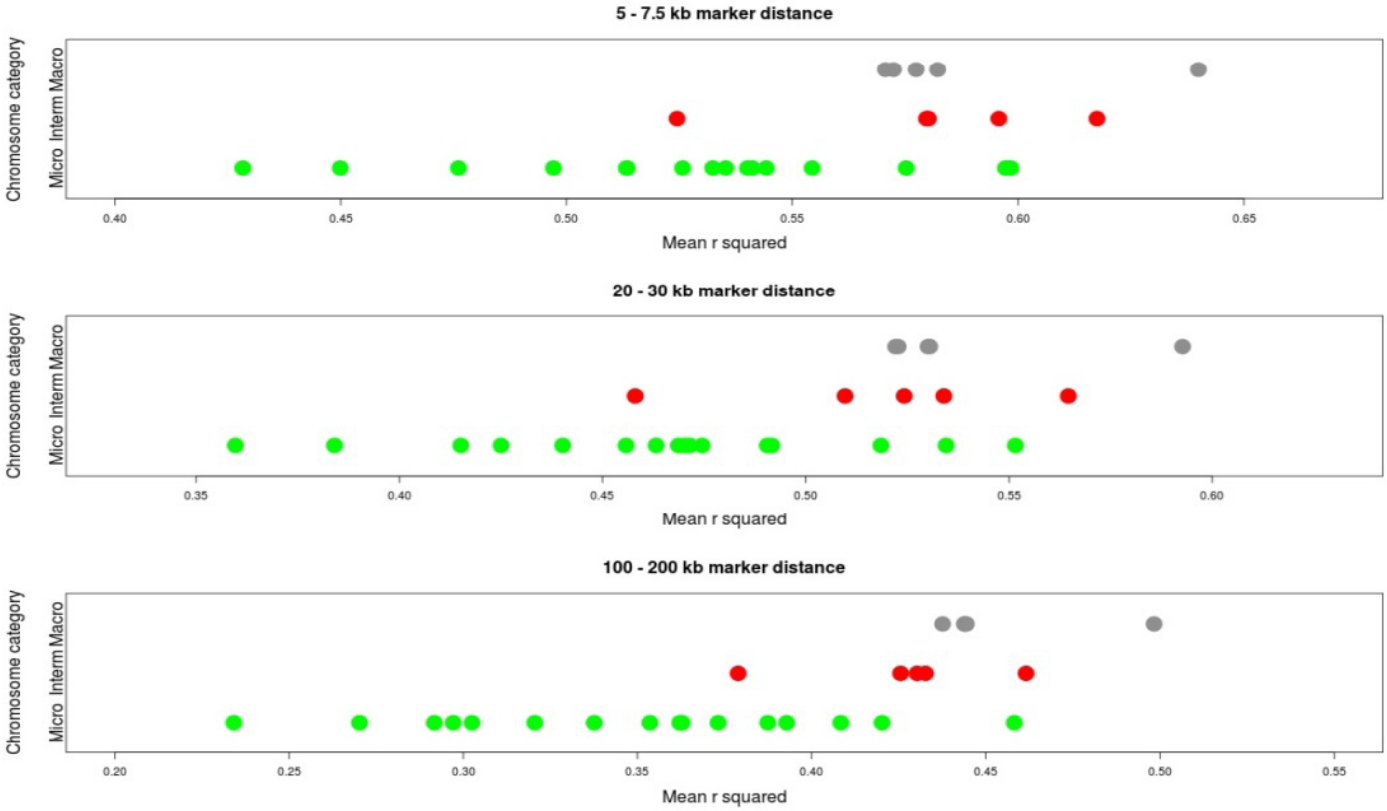


Figure 11 : Étude de la persistance du déséquilibre de liaison pour différentes distances entre couples de marqueurs en fonction des catégories de chromosomes (macro-chromosomes, chromosomes intermédiaires et micro-chromosomes) (Robert et al., 2015)

Chaque point représente un chromosome. Dans cette étude sont considérés comme macro-chromosomes (gris) les chromosomes 1 à 5, comme chromosomes intermédiaires (rouge) les chromosomes 6 à 10, et comme micro-chromosomes (vert) les chromosomes 11 à 28.
Une bonne persistance du déséquilibre de liaison entre couples de marqueurs pour les trois distances étudiées est observée chez les macro-chromosomes. La persistance du déséquilibre de liaison entre couples de marqueurs pour une distance faible (5 -7,5 kb) est aussi observée chez les micro-chromosomes mais elle devient plus variable avec une augmentation de la distance entre couples de marqueurs : $0,36 < r^2 < 0.55$ pour une distance intermédiaire (20 – 30 kb) et $0,23 < r^2 < 0,46$ pour une longue distance (100 – 200 kb)

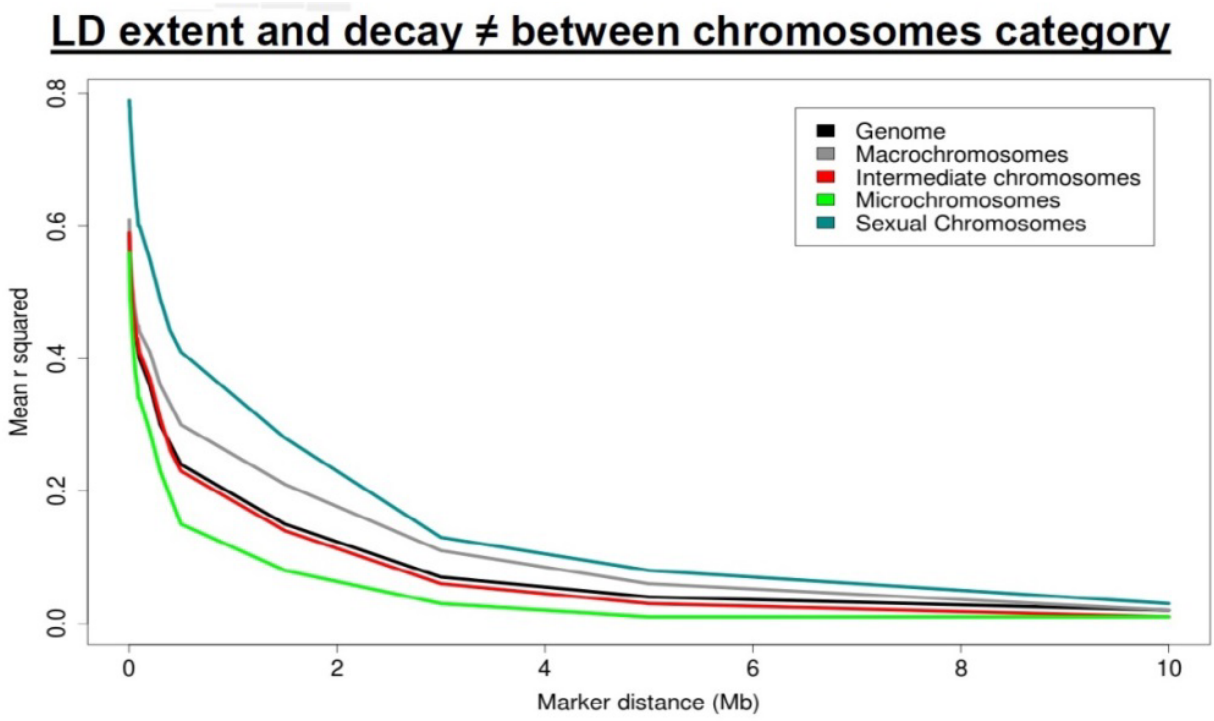


Figure 12 : Étendue du déséquilibre de liaison en fonction des catégories de chromosomes (Robert et al., 2015)

Dans cette étude sont considérés comme macro-chromosomes les chromosomes 1 à 5, comme chromosomes intermédiaires les chromosomes 6 à 10, et comme micro-chromosomes les chromosomes 11 à 28.
Une chute rapide du DL est observée chez les micro-chromosomes ($r^2 = 0,15$ pour une distance de 0,5 Mb) alors qu'il se maintient à plus grande distance sur les chromosomes intermédiaires ($r^2 = 0,15$ pour une distance de 1,5 Mb) et encore plus sur les macro-chromosomes ($r^2 = 0,15$ pour une distance de 2,5 Mb). Le chromosome Z (chromosome sexuel) possède le DL le plus fort.

avicole, la persistance du DL varie beaucoup le long du génome.

2) Les particularités du génome avicole

La poule est la première espèce d'élevage à avoir été séquencée en 2004 (International Chicken Genome Sequencing Consortium, 2004a ; International Chicken Genome Sequencing Consortium, 2004b). Depuis, quatre nouvelles versions de l'assemblage ont été publiées, la cinquième et dernière version venant d'être publié en 2016. Le caryotype de la poule (Figure 10) présente la particularité d'être sous-divisé en 9 paires de macro-chromosomes, dont les chromosomes sexuels, et 30 paires de micro-chromosomes. Les mâles sont homogamétiques ZZ et les femelles hétérogamétiques ZW (Bloom et al., 1993 ; Vignal, 2000). Les micro-chromosomes sont caractéristiques des oiseaux et d'une partie des reptiles et des poissons. Ils présentent une densité de gènes plus élevée que les macro-chromosomes mais la densité des SNP est environ la même pour tous les chromosomes (5 SNP.kb^{-1}). Enfin, le génome de la poule représente 1Gb soit environ un tiers de celui des mammifères (International Chicken Genome Sequencing Consortium, 2004a ; International Chicken Genome Sequencing Consortium, 2004b).
De plus, Ytournel (2008) a montré que la structure du déséquilibre de liaison était différente entre différentes espèces d'élevage, en particulier chez la poule où l'on observe un DL très fort entre marqueurs distants de moins de 5 cM et pas de déséquilibre de liaison entre locus à grande distance (Heifetz et al., 2005). Toutefois, des travaux plus récents ont permis de montrer la présence d'un déséquilibre de liaison à grande distance sur certains chromosomes (Robert et al., 2015). Chez les macro-chromosomes, on observe une plus grande persistance du déséquilibre de liaison avec un r^2 compris entre 0.43 et 0.50 pour une distance entre couple de marqueurs comprise entre 100 et 200kb. Ce r^2 est beaucoup plus variable chez les micro-chromosomes (entre 0.23 et 0.46) (Figure 11). On peut donc observer une étendue du DL en moyenne plus petite chez les micro-chromosomes que chez les macro-chromosomes (Figure 12). Cette structure particulière du déséquilibre de liaison chez les macro et micro-chromosomes suggère l'intérêt de prendre en compte le DL dans la construction des puces de génotypage basse-densité.

3) Objectifs du stage

Au cours du stage, de nombreuses imputations à partir de différentes puces basse densité et de plusieurs scénarii populations ont été réalisées. Deux logiciels d'imputations ont été comparés et six stratégies différentes ont été étudiées afin de regarder leur influence sur l'efficacité de l'imputation. Cette efficacité a été mesurée par le taux d'erreur génotypique, le taux d'erreur allélique, les corrélations entre génotypages haute densité et génotypages imputés. L'efficacité des imputations a aussi été analysée après réalisations des évaluations génomiques à partir des génotypages imputés. L'optimisation des schémas de sélection passant par une minimisation du coût de la sélection et une maximisation de la précision des évaluations génomiques, ces différentes études ont ainsi été menées dans le but de choisir la stratégie de génotypages basse-densité la mieux adaptée à la lignée de poule pondeuse de la société Novogen.

II) Matériels et méthodes

A) Population d'étude

La population d'étude (Figure 13) est constituée de trois générations de coqs et de poules pondeuses issues d'une lignée pure commerciale créée et sélectionnée par la société NOVOGEN du Groupe Grimaud (Le Fœil, Côtes d'Armor).
La première génération (G0) de l'étude est constituée de 437 coqs dont 134 ont été mis en reproduction pour produire la génération G1. Ces animaux ont été élevés en trois lots en Mai 2010, Novembre 2010 et Mai 2011. La deuxième génération (G1) est constituée de 565 coqs dont 125 sont les pères de la génération G2. Ces animaux ont été élevés en trois lots en Novembre 2011, Mai 2012 et Novembre 2012. La troisième génération est constituée de 132 coqs et 635 poules

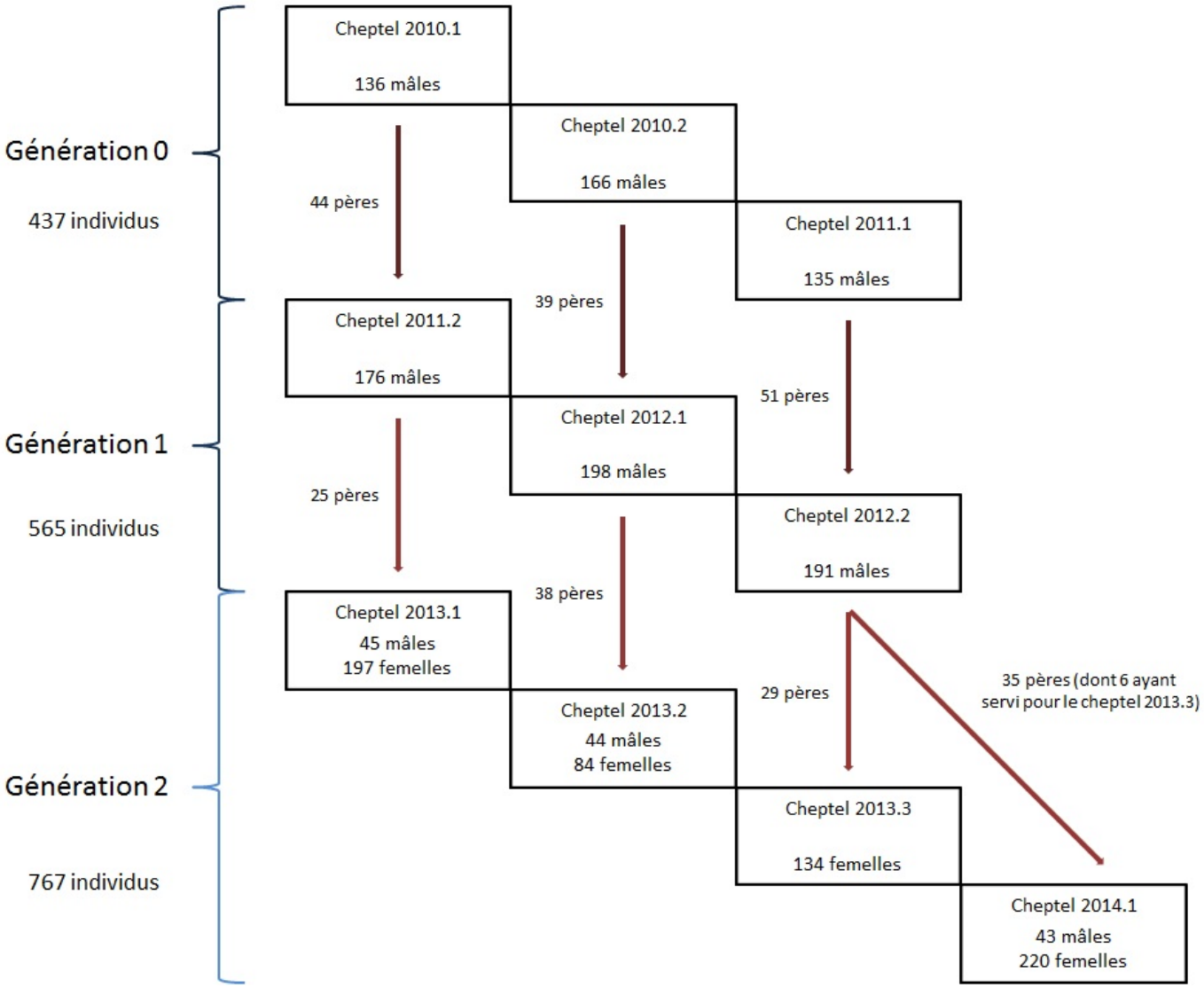


Figure 13 : Organisation en lots et en générations de la population d'étude

pondeuses. Ces animaux ont été élevés en 4 lots en Mai 2013, Novembre 2013, Mars 2014 et Mai 2014. Un bâtiment d'élevage de poules pondeuses du lot 2013.2 ayant pris feu en janvier 2014, un autre lot de poules pondeuses (2013.3) issus de parents 2012.2 a dû être produit en mars 2014. Ce lot a rejoint à 18 semaines le lot 2013.2 en bâtiment cages multiples. Les générations G0 et G1 correspondaient à la population étudiée dans le cadre du projet UtOpIGe.
Après 18 semaines en poussinière, les poules pondeuses ont été élevées en cages multiples de 5 pleines-sœurs. À 60 semaines est intervenue une première sélection sur des index (ou Estimated Breeding Values) calculés à partir de performances de poids et de qualité d'œufs. Les poules ont ensuite été élevées en cage individuelle jusqu'à 85 semaines. Aux alentours de 70 semaines a eu lieu la deuxième étape de sélection sur des EBV calculées sur des performances individuelles de qualité d'œufs et d'intensité de ponte. Les mâles quant à eux ont été sélectionnés après 19 semaines en poussinière sur des GEBV calculées avec des informations sur ascendance et leurs génotypes, puis ont été élevés en cage individuelle jusqu'à 30 semaines où est intervenue la deuxième sélection sur la base des poids individuels des animaux et l'absence de défauts physiques. À l'issue des deuxièmes étapes de sélection, femelles et mâles ont été mis en reproduction afin de produire la génération suivante.

B) Génotypages

Le sang a été prélevé au niveau de la veine brachiale des coqs et des poules pondeuses et l'ADN a été extrait et hybridé sur la puce de génotypage 600K Affymetrix® Axiom® HD par le laboratoire Ark-Genomics (Édimbourg, Royaume-Uni). Au total, 437 coqs pour la génération (G0), 565 coqs pour la génération (G1), et 132 coqs et 635 poules pondeuses pour la génération (G2) ont été génotypés pour 580 961 SNP. D'après la cinquième version de l'assemblage du génome de la poule (NCBI Gallus gallus Annotation Release 103, 2016), ces SNP sont distribués le long des chromosomes 1 à 28, du chromosome 33, d'un groupe de liaison (LGE64), des deux chromosomes sexuels Z et W, ainsi qu'un groupe de 3 724 SNP ayant des positions non précises. En amont du stage, les génotypages ont été filtrés par un contrôle qualité selon six étapes successives :

- 1) 14 SNP sur le chromosome W avec un call rate (pourcentage de génotypes présents dans la population) inférieur à 5% ont été exclus
- 2) Aucun animal ne présentait de call rate inférieur à 95%
- 3) 260 945 SNP ont été exclus du fait d'une MAF inférieure à 0,05
- 4) 9 041 SNP ont été éliminés à cause d'un call rate inférieur à 95%
- 5) 26 318 SNP déviant significativement (P < 5%) de l'équilibre de Hardy-Weinberg ont été supprimés
- 6) 29 SNP doublons et 1686 SNP avec des positions non précises ont été repérés et supprimés

In fine, 282 928 SNP ont été sélectionnés pour les analyses et seront par la suite désignés sous l'appellation 300K.
Dans la génération (G2), 71 individus n'avaient pas de pères G1 génotypés. Ils ont été supprimés pour la suite des études, ce qui a abaissé la population G2 à 124 coqs et 572 pondeuses.
Enfin, 16 incompatibilités pères-descendants ont été repérées. 6 incompatibilités ont pu être corrigées en réassignant le bon père aux descendants, les 10 autres descendants (G1) ont été supprimés.

C) Puces BD

1) Puces construites sur la structure du DL

À partir de la puce HD de 282 928 SNP, sept puces BD ont été simulées en « effaçant » certains SNP de la puce HD. Compte tenu de la structure du déséquilibre de liaison particulière dans le génome avicole, quatre puces BD avec différents seuils de DL ont été créées.

Scénario	Population de référence	Population candidate
Utopige	G0	G1
Population totale	G0 + G1	G2
G1-G2	G1	G2
Saut de génération	G0	G2

Tableau 2 : Organisation des populations de référence et candidate des différents scénarii populations

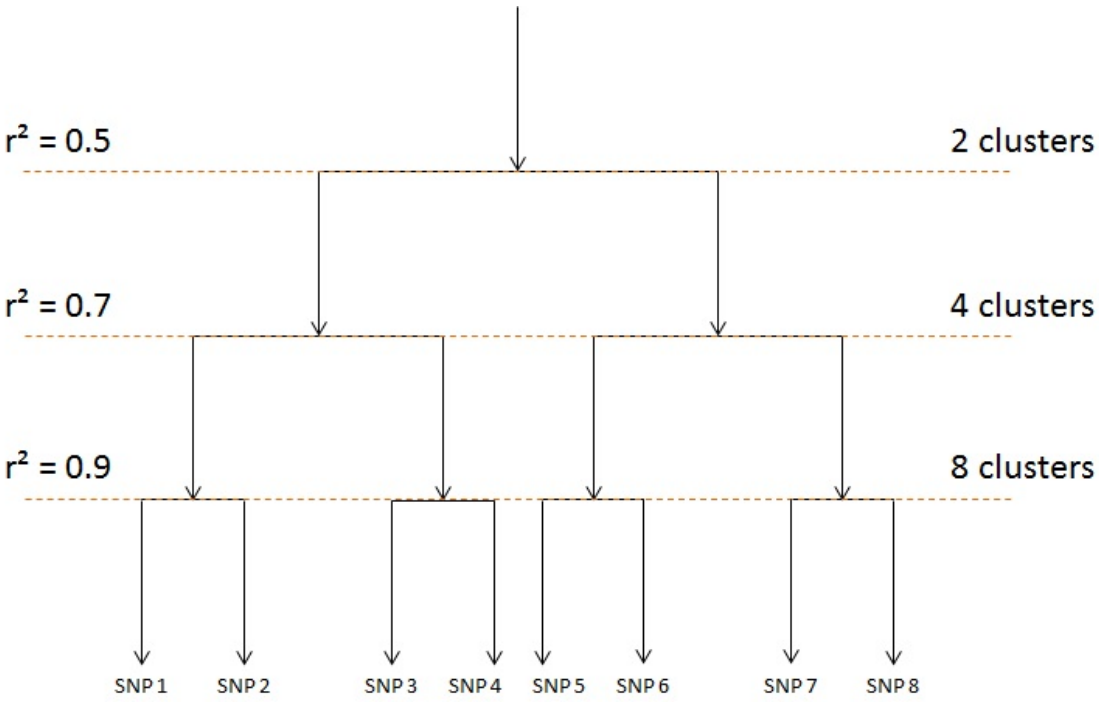


Figure 14 : Évolution du nombre de cluster en fonction du seuil de DL utilisé pour le clustering

Après avoir obtenu la matrice des valeurs de déséquilibre de liaison (r^2) entre SNP sous Plink (Purcell et al., 2007), les données ont été importées sous R (R Core Team, 2015). La sélection des SNP à garder sur la base du DL a été réalisée avec le package R hclust et la méthode « Ward.D » (Murtagh and Legendre, 2011). Ce package permet l'obtention de clusters de SNP selon la méthode Ward qui crée et agrège des clusters en maximisant la variance inter-cluster et en minimisant la variance intra-cluster. Le clustering permet donc de définir des groupes de SNP en très fort DL les uns avec les autres. Le SNP avec la plus grande MAF est ensuite conservé et sert de représentant du cluster. Plus le seuil de DL augmente, plus le nombre de clusters augmente et plus le nombre de SNP sélectionnés sur la puce BD augmente (Figure 14). Quatre puces BD sur la base du DL ont ainsi été créées :

- Puce DL 0.5 avec 9 820 SNP
- Puce DL 0.2 avec 5 224 SNP
- Puce DL 0.1 avec 3 988 SNP
- Puce DL 0.05 avec 3 357 SNP

2) Puce intégrant les QTL

Chez la poule pondeuse, de nombreux QTL ont été mis en évidence et ont des effets sur plusieurs caractères de production d'œufs ou de qualité des œufs. Les SNP marqueurs de QTL ayant une faible MAF (Hickey et al., 2012) ils seront mal imputés. Ceci suggère l'intérêt de prendre en compte les QTL dans l'imputation en incluant les SNP marqueurs des QTL sur la puce BD afin de ne pas se tromper lors de la réalisation des imputations. Romé et al. (2015) ont réalisé une détection de QTL sur la population UtOpIGe. Ces QTL ont un effet fort sur 16 caractères correspondants à des caractères de production d'œufs ou de qualité des œufs. 294 SNP associés aux plus forts effets ont été ajoutés sur la puce DL 0.5 dans le but d'étudier l'intérêt de la prise en compte des QTL dans l'imputation. La cinquième puce contient 10 114 SNP.

3) Puces avec SNP équidistants

Enfin, de nombreux travaux (majoritairement dans les filières non avicoles) ont étudié différentes densités de puces BD en choisissant les SNP à intervalles réguliers le long du génome. Deux puces avec des SNP équidistants ont été réalisées avec des densités de 3K et 10K SNP (notées puces 3Kequi et 10Kequi). En fonction du nombre de SNP désirés, le génome est séparé respectivement en cluster de 290 000 et 100 000 pdb. Le SNP ayant la plus forte MAF et, à MAF équivalente, situé le plus à gauche dans chaque intervalle est choisi comme représentant du cluster. On obtient alors deux puces équidistantes :

- Puce 10Kequi avec 9 352 SNP
- Puce 3Kequi avec 3 337 SNP

D) Scénarii populations

Quatre scénarii ont été mis en place et diffèrent selon la population de référence et la population candidate (Tableau 2). Le premier scénario « Utopige » correspond à la génération G0 comme population de référence et la génération G1 comme population candidate.

Le deuxième scénario « population totale » consiste à intégrer la génération G1 dans la population de référence (G0-G1) et à considérer la génération G2 comme population candidate.

Le troisième scénario « G1-G2 » est similaire au scénario « Utopige » en prenant la génération G1 comme population de référence et la génération G2 comme population candidate.

Enfin le quatrième scénario « saut de génération » correspond à la population G0 comme population de référence et la population G2 comme population candidate.

E) Stratégies étudiées

À partir des 7 puces créées et des 4 scénarii populations retenus, 6 stratégies ont été étudiées afin de regarder leur influence sur l'efficacité de l'imputation (Tableau 3). Les quatre premières stratégies ont été étudiées sur le scénario « Utopige » et concernent le type de puce utilisée.

Scénario	Puce DL 0.5	Puce DL 0.2	Puce DL 0.1	Puce DL 0.05	Puce QTL	Puce 10K équi	Puce 3K équi
Utopige	(1) (2) (3) (4) (5)	(1) (3)	(1) (3)	(1) (3)	(2) (4) (5)	(3) (4) (5)	(3)
Population totale	(5) (6)				(5) (6)	(5) (6)	
G1-G2	(6)				(6)	(6)	
Saut de génération	(6)				(6)	(6)	

Tableau 3 : Tableau croisé des stratégies étudiées en fonction des différentes puces et scénarii étudiés

En bleu : Effet de puces ; En marron : Effet de population
(1) : Étude de l'effet du seuil de DL
(2) : Étude de l'effet des QTL
(3) : Étude de l'effet de la densité de marqueurs
(4) : Étude de l'intérêt du choix des SNP sur la base du DL, DL+QTL, ou de la distance entre SNP
(5) : Étude de l'effet de la taille de la population de référence
(6) : Étude de l'effet des relations de parenté entre population de référence et population candidate

Taux d'erreur génotypique	Taux d'erreur allélique	Corrélation
Pour chaque SNP : $Si\ SNP_{imp}(i) = SNP_{HD}(i),\ a(i) = 0$ $Si\ SNP_{imp}(i) \neq SNP_{HD}(i),\ a(i) = 1$	Pour chaque SNP : $Si\ \lvert SNP_{imp}(i) - SNP_{HD}(i)\rvert = 0,\ a(i) = 0$ $Si\ \lvert SNP_{imp}(i) - SNP_{HD}(i)\rvert = 1,\ a(i) = 0{,}5$ $Si\ \lvert SNP_{imp}(i) - SNP_{HD}(i)\rvert = 2,\ a(i) = 1$	Pour les génotypages HD : SNP (1, 2, …, 282927, 282928) Individu (1, …, n) $\begin{pmatrix} 1 & 2 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 2 & \cdots & 0 & 2 \end{pmatrix}$ On dispose du même type de matrice pour les génotypages imputés
Puis on somme : $\sum_{i=1}^{282\,928} a(i) = nb\ erreur$		Puis on calcule les corrélations entre les deux matrices, colonne par colonne : $corr\ (Vect_{imp}(i), Vect_{HD}(i))$
$T = \frac{nb\ erreur}{(282\,928\ - nb\ SNP_{puce10K}) * nb\ ind_{imputé}}$		

Tableau 4 : Calcul des critères de mesure de l'efficacité de l'imputation

1) Dans la première étude, la densité des SNP sur la puce BD est étudiée à partir des 4 puces créées sur la base du DL ainsi que sur les deux puces avec des SNP équidistants.
2) La deuxième étude correspond à l'influence du seuil de DL sur l'efficacité de l'imputation. Cette étude est menée sur les 4 puces créées sur la base du DL.
3) La troisième étude réalisée avec la puce QTL et la puce DL 0.5 vise à observer l'effet de l'intégration des QTL sur l'efficacité de l'imputation.
4) Enfin, une dernière analyse sur l'efficacité de l'imputation est menée afin d'étudier l'intérêt de choisir les SNP sur la base du DL ou bien sur la distance entre SNP. Une analyse plus approfondie est réalisée en comparant chaque chromosome.

Les deux dernières stratégies concernent des effets de populations.

5) La cinquième étude vise ainsi à analyser l'effet de la taille de la population de référence sur l'imputation. Cette étude est réalisée par comparaison entre le scénario « Utopige » et le scénario « population totale » sur les puces DL 0.5, QTL et 10Kequi.
6) Enfin la dernière étude cherche à observer l'effet des relations de parenté entre population de référence et population candidate. Elle est menée par comparaison entre le scénario « population totale », « G1-G2 » et « saut de génération » sur les puces DL 0.5, QTL et 10Kequi.

F) Logiciels d'imputation étudiés

Deux logiciels d'imputation ont été utilisés et comparés sur les 6 stratégies citées précédemment. Il s'agit de FImpute V2.2 (Sargolzaei et al., 2014) et Beagle V4.1 (Browning and Browning, 2016).
FImpute est un logiciel développé pour les espèces d'élevage et réalise les imputations avec la méthode de la fenêtre glissante chevauchante. Beagle, quant à lui, a été développé sur le modèle humain et utilise la méthode des modèles de Markov cachés. Il a par la suite été adapté aux populations animales avec la possibilité d'intégrer la connaissance d'un pedigree.
FImpute a été utilisé sur l'ensemble des études, et Beagle uniquement sur le scénario Utopige.

G) Mesure de l'efficacité de l'imputation

Quatre critères ont été pris en compte pour étudier l'efficacité des imputations réalisées par les deux logiciels sur les 6 stratégies précédentes (Tableau 4).
Le taux d'erreur génotypique est obtenu par comparaison, pour chaque SNP, des génotypages imputés avec les génotypages HD (300K). Lorsque l'on observe une différence, on considère l'imputation comme fausse. On obtient ensuite le taux d'erreur par SNP en sommant le nombre d'erreurs que l'on divise par le nombre total de SNP imputés (282928-nombre de SNP sur la puce BD) multipliés par le nombre de candidats.
Le taux d'erreur allélique permet d'apporter une précision en regardant les deux allèles au niveau d'un locus. Lorsqu'un allèle sur les deux est bien imputé, on considère que l'on a une « demie-erreur ». Lorsque les deux allèles sont mal imputés, on considère que l'on a une erreur. On obtient ensuite le taux d'erreur en sommant le nombre d'erreur que l'on divise par le nombre total de SNP imputés (282928-nombre de SNP sur la puce BD) multipliés par le nombre de candidats.
Les corrélations entre génotypes imputés et HD étant plus indépendantes de la MAF que les taux d'erreur génotypique et allélique (Dassonneville et al., 2012 ; Hickey et al., 2012), elles ont aussi été calculées, par SNP, avec la méthode de Pearson sur l'ensemble des candidats. Les différences de corrélations obtenues entre puces BD ont été testées par des tests de Student au seuil de première espèce $\alpha = 0.1\%$.
Pour ces trois critères, des moyennes sur l'ensemble des SNP imputés sont ensuite effectuées pour obtenir des résultats à l'échelle de la puce BD.
Enfin, le dernier critère permettant d'étudier l'efficacité des imputations est l'analyse de l'impact sur les évaluations génomiques. Pour cela, une comparaison entre les résultats des GEBV des candidats calculées à partir des génotypages HD (300K) et des GEBV des candidats calculées à partir des génotypages imputés a été réalisée, sur le scénario Utopige en calculant des

corrélations de Pearson et de Spearman. La corrélation de Spearman permet d'évaluer l'éventuel reclassement des candidats selon leur GEBV et a été calculée sur les 150 candidats ayant la plus grande GEBV à partir des génotypages HD 300K. La significativité des différences de corrélations obtenues avec les différentes puces BD a été testée sur R avec la fonction paired.r du package Psych (Revelle, 2016). Cette fonction transforme les corrélations en z score et teste la significativité des z scores avec un test de Student (le risque α = 5% a été retenu).

H) Étude des évaluations génomiques

Les évaluations génomiques ont été menées avec un BLUP Single Step (Legarra et al., 2009) afin d'étudier l'impact de l'imputation sur le calcul des GEBV de trois caractères présentant des déterminismes génétiques différents (Romé et al., 2015) : l'intensité de ponte (IP), le poids d'œuf (PO) et la couleur des œufs (LAB). La méthode du Single Step permet de combiner des informations hétérogènes au sein de la population d'étude, certains animaux ne disposant que d'informations de performances (phénotypages), d'autres d'informations de génotypages, et d'autres de phénotypages et de génotypages. Cela résulte dans la création d'une matrice H. Elle correspond à l'extension de la matrice de parenté « classique » A, obtenue d'après le pedigree, et de la matrice de parenté génomique G à l'ensemble des individus génotypés et/ou phénotypés.

$$H = \begin{pmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{pmatrix}$$

Les indices 1 et 2 correspondent respectivement aux animaux non génotypés et génotypés.
Un modèle « animal » avec des effets fixes est utilisé. Les effets fixes sont le lot, la batterie intra-lot, la longueur de la batterie, la date d'éclosion intra-lot, l'âge réel des animaux (pour l'intensité de ponte et le poids d'œuf) et la différence de jours entre jour de ponte de l'œuf et jour de pesée de l'œuf (pour le poids d'œufs). L'effet animal est aléatoire.
Le modèle est ajusté avec l'utilisation du programme Renumf90 (Misztal et al., 2002) qui prépare tous les fichiers pour Remlf90 (Misztal et al., 2002) qui a ensuite réalisé l'estimation des paramètres génétiques et les évaluations génomiques.

III) Résultats et discussion

A) Comparaison des mesures d'efficacité de l'imputation

Une comparaison des taux d'erreurs alléliques et génotypiques (Figure 15), ainsi que des corrélations (Figure 16) sur les puces DL 0.5, QTL et 10Kequi a été réalisée sur le scénario Utopige. On constate pour les trois puces que le taux d'erreur génotypique est 1.98 fois supérieur au taux d'erreur allélique. Quant aux corrélations, les valeurs associées aux 3 puces sont très proches et se révèlent peu discriminantes. Dans le détail, on observe que le taux d'erreur génotypique de la puce 10Kequi (3.08%) est supérieur au taux d'erreur génotypique de la puce DL 0.5 (2.36%) qui est lui-même supérieur au taux d'erreur génotypique de la puce QTL (2.31%). La même observation est faite pour le taux d'erreur allélique (1.56% > 1.19% > 1.16%). En effet, la majorité des erreurs d'imputation (98%) n'impacte qu'un seul des 2 allèles du génotype (par exemple, une imputation AG alors que le génotypage HD est AA). Dans le comptage des erreurs alléliques un problème au niveau d'un seul allèle vaut une demi-erreur. Cela explique donc que le taux d'erreur génotypique soit quasiment le double du taux allélique (1.98).
Concernant les corrélations avec les génotypages HD, on observe une moins bonne corrélation pour la puce 10Kequi (0.9734) qu'avec la puce DL 0.5 (0.9801) qui a elle-même une corrélation moins bonne que la puce QTL (0.9807). Ces différences sont bien significatives selon un test de Student sur les différences des moyennes des corrélations, au seuil de première espèce α = 0.1%, avec des p-value toutes inférieures à 10^{-13}, du fait notamment de la grande taille de la population candidate (565 individus). Une moins bonne corrélation est synonyme de plus d'erreurs d'imputation et de taux d'erreur génotypique et allélique plus élevés. On constate donc que les trois critères de mesure d'efficacité de l'imputation aboutissent aux mêmes conclusions, à savoir

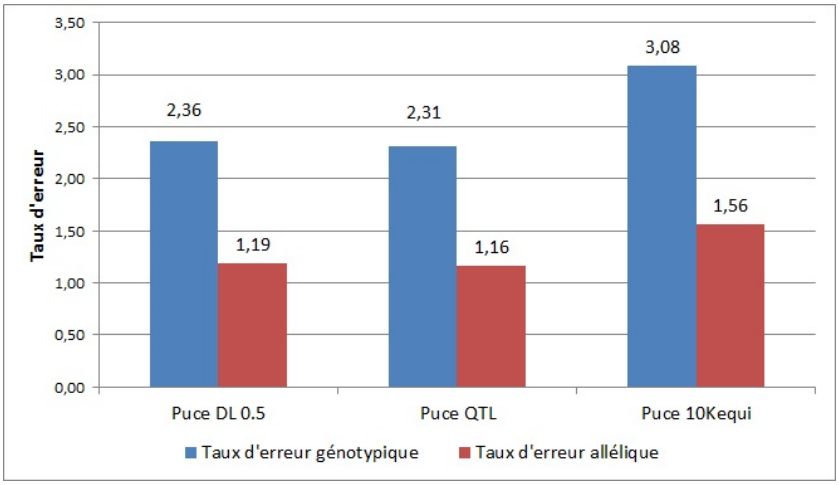


Figure 15 : Comparaison des taux d'erreurs génotypiques et alléliques sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi

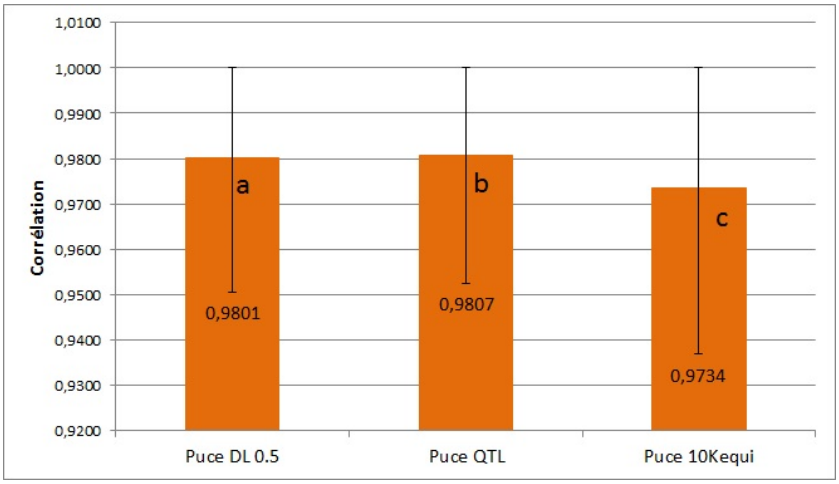


Figure 16 : Comparaison des corrélations entre génotypages imputés et génotypages HD sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi

Les lettres a, b et c indiquent des différences de moyennes de corrélations significatives d'après des tests de Student au seuil de première espèce α = 0.1%

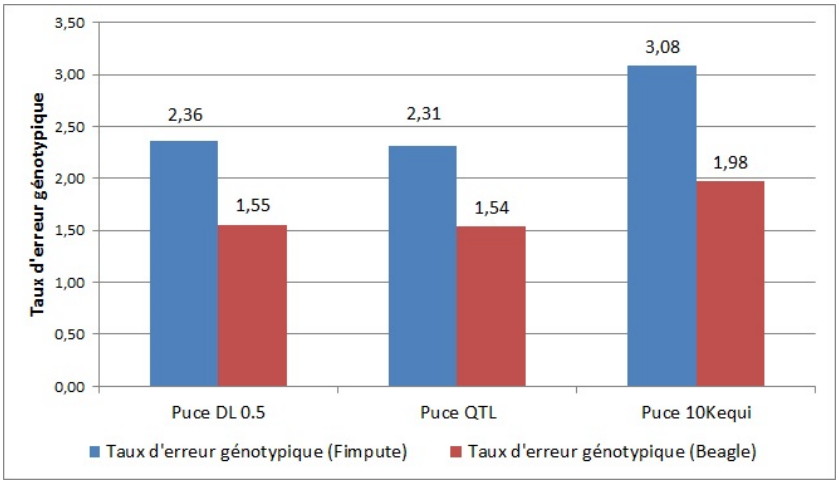


Figure 17 : Comparaison des taux d'erreurs génotypiques obtenue avec les logiciels FImpute et Beagle sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi

une meilleure imputation à partir de la puce QTL.
Ces trois critères sont utilisés dans la littérature mais ont des particularités. La corrélation peut permettre de présenter des bons résultats avec des valeurs très élevées et très proches, et n'est donc pas un critère discriminant. Le taux d'erreur allélique peut être vu comme une manière de présenter de meilleurs résultats avec des taux d'erreur plus faibles que le taux d'erreur génotypique (Dassonneville, 2012). De plus, nous pouvons considérer que lorsque l'imputation n'est pas bonne au niveau d'un des deux allèles, l'erreur d'imputation pourra avoir des conséquences lourdes si elle concerne un SNP avec un effet fort. On préfère alors considérer l'imputation comme mauvaise si au moins un des deux allèles est mal imputé. C'est pourquoi dans la suite des résultats nous n'utiliserons que le taux d'erreur génotypique pour mesurer l'efficacité des imputations réalisées.

B) Comparaison de FImpute et Beagle

Une comparaison des imputations réalisées sur le scénario Utopige à partir des puces DL 0.5, QTL et 10Kequi avec les logiciels FImpute et Beagle a été menée (Figure 17). On note, pour l'ensemble des trois puces, que le taux d'erreur génotypique est à chaque fois plus faible avec le logiciel Beagle qui réalise donc de meilleures imputations que FImpute. Lorsque l'on regarde le détail des imputations avec Beagle sur les trois puces, on constate à nouveau que le taux d'erreur génotypique est plus faible pour la puce QTL (1.54%) que pour la puce DL 0.5 (1.55%) que pour la puce 10Kequi (1.98%). Cependant, bien que l'imputation réalisée avec Beagle soit meilleure que celle avec FImpute, pour aboutir aux mêmes conclusions, le temps nécessaire pour réaliser les imputations est bien différent. FImpute réalise pour la puce QTL une imputation de 272 814 SNP sur 565 individus en 3 minutes et 29 secondes, contre 24 heures et 2 minutes avec Beagle. Ces résultats sont en accord avec la littérature. Par exemple, pour une imputation de 6 000 SNP vers 50 000 SNP et pour 2000 individus candidats et 10 000 individus référence, Sargolzaei et al. (2014) ont obtenu une imputation complète avec FImpute au bout de 3 minutes contre plus de 15 heures avec Beagle. De même, Ventura et al . (2014) ont montré en bovins que l'imputation avec FImpute d'une puce 6 000 SNP vers une puce 50 000 SNP, à partir de 4 886 animaux de référence et 146 candidats durait 5 minutes et 53 secondes versus 7 heures 35 minutes et 43 secondes avec Beagle. L'utilisation de Beagle pour mener des travaux de recherche, sur l'optimisation des imputations par exemple, peut être réalisable. En revanche, une utilisation en routine, lors d'évaluations génomiques par exemple, sera beaucoup plus compliquée à cause du temps de calcul et le gain en terme d'efficacité ne justifie pas une telle contrainte. C'est pourquoi dans la suite des résultats nous nous concentrerons uniquement sur le logiciel FImpute.

C) Influence de la densité de marqueurs

La figure 18 illustre, sur le scénario Utopige, l'évolution du taux d'erreur génotypique en fonction du nombre de SNP présents sur la puce BD. On constate une augmentation du taux d'erreur à mesure que l'on diminue le nombre de SNP sur puce BD. Ceci est valable pour les puces avec des SNP choisis selon la base du déséquilibre de liaison ou bien selon la distance entre SNP. En effet, avec des SNP choisis à partir du DL, pour 3 357 SNP et 9 820 SNP les taux d'erreur sont respectivement de 6.00% et 2.36%. De même avec des SNP équidistants, pour 3 337 SNP et 9 352 SNP, les taux d'erreur sont respectivement de 7.18% et 3.08%. Ces résultats sont en accord avec la littérature (Dassonneville et al., 2012 ; Carvalheiro et al., 2014) où de meilleures imputations sont réalisées avec l'augmentation du nombre de SNP sur la puce BD. Ceci s'explique par le fait qu'avec un plus grand nombre de SNP sur puce BD, on augmente le nombre de points ancrants permettant d'identifier les haplotypes de référence correspondants, et on diminue la probabilité d'identifier par hasard des haplotypes en commun entre la population de référence et la population cible.

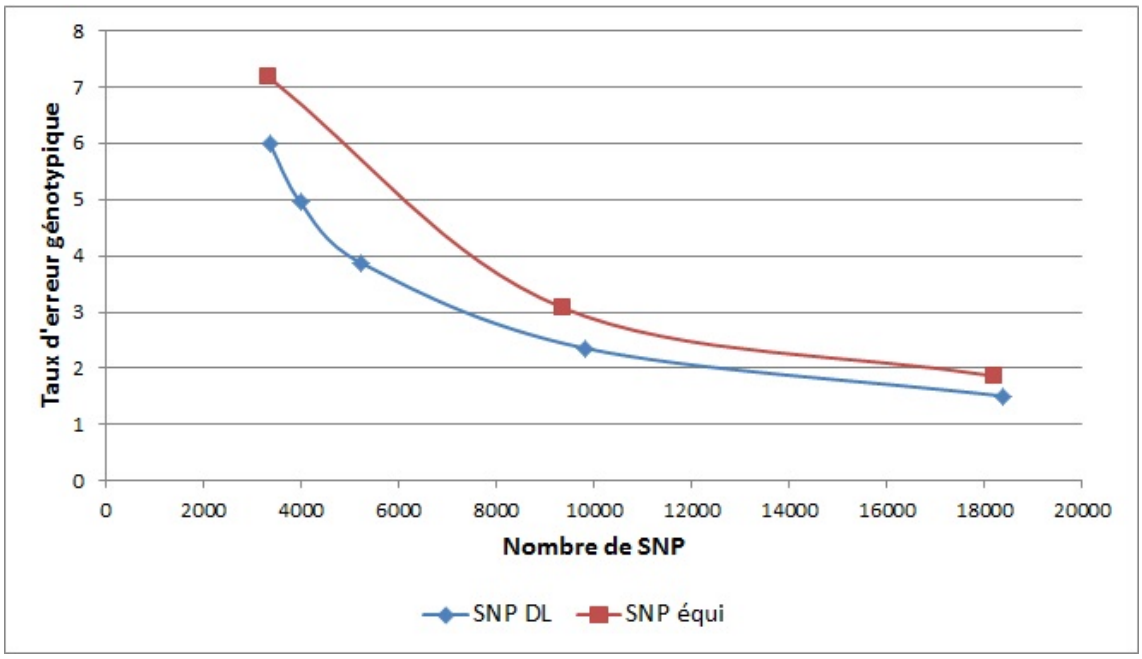


Figure 18 : Évolution du taux d'erreur génotypique en fonction du nombre de SNP pour des puces basées sur le seuil du DL ou bien sur la distance entre SNP

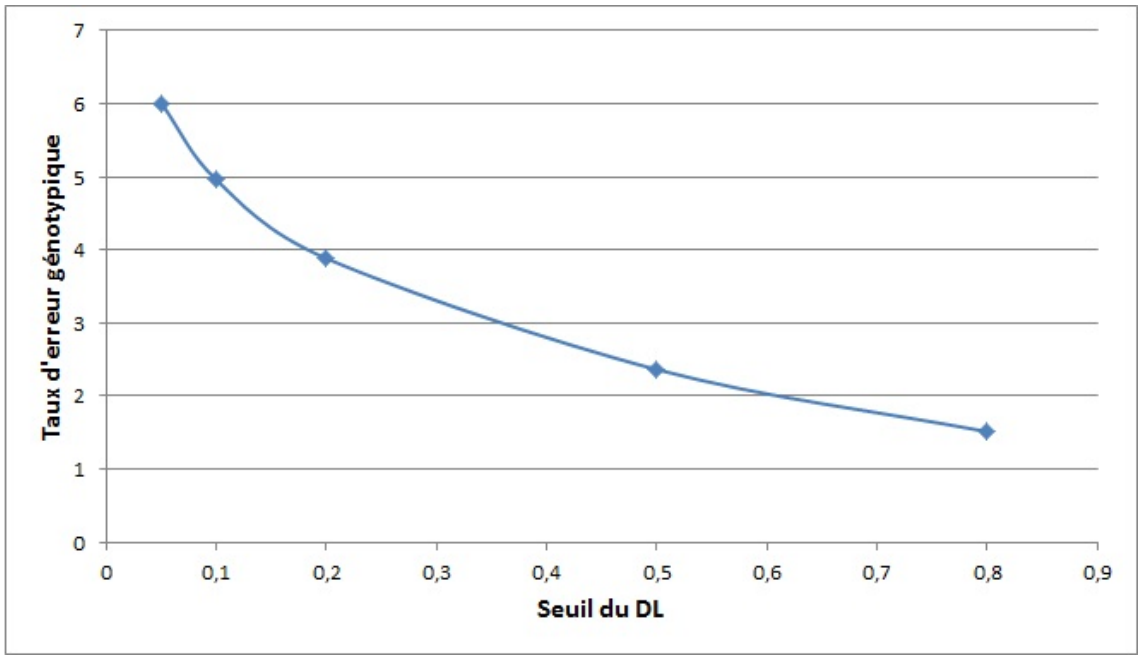


Figure 19 : Évolution du taux d'erreur génotypique en fonction du seuil de DL utilisé pour la construction des puces BD basées sur le seuil de DL

D) Influence du seuil de déséquilibre de liaison

La figure 19 montre que le taux d'erreur génotypique diminue avec une augmentation du seuil de DL utilisé pour le clustering. En effet, pour les puces DL 0.05, DL 0.2 et DL 0.5, on obtient respectivement un taux d'erreur de 6.00%, 3.88% et 2.36%. On augmente donc l'efficacité de l'imputation en augmentant le seuil de DL lors du clustering. Ceci s'explique par la façon dont le clustering se fait (Figure 14). Lorsque le seuil de DL pour le clustering est très élevé (0.9 par exemple), le nombre de clusters est élevé car on ne va regrouper dans un même cluster que des paires de SNP en très fort DL. Il y aura alors beaucoup de clusters créés et un nombre élevé de SNP sur la puce BD. Lorsque le seuil de DL pour le clustering diminue (0.7 par exemple), on peut agréger des clusters de paires de SNP avec des DL plus faibles. On diminue alors le nombre de clusters, ce qui diminue le nombre de SNP à inclure sur la puce BD. Or, comme nous l'avons vu précédemment, plus le nombre de marqueurs sur les puces BD est élevé, plus l'imputation sera bonne. De plus, en diminuant le DL, on choisit un SNP faiblement « associé » aux autres SNP du cluster, et donc un moins bon SNP pour l'imputation.

E) Influence des QTL

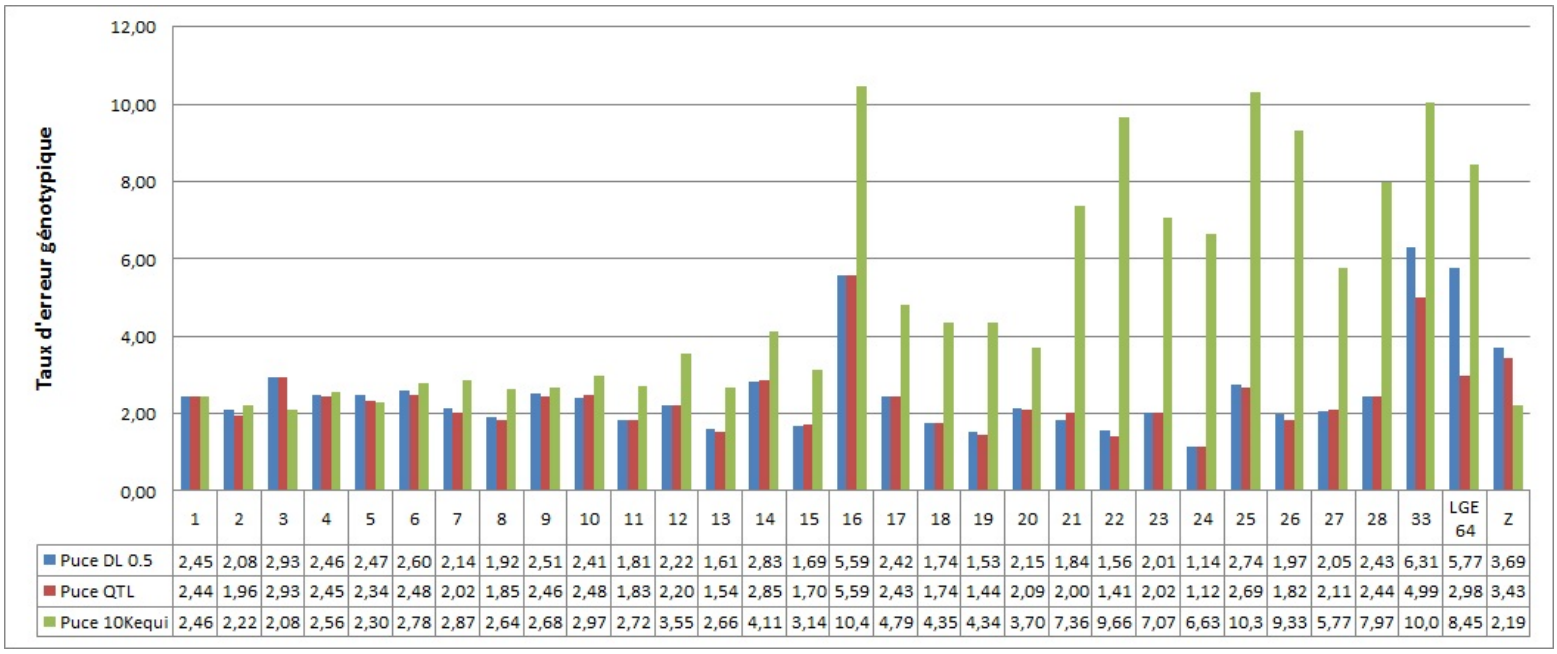
Pour tester l'influence des QTL sur l'imputation, une puce QTL a été créée. Elle correspond à la puce DL 0.5 sur laquelle 294 SNP associés aux QTL ayant les plus forts effets sur plusieurs caractères de production d'œufs ou de qualité des œufs ont été ajoutés. D'après la Figure 15, on constate que le taux d'erreur de la puce DL 0.5 est de 2.36% et celui de la puce QTL est de 2.31%. On augmente donc légèrement l'efficacité de l'imputation en rajoutant les QTL. Cette diminution du taux d'erreur peut s'expliquer par deux points entre lesquels il peut y avoir une confusion au niveau des effets. La premier point est qu'en rajoutant les SNP marqueurs des QTL sur la puce DL 0.5, on augmente le nombre de SNP présents sur la puce BD. Comme vu précédemment, une augmentation du nombre de marqueurs sur la puce BD se traduit par une diminution du taux d'erreur. Le deuxième point développé dans la littérature (Hozé et al., 2013 ; Carvalheiro et al., 2014, Heidaritabar et al., 2014 ; Heidaritabar et al., 2015) est que les SNP marqueurs des QTL ont une faible MAF et seront donc moins bien imputés si on ne retrouve pas le bon haplotype dans la librairie de référence. De plus, si l'imputation de ces SNP avec une faible MAF n'est pas bonne, cela peut entraîner de mauvaises estimations des évaluations génomiques à cause des effets forts qui leurs sont associés. En les incluant sur la puce DL 0.5, on évite la possibilité de se tromper dans l'imputation de ces SNP. L'imputation de l'ensemble des SNP sera donc meilleure.

F) Intérêt de choisir les SNP sur la base du DL ou sur la distance entre SNP

Une comparaison des puces DL 0.5, QTL et 10Kequi sur le scénario Utopige (Figure 15) permet de juger de l'intérêt de choisir les SNP sur la base du DL ou sur la notion de distance entre SNP. Lorsque l'on regarde les résultats d'imputation de la puce DL 0.5 avec la puce 10Kequi, on constate que le taux d'erreur est plus élevé avec la puce 10Kequi (3.08%) qu'avec la puce DL 0.5 (2.36%). Cette différence pourrait être due au nombre de SNP qui est inférieur sur la puce 10Kequi, la puce DL 0.5 ayant 468 SNP en plus.

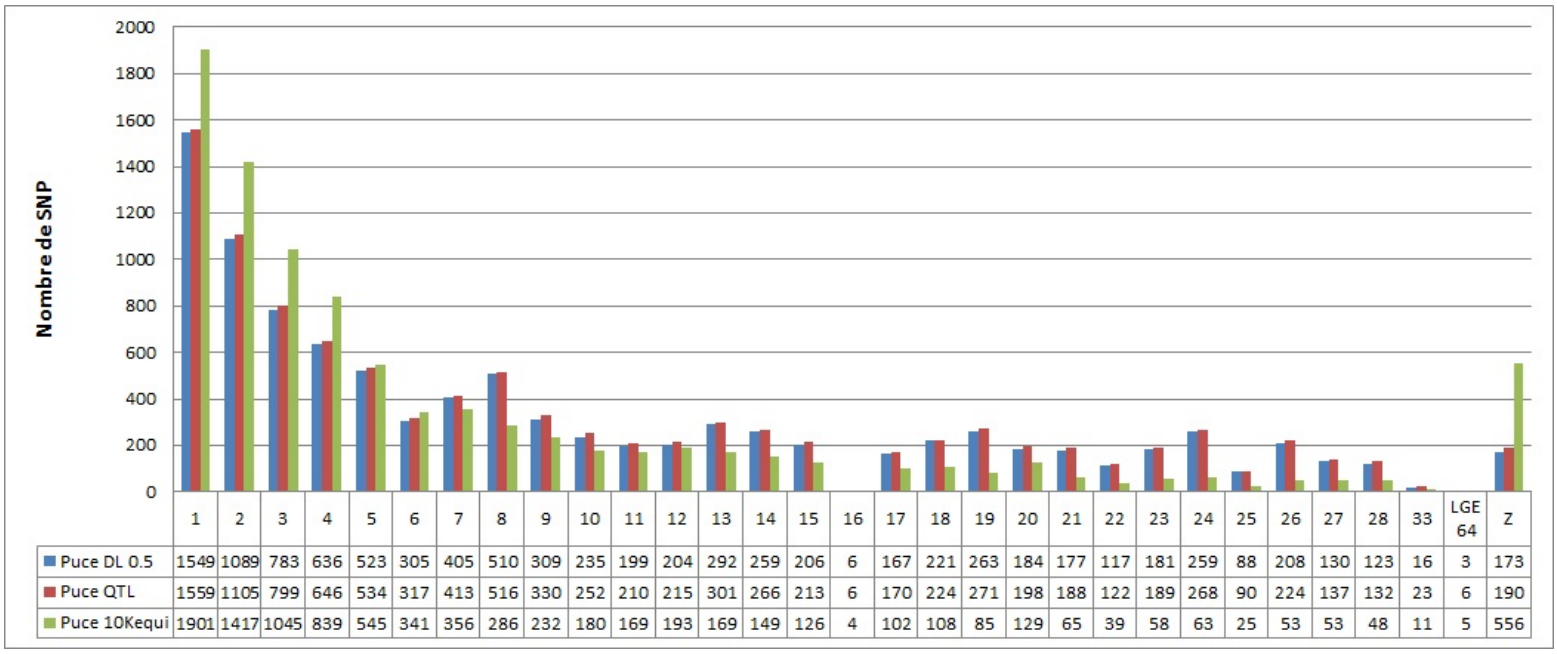
Une deuxième étude a donc été menée en créant une puce DL 0.05 avec 3 357 SNP et une puce 3Kequi avec 3 337 SNP. On observe à nouveau que les résultats d'imputations sont meilleurs avec la puce DL 0.05 qui présente un taux d'erreur de 6.00%, qu'avec la puce 3Kequi avec un taux d'erreur de 7.18%. Ainsi, pour réaliser une bonne imputation, il est préférable de choisir les SNP sur la base du DL plutôt que sur la distance entre SNP. Ce résultat est cohérent avec ceux trouvés dans la littérature (Carvalheiro et al., 2014). Ceci est à mettre en lien avec la structure particulière du DL chez les poules pondeuses, c'est pourquoi nous avons étudié sur le scénario Utopige l'évolution du taux d'erreur génotypique en fonction de chaque chromosome imputé à partir des puces DL 0.5, QTL et 10Kequi (Figure 20).

Pour les 6 premiers chromosomes, les taux d'erreurs génotypiques sur les puces DL 0.5, QTL et



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	33	LGE 64	Z
Puce DL 0.5	2,45	2,08	2,93	2,46	2,47	2,60	2,14	1,92	2,51	2,41	1,81	2,22	1,61	2,83	1,69	5,59	2,42	1,74	1,53	2,15	1,84	1,56	2,01	1,14	2,74	1,97	2,05	2,43	6,31	5,77	3,69
Puce QTL	2,44	1,96	2,93	2,45	2,34	2,48	2,02	1,85	2,46	2,48	1,83	2,20	1,54	2,85	1,70	5,59	2,43	1,74	1,44	2,09	2,00	1,41	2,02	1,12	2,69	1,82	2,11	2,44	4,99	2,98	3,43
Puce 10Kequi	2,46	2,22	2,08	2,56	2,30	2,78	2,87	2,64	2,68	2,97	2,72	3,55	2,66	4,11	3,14	10,4	4,79	4,35	4,34	3,70	7,36	9,66	7,07	6,63	10,3	9,33	5,77	7,97	10,0	8,45	2,19

Figure 20 : Évolution du taux d'erreur génotypique en fonction des chromosomes sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	33	LGE 64	Z
Puce DL 0.5	1549	1089	783	636	523	305	405	510	309	235	199	204	292	259	206	6	167	221	263	184	177	117	181	259	88	208	130	123	16	3	173
Puce QTL	1559	1105	799	646	534	317	413	516	330	252	210	215	301	266	213	6	170	224	271	198	188	122	189	268	90	224	137	132	23	6	190
Puce 10Kequi	1901	1417	1045	839	545	341	356	286	232	180	169	193	169	149	126	4	102	108	85	129	65	39	58	63	25	53	53	48	11	5	556

Figure 21 : Évolution du nombre de SNP par chromosome retenus sur les puces DL 0.5, QTL et 10Kequi sur le scénario Utopige

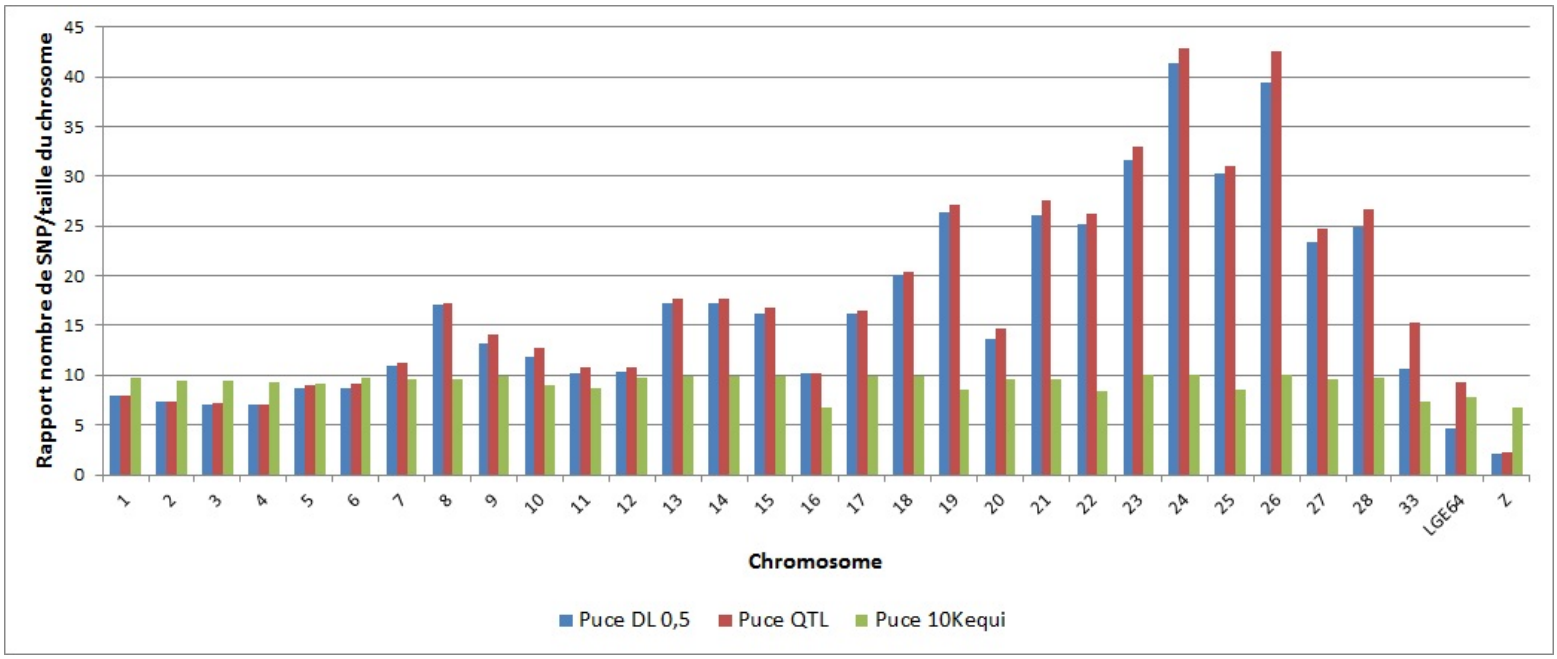


Figure 22 : Évolution du rapport (nombre de SNP/taille du chromosome) en fonction des chromosomes sur le scénario Utopige pour les puces DL 0.5, QTL et 10Kequi

10Kequi sont proches. Or lorsque l'on regarde le nombre de SNP des six premiers chromosomes présents sur les puces BD (Figure 21), le nombre de SNP entre puces basées sur le DL et puce 10Kequi est différent. Par exemple, par rapport à la puce QTL, la puce 10Kequi présente 342 SNP de plus sur le chromosome 1 et 312 SNP de plus sur le chromosome 2.
On note ensuite une forte augmentation du taux d'erreur sur la puce 10Kequi par rapport au taux d'erreur sur la puce DL 0.5 et QTL à partir du 7ème chromosome. Ce décrochement s'accentue ensuite à partir du chromosome 16. Quand on met en lien ces observations avec le nombre de SNP de chaque chromosome présent sur les puces BD (Figure 21), on constate qu'à partir du 7ème chromosome, il y a beaucoup plus de SNP sur les puces tenant compte du DL que sur la puce 10Kequi. Le décrochement du taux d'erreur par rapport aux puces DL 0.5 et QTL peut alors être dû au nombre insuffisant de SNP présents sur les micro-chromosomes pour la puce 10Kequi.
On peut ainsi noter ici tout l'intérêt d'utiliser le déséquilibre de liaison pour créer des puces BD. En utilisant une méthode basée sur la distance entre SNP, on ne tient pas compte de la structure du génome avicole et une fois la distance entre SNP choisie, on conserve sur la puce BD un nombre de SNP proportionnel à la taille du chromosome. Avec une méthode basée sur le DL, on tient compte de la structure particulière du DL chez la poule pondeuse et le nombre de SNP retenus sur la puce BD n'est pas proportionnel à la taille des chromosomes comme le montre la Figure 22. Pour un seuil de DL fixé, on constate d'après la figure 12 une persistance du DL plus forte chez les macro-chromosomes que chez les micro-chromosomes. Comme la persistance est forte pour un macro-chromosome, il faudra en proportion peu de SNP sur la puce BD pour couvrir tout ce macro-chromosome. À l'inverse, la persistance du DL étant peu élevée pour les micro-chromosomes, il faut en proportion beaucoup de SNP sur la puce BD pour couvrir tous les micro-chromosomes. Cela explique alors la densification observée des SNP pour les micro-chromosomes sur les puces basées sur le DL.
En mettant toutes ces observations en lien avec celles faites sur la densité de marqueurs sur puces BD, on peut comprendre pourquoi le taux d'erreur génotypique est faible pour les micro-chromosomes lorsque les SNP sont choisis selon la base du DL. En rajoutant en plus sur chaque chromosome les marqueurs SNP des QTL, on peut diminuer encore plus le taux d'erreur génotypique sur chaque chromosome. En effet, comme expliqué précédemment, on augmente à nouveau le nombre de SNP sur la puce BD et ces SNP marqueurs des QTL ont une faible MAF et seront mal imputés si on ne les inclut pas sur la puce BD.
Ainsi, un choix des SNP sur la base du DL et en tenant compte des QTL plutôt que sur la distance entre SNP permet donc d'optimiser le nombre de SNP sur les macro-chromosomes et de densifier les SNP sur les micro-chromosomes afin d'avoir des taux d'erreurs faibles.

G) Effet de la taille de la population de référence

L'effet de la taille de la population a été étudié en comparant les scénarii Utopige et Population totale sur les puces DL 0.5, QTL et 10Kequi (Figure 23). Le scénario Utopige est constitué des 437 coqs G0 comme population de référence et des 565 coqs G1 comme population candidate. Dans le scénario Population totale, la population de référence est constituée des 437 coqs G0 et 565 coqs G1. La population candidate correspond aux 124 coqs et 572 pondeuses en G2. On augmente donc la taille de la population de référence dans le scénario Population totale.
Pour la puce DL 0.5, le taux d'erreur sur le scénario Utopige est de 2.36% et de 2.03% pour le scénario Population totale. On observe la même diminution du taux d'erreur sur la puce QTL avec un passage d'un taux d'erreur de 2.31% sur le scénario Utopige à un taux d'erreur de 2.00% sur le scénario Population totale. Enfin, pour la puce 10Kequi, on observe une diminution plus prononcée du taux d'erreur qui passe de 3.08% à 2.33%. Par conséquent, pour les 3 puces, le taux d'erreur est inférieur avec le scénario Population totale qu'avec le scénario Utopige, ce qui peut s'expliquer par une population de référence plus grande en utilisant la population totale. En effet, en augmentant la taille de la population de référence, on augmente la taille de la librairie d'haplotypes

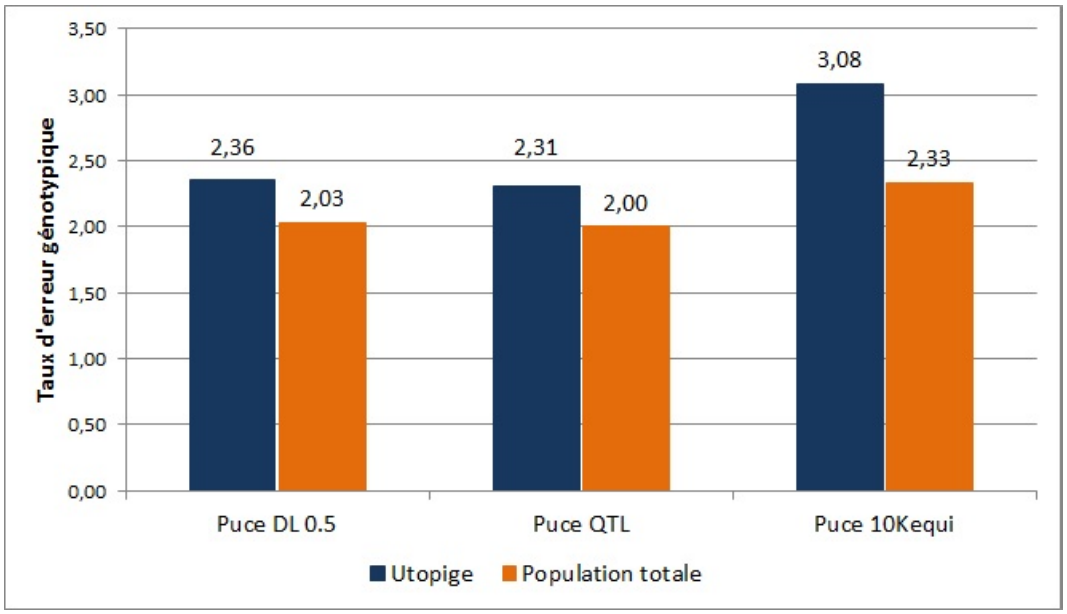


Figure 23 : Évolution du taux d'erreur génotypique sur les puces DL 0.5, QTL et 10Kequi avec une augmentation de la taille de la population de référence

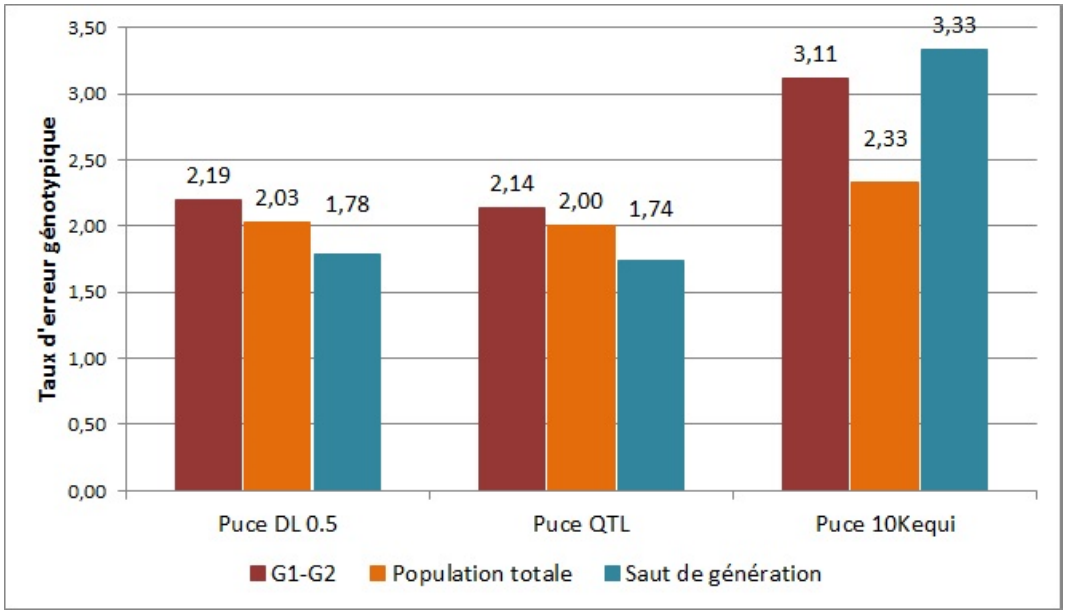


Figure 24 : Évolution du taux d'erreur génotypique sur les puces DL 0.5, QTL et 10Kequi en fonction du degré de parenté observé entre population de référence et population candidate

de référence. La probabilité de retrouver des fragments d'haplotypes de la population cible dans la librairie construite sera alors plus grande. Cela explique pourquoi en augmentant la taille de la population de référence, on diminue le taux d'erreur génotypique sur chaque puce. Ces résultats sont concordants avec ceux présents dans la littérature (Heidaritibar et al., 2014 ; Ventura et al., 2014 ; Heidaritibar et al., 2015).
Par ailleurs, le taux d'erreur sur les trois puces évolue sur le scénario Population totale comme sur le scénario Utopige. On obtient de meilleurs résultats d'imputations avec des puces basées sur le DL plutôt que sur la distance entre SNP. En rajoutant les marqueurs SNP des QTL, on peut diminuer encore plus le taux d'erreur génotypique par rapport à la puce DL 0.5, ce qui est cohérent avec les résultats précédents.

H) Effet des relations de parenté entre population de référence et population cible

Les relations de parenté entre population de référence et population cible ont aussi pu être étudiées sur les puces DL 0.5, QTL et 10Kequi en comparant les scénarii G1-G2, Population totale et Saut de génération (Figure 24).
Le scénario G1-G2 correspond à la population G1 comme population de référence et la population G2 comme candidate. Dans le scénario Population totale, on rajoute la population G0 dans la population de référence, ce qui permet d'augmenter les relations de parenté entre population de référence et population cible. Enfin, pour diminuer les relations de parenté par rapport au scénario G1-G2, le scénario Saut de population a été mis en place en considérant uniquement la population G0 comme la population de référence et en conservant comme population cible la population G2.
Pour la puce DL 0.5 sur le scénario G1-G2, le taux d'erreur est de 2.19%. En augmentant les relations de parenté avec le scénario Population totale, le taux d'erreur diminue à 2.03%. Enfin, avec le scénario Saut de génération et donc en diminuant les relations de parenté par rapport au scénario G1-G2, on obtient un taux d'erreur de 1.78%. Sur la puce QTL, l'évolution sur ces trois scénarii est la même avec des taux d'erreurs respectifs de 2.14%, 2.00% et 1.74%. Enfin, pour la puce 10Kequi, le taux d'erreur sur le scénario G1-G2 est de 3.11%, de 2.33% pour le scénario Population totale et de 3.33% sur le scénario Saut de génération.
Lorsque l'on compare le scénario G1-G2 avec le scénario Population totale, on constate donc pour les trois puces une amélioration du taux d'erreur génotypique. Or, en passant du scénario G1-G2 au scénario Population totale, on a rajouté la génération G0 dans la population de référence. Cette génération correspond aux pères de la G1 qui elle-même correspond aux pères de la G2. En rajoutant G0 dans la population de référence, on augmente donc les relations de parenté entre la population de référence et la population candidate. Ce résultat est cohérent avec ceux notés dans la littérature (Dassonneville et al., 2011 ; Hayes et al., 2011 ; Hozé et al., 2013 ; Bouquet et al., 2015). En effet, plus les relations de parenté entre population de référence et population cible sont proches, plus les individus ont en commun des fragments d'haplotype de grande taille. À l'inverse, avec une diminution des relations de parenté, les fragments d'haplotype en commun se réduisent à cause des recombinaisons au fil des générations. Les individus auront alors en commun des fragments d'haplotypes de plus petites tailles. Ainsi, plus les fragments d'haplotypes sont de grandes tailles, plus la probabilité d'identifier par hasard un mauvais fragment d'haplotype sera faible et plus l'imputation sera bonne. Et plus la taille des fragments d'haplotypes diminue, plus la probabilité d'identifier par hasard un mauvais fragment d'haplotype augmente et plus l'imputation sera mauvaise.
On s'attend donc à ce qu'une diminution des relations de parenté entre population de référence et population cible (Saut de génération) entraîne une augmentation du taux d'erreur. Cette augmentation est observée pour la puce 10Kequi où on note une augmentation de 0.22% par rapport au scénario G1-G2. En revanche pour les puces DL 0.5 et QTL, on note une amélioration du taux d'erreur de 0.41% et de 0.40% respectivement ! Toutefois en filière de ponte, Heidaritabar

Scénario G0G1					
	Nombre de SNP	Taux d'erreur génotypique	Intensité de ponte	Couleur LAB	Poids d'œuf
Puce DL 0.5	9820	2,36	0,9960[a]	0,9945[a]	0,9966[a]
Puce QTL	10114	2,31	0,9960[a]	0,9945[a]	0,9964[a]
Puce 10Kequi	**9352**	**3,08**	**0,9967[b]**	**0,9967[b]**	**0,9971[b]**
Puce DL 0.2	5224	3,88	0,9903[c]	0,9912[c]	0,9918[c]
Puce DL 0.1	3988	4,95	0,9861[d]	0,9861[d]	0,9871[d]
Puce DL 0.05	3357	6,00	0,9804[e]	0,9820[e]	0,9840[e]
Puce 3Kequi	3337	7,18	0,9832[f]	0,9859[f]	0,9860[f]

Tableau 5 : Corrélations de Pearson entre les GEBV des 565 candidats calculées à partir des génotypages HD (300K) et les GEBV des candidats calculées à partir des génotypages imputés, pour les 7 puces étudiées sur le scénario Utopige

Les lettres différentes entre puces et par caractère indiquent, après transformation des corrélations en z scores par la fonction r.paired du package Psych (Revelle, 2016), des différences de z scores significatives selon des tests de Student au seuil de première espèce α = 5%

et al. (2014) ont montré qu'il n'y avait pas de dégradation de l'efficacité de l'imputation en diminuant les relations de parenté. C'est bien ce que l'on observe ici. Une explication donnée par Heidaritabar et al. (2014) est que la persistance du DL à travers les générations est très forte avec une corrélation de 0.93 entre le DL de leur génération G0 et leur génération G2. Par manque de temps, nous n'avons pu étudier la persistance du DL à travers nos trois générations. C'est une étude qui pourra être menée à la suite de ce stage afin de confirmer l'hypothèse.
Enfin, comparativement à la puce DL 0.5, la puce QTL présente de meilleurs résultats d'imputations pour les mêmes raisons que précédemment.

I) Impact sur les évaluations génomiques

Après avoir étudié les stratégies pouvant influencer l'efficacité de l'imputation, nous avons étudié les impacts des erreurs d'imputation sur la précision des évaluations génomiques de trois caractères présentant des déterminismes génétiques différents (Romé et al., 2015): l'intensité de ponte, la couleur des œufs et le poids d'œufs. Ces études ont été menées sur le scénario Utopige et sur les 7 puces étudiées. Pour cela, nous avons calculé des corrélations de Pearson et de Spearman pour réaliser une comparaison entre les GEBV des 565 candidats calculées à partir des génotypages HD (300K) et les GEBV des candidats calculées à partir des génotypages imputés.
Lorsque l'on regarde les corrélations de Pearson (Tableau 5), on constate que les valeurs obtenues sont très élevées et toutes supérieures à 0,98, signe d'une très bonne précision des évaluations génomiques avec toutes les puces. Dans le détail, on peut néanmoins observer des différences significatives (excepté pour les puces DL0.5 et QTL qui ont des résultats similaires pour les 3 caractères). À mesure que le seuil de DL utilisé pour construire les puces baisse et que le taux d'erreur génotypique augmente, les corrélations diminuent significativement. Pour l'intensité de ponte, la corrélation passe ainsi de 0,9960 à 0,9903 puis 0,9861 et enfin 0,9804, en passant respectivement d'un r^2 de 0.5 à 0.2, puis 0.1 et 0.05. Parallèlement, le taux d'erreur génotypique augmente respectivement de 2,36 à 3,88, 4,95 et 6,00 pour ces mêmes seuils de DL.
On observe la même évolution pour les puces basées sur la distance entre SNP, de meilleures corrélations étant calculées sur les trois caractères pour la puce 10Kequi que pour la puce 3Kequi. Ces évolutions sont bien conformes à celles notées dans la littérature (Dassonneville et al., 2011 ; Wolc et al., 2011) où une augmentation de la précision des évaluations génomiques est observée lorsque l'efficacité de l'imputation augmente.
Avec des nombres de SNP équivalents, les puces DL 0.5, QTL et 10Kequi présentent des corrélations très proches ; il faut regarder le 4ème chiffre significatif pour noter des différences. De façon surprenante, on constate que les corrélations entre les GEBV des candidats calculées avec les génotypages HD (300K) et avec les génotypages imputés sont significativement meilleures pour la puce 10Kequi que pour la puce DL 0.5 (ou la puce QTL). En effet, pour la puce 10Kequi, la corrélation pour l'intensité de ponte est de 0.9967 (+0.0007 par rapport à la puce QTL), de 0.9967 pour la couleur (+0.0022) et de 0.9971 pour le poids d'œufs (+0.0007). Or les taux d'erreurs que nous avons calculés sont meilleurs sur la puce QTL (2.31%) que sur la puce 10Kequi (3.18%). Une meilleure imputation ne garantirait donc pas d'obtenir une meilleure précision dans le calcul des évaluations génomiques. On obtient ainsi des résultats qui sont opposés à ceux trouvés dans la littérature (Dassonneville et al., 2011 ; Wolc et al., 2011).
De même, quand on compare la puce 3Kequi avec la puce DL 0.05, de meilleures corrélations sont calculées sur la puce 3Kequi alors que le taux d'erreur génotypique de l'imputation pour cette puce (7,18%) est moins bon que celui de la puce DL 0.05 (6.00%).
Cependant, les valeurs des corrélations étant très élevées, nous avons étudié les corrélations de Spearman sur les 150 candidats ayant la plus grande GEBV à partir des génotypages HD 300K (Tableau 6) afin d'évaluer leur éventuel reclassement selon leur GEBV calculées à partir des génotypages imputés. En effet, les sélectionneurs travaillent sur le classement des individus les uns par rapport aux autres et non sur leurs valeurs génétiques calculées de façon absolue. Nous

Scénario G0G1					
	Nombre de SNP	Taux d'erreur génotypique	Intensité de ponte	Couleur LAB	Poids d'œuf
Puce DL 0.5	9820	2,36	0,9647^{a}	0,9460^{a}	0,9741^{a}
Puce QTL	10114	2,31	0,9656^{a}	0,9499^{a}	0,9755^{a}
Puce 10Kequi	**9352**	**3,08**	**0,9739^{a}**	**0,9695^{b}**	**0,9855^{b}**
Puce DL 0.2	5224	3,88	0,9211^{b}	0,9258^{c}	0,9542^{c}
Puce DL 0.1	3988	4,95	0,9150^{bc}	0,8770^{de}	0,9258^{d}
Puce DL 0.05	3357	6,00	0,8591^{d}	0,8551^{e}	0,9226^{d}
Puce 3Kequi	3337	7,18	0,8842^{cd}	0,9017^{cd}	0,9275^{d}

Tableau 6 : Corrélations de Spearman entre les GEBV des 150 meilleurs candidats calculées à partir des génotypages HD (300K) et les GEBV des candidats calculées à partir des génotypages imputés, pour les 7 puces étudiées sur le scénario Utopige

Les lettres différentes entre puces et par caractère indiquent, après transformation des corrélations en z scores par la fonction r.paired du package Psych (Revelle, 2016), des différences de z scores significatives selon des tests de Student au seuil de première espèce α = 5%

rappelons ici que l'objectif de la sélection génétique est de choisir les individus ayant le meilleur potentiel génétique pour un ou des caractères.

La première observation que nous pouvons faire est que les corrélations sont toujours élevées mais des différences plus grandes sont observées entre les puces. La corrélation de Spearman permet de mieux discriminer les résultats.
Comme précédemment, les valeurs des corrélations baissent avec le seuil de DL. Ainsi pour la couleur (LAB), la corrélation passe de 0,9460 à 0,9258, puis 0,8770 et enfin 0,8551 en passant respectivement d'un r^2 de 0,5 à 0,2, puis 0,1 et 0,05. De même, les corrélations baissent avec la densité de SNP sur la puce BD car, pour la puce 10Kequi, la corrélation est de 0.9695 pour la couleur versus 0,9017 pour la 3Kequi. Dans ces 2 cas, nous confirmons l'hypothèse qu'une meilleure imputation permet d'obtenir une meilleure précision des évaluations génomiques puisque les puces DL 0.5 et 10Kequi présentent respectivement de meilleurs résultats d'imputations que les puces DL 0.05 et 3Kequi.
Par ailleurs, lorsque l'on regarde les puces DL 0.5, QTL et 10Kequi, les corrélations sont toujours élevées (>0,94) ce qui indique qu'il n'y a pas beaucoup de reclassements des candidats à la sélection. Ceci est conforme avec les résultats de Dassonneville et al. (2011) qui montrent aussi qu'il n'y a que très peu de changements dans le classement des animaux lorsque les taux d'erreurs sont faibles.
Toutefois, à densité de SNP équivalente, quand on compare la puce QTL avec la puce 10Kequi, on constate que les corrélations entre les GEBV des 150 meilleurs candidats calculées à partir des génotypages HD 300K et les GEBV de ces 150 mêmes candidats calculées à partir des génotypages imputés sont à nouveau meilleures, de façon significative, pour la puce 10Kequi pour la couleur et le poids d'œufs. En effet, pour la puce 10Kequi la corrélation pour la couleur est supérieure de +0.0196 par rapport à la puce QTL et de +0.0100 pour le poids d'œufs. De même, pour le caractère de la couleur, la puce 3Kequi présente de meilleures corrélations par rapport à la puce DL 0.05. Or, nous avons vu que les taux d'erreurs génotypiques sont plus faibles pour les puces QTL et DL 0.05 que pour les puces 10Kequi et 3Kequi. Ainsi, on confirme la conclusion qu'une meilleure imputation ne serait pas toujours indicatrice d'une meilleure précision des évaluations génomiques.

Devant ces résultats particulièrement inattendus, deux hypothèses peuvent être proposées. La première serait que certaines erreurs d'imputations ont plus d'effets que d'autres. Si l'imputation se trompe sur un SNP à effet fort sur les caractères étudiés, cette erreur peut avoir plus d'influence sur le calcul des évaluations génomiques que si l'imputation se trompe sur un SNP à effet plus faible sur les caractères étudiés. C'est ce qui pourrait se produire dans notre cas et qui pourrait expliquer pourquoi on obtient de meilleures évaluations génomiques avec les puces basées sur la distance entre SNP alors que les taux d'erreurs sont meilleurs sur les puces basées sur le DL. Ceci serait finalement cohérent avec le principe de la sélection qui sélectionne des régions en fort DL de génération en génération. Les erreurs d'imputations réalisées sur les puces DL auraient ainsi plus d'importances que les erreurs réalisées sur les puces basées sur la distance entre SNP, et donc plus de conséquences dans le calcul des évaluations génomiques.
La deuxième hypothèse est que lors du génotypage sur puces HD en laboratoire, des erreurs techniques peuvent se produire et subsister malgré le contrôle qualité (d'autant plus qu'ici les mères ne sont pas génotypés ce qui limite la vérification de la transmission mendélienne). Or lorsque l'on calcule le taux d'erreur génotypique, on compare les imputations aux génotypages issu du laboratoire qui ne sont pas les « vrais génotypes ». Ainsi, avec les puces basées sur le DL, on se rapprocherait peut être plus de la vérité.
Pour vérifier la première hypothèse, il faudrait étudier les effets des SNP calculés via le logiciel

d'évaluation génomique et regarder si, dans le cas des puces basées sur le DL, l'imputation fait des erreurs dans des régions à fort effet. Pour étudier la deuxième hypothèse, il faudrait réaliser des évaluations génomiques en intégrant les performances des filles des candidats afin de s'approcher le plus possible de la valeur génétique vraie des candidats. Il faudrait ensuite calculer la corrélation entre ces valeurs génétiques et les GEBV des candidats à la naissance calculées avec les génoypages 300K, les génotypages 3Kequi et DL0.05 par exemple.
Ces travaux n'ont pas pu être réalisés pendant le stage, faute de temps, mais seront réalisés par la suite.

Conclusion

L'optimisation des schémas de sélection passant par une minimisation du coût de la sélection et une maximisation de la précision des évaluations génomiques, ce stage a été mené dans le but de choisir la stratégie de génotypages basse-densité la mieux adaptée à la lignée de poule pondeuse de la société Novogen.
Les premiers travaux nous ont permis de définir les meilleurs outils et critères d'évaluations pour mener les imputations et juger de leur qualité. Il s'est avéré que le logiciel Beagle, bien qu'aboutissant à des résultats légèrement supérieurs, ne pourra pas être utilisé en routine du fait de temps de calcul très longs. Le logiciel FImpute se révèle donc un bon compromis entre efficacité et temps d'exécution. Quant aux critères d'efficacité de l'imputation, nous avons choisi le taux d'erreur génotypique, un critère plus sévère et plus discriminant que le taux d'erreur allélique et les corrélations.
Nous avons ensuite étudié l'intérêt de différentes puces : des puces BD basées sur le seuil de DL, pour lesquelles on rajoutait ou non des SNP marqueurs de QTL à effets forts sur des caractères de production et de qualité des œufs, et des puces BD basées sur la notion de distance entre SNP. Il a ainsi été montré que plus la densité de marqueurs sur les puces BD augmente, plus les imputations sont facilitées. De même, plus le seuil de DL augmente, plus les imputations sont de bonnes qualités. Enfin la prise en compte des SNP marqueurs de QTL permet d'obtenir des résultats supérieurs car ces SNP ont une faible MAF et sont mal imputés. Les inclure sur la puce BD permet donc d'éviter une mauvaise imputation par le logiciel. Enfin, la comparaison des taux d'erreurs génotypiques des différentes puces a permis de conclure à l'intérêt de prendre en compte la structure particulière du génome avicole en créant des puces BD basées sur le seuil de DL, tout en conservant des SNP marqueurs de QTL. En effet, un choix des SNP sur la base du DL et en tenant compte des QTL plutôt que sur la distance entre SNP permet d'optimiser le nombre de SNP sur les macro-chromosomes et de densifier les SNP sur les micro-chromosomes afin d'avoir des taux d'erreurs faibles et donc de réaliser de bonnes imputations.
Des études sur les populations de référence et populations candidates ont aussi été menées. Nous avons pu voir que plus la taille de la population de référence augmente, plus les imputations sont de bonnes qualités. De même, plus les relations de parenté entre population de référence et population candidate augmente, plus les imputations sont facilitées. Toutefois, en diminuant les relations de parenté, l'efficacité de l'imputation ne semble pas affectée. Il sera intéressant pour la suite d'aller étudier l'effet d'un saut de génération plus important, avec des sauts sur deux ou trois générations, pour ensuite analyser l'impact sur l'efficacité de l'imputation.
L'objectif des sélectionneurs étant de choisir les individus ayant le meilleur potentiel génétique pour plusieurs caractères d'intérêt, nous avons ensuite étudié ces résultats en lien avec les impacts sur la précision des évaluations génomiques de trois caractères présentant des déterminismes génétiques différents (intensité de ponte, couleur des œufs et poids d'œufs). Pour cela, nous avons analysé les corrélations de Spearman sur les 150 candidats ayant la plus grande GEBV à partir des génotypages HD 300K, afin d'évaluer leur éventuel reclassement selon leur GEBV calculées à partir des génotypages imputés. Nous avons constaté que de meilleurs

résultats sont obtenus pour les puces basées sur la distance entre SNP pour la couleur et le poids d'œufs. Or, les taux d'erreurs d'imputation sont plus élevés pour les puces basées sur la notion de distance entre SNP plutôt que sur le seuil de DL, pour des densités de SNP équivalentes. Ces résultats surprenants montrent qu'une meilleure imputation ne serait alors pas toujours indicatrice d'une meilleure précision des évaluations génomiques.
Deux hypothèses peuvent expliquer ces derniers résultats et seront à étudier par la suite. La première est que certaines erreurs d'imputations auraient plus d'effets que d'autres. Cela permettrait d'expliquer pourquoi, alors que le taux d'erreur est plus faible sur les puces DL, on obtient de meilleurs résultats sur les évaluations génomiques avec les puces basées sur la distance entre SNP. La deuxième hypothèse est que lors du génotypage sur puces HD en laboratoire, des erreurs techniques peuvent se produire et subsister malgré les contrôles qualités. Les génotypages imputés sont alors comparés aux génotypages HD qui ne sont donc pas totalement les génotypages réels. Cela expliquerait qu'avec les puces basées sur le DL, on obtienne des taux d'erreurs plus faibles et qu'on se rapprocherait peut être plus de la vérité.

Ce stage a donc permis de poser les bases d'une problématique qui sera à approfondir par la suite : faut-il choisir une puce BD basée sur le seuil de DL ou bien sur la notion de distance entre SNP ? En effet, nous avons pu voir que des taux d'erreurs plus faibles sont obtenus avec les puces DL plutôt qu'avec les puces avec des SNP équidistants. Toutefois, l'objectif de la sélection génétique reste d'aller choisir les individus ayant le meilleur potentiel génétique pour les caractères étudiés. Nous avons pu voir que les résultats des évaluations génomiques étaient très bons, et même meilleurs pour les puces avec des SNP équidistants. On ne peut donc actuellement pas conclure sur l'intérêt d'utiliser une puce BD basée sur le seuil de DL ou bien sur la notion de distance entre SNP.
Ce stage permet aussi d'ouvrir sur la possibilité de diminuer la densité des SNP sur les puces BD. En effet, même si les taux d'erreurs augmentent lorsque l'on diminue le nombre de SNP sur les puces BD, nous avons vu que pour des taux d'erreurs supérieurs à 5%, donc élevés, les résultats des évaluations génomiques restaient très bons. De plus, le prix des puces BD diminue lorsque l'on passe sous la barre des 3 000 SNP car la conception de ces puces pourrait alors se faire avec une autre technologie moins coûteuse (Affymetrix INC, 2016). Il pourrait donc être intéressant d'essayer de diminuer la densité des puces BD jusqu'à 3 000 SNP afin de diminuer le coût de conception des puces BD tout en maintenant de bons résultats d'évaluations génomiques. À ce moment, se reposera la question d'une puce basée sur le DL ou bien avec des SNP équidistants.
Par ailleurs, l'utilisation des séquences basses profondeurs comme alternatives aux puces BD (VanRaden et al., 2015) sera intéressante à étudier, le prix du séquençage devenant aujourd'hui de plus en plus accessible. Le séquençage des reproducteurs à une plus grande profondeur pourra alors également être envisagé si le gain en matière de précision des évaluations génomiques est conséquent.
Nous avons aussi pu voir l'intérêt d'utiliser une population de référence de grande taille pour réaliser de bonnes imputations. La stratégie des relations de parenté pourra être approfondie par la suite. Il sera intéressant d'étudier l'évolution des taux d'erreurs et des impacts sur les évaluations génomiques avec des sauts sur au moins deux ou trois générations. Enfin, l'entreprise Novogen ayant commencé le génotypage des femelles, il pourra être intéressant de se demander dans quelle mesure l'ajout de ces informations dans la population de référence va modifier les imputations et les évaluations génomiques. Les génotypages se faisant pour de nombreux animaux n'ayant pas de descendants présents dans la population de candidats, il sera aussi utile de voir si la restriction de la population de référence aux seules informations portées par les pères et les mères des candidats modifie beaucoup les imputations et les évaluations génomiques.

Bibliographie

ALLAIS, S., LAGARRIGUE, S. et LECERF, F., 2015. Notes de cours : Amélioration génétique des espèces d'élevage. Cours de spécialisation d'ingénieurs Ingénierie Zootechnique et master SAED. 145 p. AGROCAMPUS OUEST.

AFFYMETRIX, INC, 2016. Eureka Genomics® | Eureka Genotyping Solutions Animal and Plant Science. [en ligne]. Consulté le 11 août 2016. Disponible à l'adresse : https://www.eurekagenomics.com/ws/testing_and_analysis/ldma/animal_plant.html

BLOOM, S. E., DELANY, M. E. et MUSCARELLA, D. E., 1993. Constant and variable features of avian chromosomes. In: Etches RJ, Gibbons AMV, editors. Manipulation of the avian genome. Boca Raton, FL: CRC Press;. pp. 39–59.

BOICHARD, D., LE ROY, P., LEVÉZIEL, H. et ELSEN, J. M., 1998. Utilisation des marqueurs moléculaires en génétique animale. *INRA Prod. Anim*. Vol. 11, n° 1, pp. 67-80.

BOUQUET, A., FEVE, K., RIQUET, J. et LARZUL., 2015. Précision de l'imputation de génotypages haute densité à partir de puces basse densité pour des individus de race pure et croisés Piétrain. *Journées Recherche Porcine*. Vol. 47, pp. 1–6.

BROWNING, B. L. et BROWNING, S. R., 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*. Vol. 84, n° 2, pp. 210-223

BROWNING, B. L. et BROWNING, S. R., 2016. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*. Vol. 98, n° 1, pp. 116-126.

CARVALHEIRO, R., BOISON, S. A., NEVES, H. HR, SARGOLZAEI, M., SCHENKEL, F. S., UTSUNOMIYA, Y. T., O'BRIEN, A. M. P., SÖLKNER, J., MCEWAN, J. C., VAN TASSELL, C. P., SONSTEGARD, T. S. et GARCIA, J. F., 2014. Accuracy of genotype imputation in Nelore cattle. *Genetics Selection Evolution*. Vol. 46, pp. 69-79.

DASSONNEVILLE, R., BRØNDUM, R. F., DRUET, T., FRITZ, S., GUILLAUME, F., GULDBRANDTSEN, B., LUND, M. S., DUCROCQ, V. et SU, G., 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *Journal of Dairy Science*. Vol. 94, n° 7, pp. 3679-3686.

DASSONNEVILLE, R., FRITZ, S., DUCROCQ, V. et BOICHARD, D., 2012. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. *Journal of Dairy Science*. Vol. 95, n° 7, pp. 4136-4140.

DASSONNEVILLE, R., 2012. *Genomic selection of dairy cows*. Génétique animale. INRA GABI, Jouy-en-Josas : AgroParisTech. Thèse de doctorat.

DENJEAN, B., DUCOS, A., DARRE, A., PINTON, A., SÉGUELA, A., BERLAND, H., BLANC, M. F., FILLON, V. and DARRE, R., 1997. Caryotypes des canards commun (Anas platyrhynchos), Barbarie (Cairina moschata) et de leur hybride. *Rev. Méd. Vét.*, 148., 695-704.

EDDY, S. R., 2004. What is a hidden Markov model ? *Nature Biotechnology*. Vol. 22, n° 10, pp. 1315-1316.

FILLON, V., MORISSON, M., ZOOROB, R., AUFFRAY, C., DOUAIRE, M., GELLIN, J. et VIGNAL, A., 1998. Identification of 16 chicken microchromosomes by molecular markers using two-colour fluorescence in situ hybridization (FISH). *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*. Vol. 6, n° 4, pp. 307-313.

FRITZ, S., BAUR, A., GUILLAUME, F., DUCROCQ, V. et BOICHARD, D., 2011. Confirmation sur descendance des premiers taureaux génotypés en France. *Renc. Rech. Rum.* 2011. Vol. 18, pp. 414.

GROENEN, M. A. M., MEGENS, H. J., ZARE, Y., WARREN, W. C., HILLIER, LD W., CROOIJMANS, R. P. M. A., VEREIJKEN, A., OKIMOTO, R., MUIR, W. M. et CHENG, H. H., 2011. The development and characterization of a 60K SNP chip for chicken. *BMC genomics*. Vol. 12, n° 1, pp. 274-283.

GUÉMÉNÉ, D., BOULAY, M., CHAPUIS, H., DESNOUES, B., RAULT, P. et SEIGNEURIN, F., 2011. Espèces avicoles et productions biologiques - Sélection génétique. *Alter Agri*. N° 105, pp. 10-23.

HALEY, C. S. et VISSCHER, P. M., 1998. Strategies to utilize marker-quantitative trait loci associations. *Journal of Dairy Science*. Vol. 81 Suppl 2, pp. 85-97.

HAYES, B. E. N. et GODDARD, M. E., 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution*.. Vol. 33, n° 3, pp. 209-229.

HAYES, B. J., BOWMAN, P. J., DAETWYLER, H. D., KIJAS, J. W. et VAN DER WERF, J. H. J., 2012. Accuracy of genotype imputation in sheep breeds: Genotype imputation in sheep. *Animal Genetics*. Vol. 43, n° 1, pp. 72-80.

HAZEL, L. N., 1943. The genetic basis for constructing selection indexes. *Genetics*. Vol. 28, n° 6, pp. 476–490.

HEIDARITABAR, M., CALUS, M. P. L., VEREIJKEN, A., GROENEN, M. A. M et BASTIAANSEN, J. W. M., 2014. High imputation accuracy in layer chicken from sequence data on a few key ancestors. *10th World Congress of Genetics Applied to Livestock Production*. 3 p.

HEIDARITABAR, M., CALUS, M. P. L., VEREIJKEN, A., GROENEN, M. A. M. et BASTIAANSEN, J. W. M., 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. *BMC Genetics*. Vol. 16, n° 1, pp. 101-114.

HEIFETZ, E. M., FULTON, J.E., ZHAO, H., DEKKERS, J. C. M. et SOLLER, M., 2005. Extent and Consistency Across Generations of Linkage Disequilibrium in Commercial Layer Chicken Breeding Populations. *Genetics*. Vol. 171, n° 3, pp. 1173-1181.

HENDERSON, C. R., 1973. Sire evaluation and genetic trends. *Journal of Animal Science*. Vol. Symposium, pp. 10–41.

HENDERSON, C. R., 1975. Rapid method for computing the inverse of a relationship matrix. *Journal of dairy science*. Vol. 58, n° 11, pp. 1727–1730.

HICKEY, J. M., CROSSA, J., BABU, R. et DE LOS CAMPOS, G., 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Science*. Vol. 52, n° 2, pp. 654-663.

HOZÉ, C., FOUILLOUX, M. N., VENOT, E., GUILLAUME, F., DASSONNEVILLE, R., FRITZ, S., DUCROCQ, V., PHOCAS, F., BOICHARD, D. et CROISEAU, P., 2013. High-density marker imputation accuracy in sixteen French cattle breeds. *Genetics Selection Evolution*. Vol. 45, n° 1, pp. 33-43.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM, 2004a. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*. Vol. 432, n° 7018, pp. 695-716.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM, 2004b. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*. Vol. 432, n° 7018, pp. 717-722.

JUSSIAU, R., PAPET, A., RIGAL, J. et ZANCHI, E., 2014. *Amélioration génétique des animaux d'élevage: génome, caractères, sélection et croisements.* Educagri. ISBN 978-2-84444-929-0.

KRANIS, A., GHEYAS, A. A, BOSCHIERO, C., TURNER, F., YU, L., SMITH, S., TALBOT, R., PIRANI, A., BREW, F., KAISER, P., HOCKING, P. M., FIFE, M., SALMON, N., FULTON, J., STROM, T. M., HABERER, G., WEIGEND, S., PREISINGER, R., GHOLAMI, M., QANBARI, S., SIMIANER, H., WATSON, K. A., WOOLLIAMS, J. A. et BURT, D. W., 2013. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics*. Vol. 14, n° 1, pp. 59-71.

LE ROY, P., CHAPUIS, H. et GUEMENE, D.l, 2014. Sélection génomique: quelles perspectives pour les filières avicoles? *INRA Prod. Anim*. Vol. 27, n° 5, pp. 331–336.

LEGARRA, A., AGUILAR, I. et MISZTAL, I., 2009. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*. Vol. 92, n° 9, pp. 4656-4663.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. et DONNELLY, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*. Vol. 39, n° 7, pp. 906-913.

MARCHINI, J. et HOWIE, B., 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*. Vol. 11, n° 7, pp. 499-511.

MISZTAL, I., TSURUTA, S., STRABEL, T., AUVRAY, B., DRUET, T. et LEE, D. H., 2002. BLUPF90 and related programs (BGF90). In : *Proceedings of the 7th world congress on genetics applied to livestock production*. Montpellier, Communication No. 28–27. pp. 21–22.

MOSER, G., KHATKAR, M. S., HAYES, B. J. et RAADSMA, H. W., 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genetics Selection Evolution*. Vol. 42, n° 1, pp. 37-51.

MURTAGH, F. et LEGENDRE, P., 2011. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*. Vol. 31, n° 3, pp. 274–295.

NCBI, 2016. NCBI Gallus gallus Annotation Release 103. [en ligne]. Consulté le 11 août 2016. Disponible à l'adresse : http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/103/

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. et SHAM, P. C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. Vol. 81, n° 3, pp. 559-575.

REVELLE, W., 2014. psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, Illinois*. 367 p.

ROBERT, R., HERAULT, F., ROMÉ, H., VARENNE, A., CHAPUIS, H., VIGNAL, A., BURLOT, T., LE ROY, P., 2015. A linkage disequilibrium study in a layer chicken population. In: *Proceedings of the 9th European Symposium on Poultry Genetics* (p. 43). Presented at 9. European Symposium on Poultry Genetics, Tuusula, FIN (2015-06-16 - 2015-06-18). http://prodinra.inra.fr/record/308552

ROBERT-GRANIÉ, C., LEGARRA, A. et DUCROCQ, V., 2011. Principes de base de la sélection génomique. *INRA Prod. Anim*. Vol. 24, n° 4, pp. 331-340.

ROMÉ, H., VARENNE, A., HÉRAULT, F., CHAPUIS, H., ALLENO, C., DEHAIS, P., VIGNAL, A., BURLOT, T. et LE ROY, P., 2015. GWAS analyses reveal QTL in egg layers that differ in response to diet differences. *Genetics Selection Evolution*. Vol. 47, n° 1, pp. 83-94.

SARGOLZAEI, M., CHESNAIS, J. P. et SCHENKEL, F. S., 2014. A new approach for efficient genotype imputation using information from relatives. *BMC genomics*. Vol. 15, n° 1, pp. 478-489.

SHRIMPTON, A. E. et ROBERTSON, A., 1988. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila Melanogaster. II. Distribution of Third Chromosome Bristle Effects within Chromosome Sections. *Genetics*. Vol. 118, n° 3, pp. 445–459.

VANRADEN, P. M., SUN, C. et O'CONNELL, J. R., 2015. Fast imputation using medium or low-coverage sequence data. *BMC Genetics*. Vol. 16, n° 1, pp. 82-93.

VENTURA, R. V., LU, D., SCHENKEL, F. S., WANG, Z., LI, C. et MILLER, S. P., 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. *Journal of animal science*. Vol. 92, n° 4, pp. 1433–1444.

VIGNAL, A., 2000. État de la carte de la poule. *INRA Prod. Anim*. Hors-série « Génétique moléculaire : principes et application aux populations animales », pp. 113-114.

WEIGEL, K. A., DE LOS CAMPOS, G., VAZQUEZ, A. I., ROSA, G. J .M., GIANOLA, D. et VAN TASSELL, C. P., 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of Dairy Science*. Vol. 93, n° 11, pp. 5423-5435.

WOLC, A., KRANIS, A., ARANGO, J., SETTAR, P., FULTON, J. E., O'SULLIVAN, N., AVENDAÑO, S., WATSON, K. A., PREISINGER, R., HABIER, D., LAMONT, S. J., FERNANDO, R., GARRICK, D. J. et DEKKERS, J. C. M., 2014. Applications of genomic selection in poultry. *10th World Congress of Genetics Applied to Livestock Production*. 6 p.

WOLC, A., KRANIS, A., ARANGO, J., SETTAR, P., FULTON, J. E., O'SULLIVAN, N., AVENDAÑO, S., WATSON, K. A., PREISINGER, R., HABIER, D., LAMONT, S. J., FERNANDO, R., GARRICK, D. J. et DEKKERS, J. C. M., 2015. Genomic selection in layer and broiler breeding. *LOHMANN Information*. Vol. 49, pp. 4-11.

WOLC, A., KRANIS, A., ARANGO, J., SETTAR, P., FULTON, J. E., O'SULLIVAN, N. P., AVENDANO, A., WATSON, K. A., HICKEY, J. M., DE LOS CAMPOS, G., FERNANDO, R. L., GARRICK, D. J. et DEKKERS, J. C. M., 2016. Implementation of genomic selection in the poultry industry. *Animal Frontiers*. Vol. 6, n° 1, pp. 23-31.

YTOURNEL, Florence, 2008. *Déséquilibre de liaison et cartographie de QTL en population sélectionnée*. Génétique animale. INRA GABI, Jouy-en-Josas : AgroParisTech. Thèse de doctorat.

Annexe I : Récapitulatif de quelques travaux d'imputation en fonction de la filière, du type d'imputation, des facteurs pouvant influencer l'imputation et des principaux résultats

Publication	Filière	Type d'imputation	Facteurs étudiés	Résultats
Dassonneville et al., 2011 Dassonneville et al., 2012	Bovine	Imputation de la puce Bovine 3K BeadChip® vers la puce Bovine SNP50 BeadChip® (50K) (2011) Imputation de diverses puces BD (3K à 6K) vers la puce Bovine SNP50 BeadChip® (50K) (2012)	Taille de la population de référence	Augmentation de l'efficacité de l'imputation avec une augmentation de la taille de la population de référence. Race Holstein, avec 3071 individus de référence, TA = 3.9% ; avec 12 078 individus de référence, TA = 2.1%
			Densité de marqueurs	Augmentation de l'efficacité de l'imputation avec une augmentation de la densité des marqueurs sur puces BD. Race Holstein, avec 3K SNP, r = 0.94 ; avec 6K SNP, r = 0.96.
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté.
			Influence sur les évaluations génomiques	Augmentation de la précision des évaluations génomiques avec une augmentation de l'efficacité de l'imputation. Peu de changements dans le classement des animaux.
Hozé et al., 2013	Bovine	Imputation de la puce Bovine SNP50 BeadChip® (50K) vers la puce BovineHD BeadChip® (777K)	Taille de la population de référence	Augmentation de l'efficacité de l'imputation avec une augmentation de la taille de la population de référence, variable selon les races. TA_{moy} = 1.36% ; Race Simmental (100 animaux) TA = 2.51% ; Race Normande (450 animaux) TA = 0.33%. Au-delà de 400 animaux, pas d'amélioration. TA_{min} = 0.7%
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté. Race Abondance, coefficient de parenté R = 0.146 et TA = 0.75% Race Brown Swiss, coefficient de parenté R = 0.074 et TA = 1.92%
			Taille de la population cible	Diminution de l'efficacité de l'imputation avec une augmentation de la taille de la population cible. Race Montbéliarde (23 animaux cibles) TA = 0.51% ; Race Limousine (185 animaux cibles) TA = 1.09% Effet plus faible que celui dû à la taille de la population de référence.
			Niveau du déséquilibre de liaison	Faible effet
			MAF	Taux d'erreur élevé (entre 5% et 20%) sur les marqueurs à faible fréquence
Ventura et al., 2014	Bovine	Imputation de la puce BovineLD BeadChip® (7K) vers la puce Bovine SNP50 BeadChip® (50K)	Taille de la population de référence	Augmentation de l'efficacité de l'imputation avec une augmentation de la taille de la population de référence. Race Elora, avec 250 animaux de référence, TG = 13% ; avec 1500 animaux de référence, TG = 5.7%
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté. Race Angus, avec 350 animaux de référence, TG = 2.4% ; 317 animaux de référence (suppression de 37 animaux les plus proches), TG = 5.8%
Carvalheiro et al., 2014	Bovine	Imputation de diverses puces BD (de 7K à 75K) vers la puce BovineHD BeadChip® (777K)	Densité de marqueurs	Augmentation de l'efficacité de l'imputation avec une augmentation de la densité des marqueurs sur puces BD. Pour 7K SNP, r = 0.9257 ; pour 55K SNP, r = 0.9931
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté. Avec une puce 7K et un coefficient de parenté R = 0.10, r = 0.87 ; pour coefficient de parenté R = 0.20, r = 0.92 Effet accru avec une augmentation du nombre de marqueurs sur puces BD. Avec une puce 55K et un coefficient de parenté R = 0.10, r = 0.985 ; pour coefficient de parenté R = 0.20, r = 0.99
			MAF	Taux d'erreur élevé sur les marqueurs à faible fréquence. r < 0.6

Bouquet et al., 2015	Porcine	Imputation de diverses puces BD (de 450 SNP à 10K) vers la puce HD PorcineSNP60 BeadChip® (60K)	Densité de marqueurs	Augmentation de l'efficacité de l'imputation avec une augmentation de la densité des marqueurs sur puces BD. Pour 450 SNP, r= 0.83 ; pour 10K SNP, r > 0.98
			Utilisation de race pure ou d'individus croisés	Meilleurs résultats d'imputation chez les individus de race pure que chez les individus croisés. Pour 10K SNP en race pure r > 0.98 ; en croisés, r = 0.89
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté. En race pure Piétrain, avec une puce 5K et les individus Utopige r = 0.98 ; avec des individus non apparentés aux individus Utopige r = 0.96 Effet accru avec une augmentation du nombre de marqueurs sur puces BD. En race pure Piétrain, avec une puce 10K et les individus Utopige r = 0.99 ; avec des individus non apparentés aux individus Utopige r = 0.98
Wolc et al., 2011	Avicole	Imputation de diverses puces BD (de 400 à 42K) vers la puce Chicken 600K Illumina BeadChip®	Densité de marqueurs	Augmentation de l'efficacité de l'imputation avec une augmentation de la densité des marqueurs sur puces BD. Avec 400 SNP, r = 0.95 ; avec 42K SNP, r > 0.97
			Influence sur les évaluations génomiques	Augmentation de la précision des évaluations génomiques avec une augmentation de l'efficacité de l'imputation
Heidaritabar et al., 2014 Heidaritabar et al., 2015	Avicole	Imputation de la puce Chicken 60K Illumina BeadChip® vers la séquence (2014) Imputation de diverses puces BD (3K et 48K) vers la puce Chicken 60K Illumina BeadChip® (2015)	Taille de la population de référence	Augmentation de l'efficacité de l'imputation avec une augmentation de la taille de la population de référence. Pour 22 animaux de référence, r = 0.78 ; pour 62 animaux de référence, r = 0.87
			Relation de parenté entre population de référence et population cible	Pas de diminution de l'efficacité de l'imputation avec un éloignement des relations de parenté entre population de référence et population cible
			MAF	Taux d'erreur élevé (entre 20% et 40%) sur les marqueurs à faible fréquence
Hayes et al., 2011	Ovine	Imputation de diverses puces BD (1K à 5K) vers la puce Ovine 50K Illumina BeadChip®	Densité de marqueurs	Augmentation de l'efficacité de l'imputation avec une augmentation de la densité des marqueurs sur puces BD. Race Border Leicester, pour 1K SNP TG = 31% ; pour 5K SNP, TG = 20%.
			Taille de la population de référence	Augmentation de l'efficacité de l'imputation avec une augmentation de la taille de la population de référence. Race Merino, avec 100 individus de référence, TG = 36% ; avec 2612 individus de référence, TG = 29%
			Relation de parenté entre population de référence et population cible	Augmentation de l'efficacité de l'imputation avec une augmentation des relations de parenté. Race Border Leicester, coefficient de parenté R = 0.34 et TG = 20% ; Race Merino, coefficient de parenté R = 0.17 et TG = 37%
			MAF	Taux d'erreur élevé (entre 5 et 40%) sur les marqueurs à faible fréquence

r : taux de corrélation ; R : coefficient de parenté ; TA : taux d'erreur allélique ; TG : taux d'erreur génotypique

AGRO CAMPUS OUEST	Diplôme : Ingénieur et Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage Spécialité : Sciences de l'Animal pour l'Élevage de Demain (SAED) Spécialisation / option : Enseignant référent : Vanessa Lollivier

Auteur(s) : Florian Herry Date de naissance* : 23/11/1992	Organisme d'accueil : NOVOGEN Adresse : Mauguerrand – CS70265 22 800 Le Foeil France Maître de stage : Sophie Allais
Nb pages : 20 Annexe(s) : 1	
Année de soutenance : 2016	

Titre français : Design d'une puce à SNP basse densité pour une lignée de poules pondeuses NOVOGEN

Titre anglais : Design of a low-density SNP chip for a line of NOVOGEN laying hens

Résumé (1600 caractères maximum) :

L'optimisation des schémas de sélection en poules pondeuses passe aujourd'hui par une minimisation du coût de la sélection et une maximisation de la précision des évaluations génomiques. Pour cela, la technique de l'imputation peut être utilisée en sélection génomique afin de déduire les génotypages haute densité des candidats à partir de leurs génotypages basse densité obtenus avec des puces à SNP basse densité coûtant moins cher que des puces à SNP haute densité. Dans cette optique, différentes études ont été menées afin de choisir la stratégie de génotypages basse densité la mieux adaptée à la lignée de poule pondeuse de la société Novogen.
Le logiciel FImpute est préféré au logiciel Beagle compte tenu d'un temps de calcul très cours avec FImpute, et le taux d'erreur génotypique, le taux d'erreur allélique et les corrélations aboutissent aux mêmes conclusions. L'étude de différentes puces basse densité montre que les imputations sont meilleures lorsque la densité des SNP augmente, lorsque le seuil de DL augmente et lorsque des SNP avec une faible MAF sont pris en compte. Les études de populations montrent que les imputations sont meilleures lorsque la taille de la population de référence augmente. En revanche, même s'il l'on observe une amélioration des imputations en augmentant les relations de parenté entre population de référence et population cible, on n'observe pas de dégradation des imputations avec une diminution des relations. Enfin, les conséquences des imputations sur les évaluations génomiques sont, en l'état, surprenantes et seront à approfondir par la suite.

Abstract (1600 caractères maximum) :

The optimization of breeding programs of laying hens pass now through minimizing the cost of selection and maximizing the accuracy of genomic evaluations. To do so, imputation can be used in genomic selection to deduce high density genotyping candidates from their low density genotyping obtained with low density SNP chips which cost less than high density SNP chips. In this regard, various studies have been conducted to choose the low density genotyping strategy the best suited to laying hens line of Novogen.
FImpute software is preferred to Beagle software given a very short calculation time with FImpute, and the rate of genotypic error, the rate of allelic error and correlations lead to the same conclusions. The study of different low density chips shows that imputations are better when the density of SNP increases, when the LD threshold increases and when SNP with low MAF are considered. Population studies show that imputations are better when the size of the reference population increases. However, even if one can see an improvement of imputations by increasing the kinship between reference population and target population, a degradation of imputations is not observed with a decreasing kinship. Finally, the impact of imputations on genomic evaluation is, as things stand, surprising and will be deepened later.

Mots-clés : Précision de l'imputation, poules pondeuses, densité de SNP, population de référence, évaluations génomiques

Key Words: Imputation accuracy, layer chickens, SNP density, reference population, genomic evaluations

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*

Document à intégrer au mémoire