

LE PROJET MILLE GENOMES GALLUS : COMMENT PARTAGER LES DONNEES DE SEQUENCES POUR MIEUX LES UTILISER

Michèle Tixier-Boichard¹, Frédéric Lecerf², Philippe Bardou³, Christophe Klopp³

¹ GABI, INRA, AgroParisTech, Université Paris-Saclay, 78352 JOUY-EN-JOSAS, ²PEGASE, Agrocampus Ouest, INRA, 35042 RENNES, ³SIGENAE, INRA, 31326 CASTANET-TOLOSAN,
michele.tixier-boichard@inra.fr

RESUME

Le séquençage du génome entier est maintenant utilisé dans de nombreux projets de recherche sur le poulet. Le projet Mille Génomes Gallus s'inspire du projet Mille Génomes Bovins et a pour objectif de rassembler les séquences du génome entier d'animaux du genre *Gallus*, produites par les équipes de recherche et les entreprises de sélection. L'intérêt est d'augmenter la puissance des analyses en augmentant le nombre de génomes comparés. Les applications concernent l'analyse de la structure du génome, la caractérisation globale de la diversité de l'espèce, l'identification de mutations causales et l'aide à la sélection génomique. Cette synthèse décrit d'abord le contexte de la génomique du poulet avant de développer le concept du projet Mille Génomes, de l'illustrer avec un projet pilote utilisant les données produites par des équipes françaises et de discuter les modalités de son extension. Le projet pilote français regroupe 207 séquences individuelles issues de 8 projets de recherches financés par des fonds publics, sur une large gamme de populations (poulets de chair, poules pondeuses, races locales, ancêtres sauvages). Les jeux de données sont décrits par des métadonnées techniques et des métadonnées liées à l'animal et sa population d'origine. Aucune information phénotypique n'est partagée. L'équipe SIGENAE a aligné l'ensemble des séquences sur la version 5 du génome de référence et a répertorié plus 40 millions de variants SNP. L'analyse de structure des populations regroupe les 207 individus en sept groupes génétiques. L'étude des SNP du gène *MC1R* montre une signature de sélection chez les pondeuses de plumage rouge. D'autres analyses sont prévues, notamment pour les variants structuraux. Les projets en cours en Europe et les contacts pris avec des équipes étrangères laissent penser que la cible de Mille Génomes peut être atteinte rapidement. Les principes d'un futur accord de consortium sont exposés.

ABSTRACT

The Mille Gallus genomes project: how to share sequence data in order to better exploit them.

The Mille Gallus genomes project aims at gathering whole genome sequence data in chicken, produced by research teams as well as breeding companies, on the model of the Mille Bull genomes project. The main interest is to increase the power of analyses by increasing the number of genomes to compare. Applications are to improve the knowledge of the chicken genome, to characterize with a high accuracy the genetic diversity at the species level, to facilitate the identification of causal mutations and to provide support to genomic selection by facilitating imputation of sequence data from SNP data. A pilot project has started in France with the collection of 207 individual genome sequences produced by eight publicly funded projects on a range of chicken populations (fast-growing and slow-growing broilers, layers, local breeds, wild ancestors). Data sets are described by technical metadata and by metadata on the animal and its population of origin. Phenotypic data are not shared. The SIGENAE team has aligned all sequences on the GalGal5 reference genome and used a common pipeline for SNP calling. More than 42 millions SNP were identified. A structure analysis led to group the 207 individuals in seven genetic clusters. The analysis of SNP in the coding sequence of the *MC1R* gene showed a selection signature in the brown-egg layers related to red plumage. Other analysis are planned, including detection of structural variants. On-going projects in Europe and feedback from international teams led to think that the target of Mille genomes can be reached quickly. The principles of a future consortium agreement are presented.

1-CONTEXTE DE LA GENOMIQUE DU POULET

1.1. Particularités de l'espèce

Le poulet est une espèce d'intérêt économique majeur pour l'alimentation humaine, avec 109 millions de tonnes de viande produites et 70 millions de tonnes d'œufs en 2017 (<http://www.fao.org/faostat>). Le poulet est aussi une espèce modèle utilisée de longue date pour la biologie du développement, grâce à l'accès relativement facile à l'embryon à tous les stades de développement. Cette conjonction d'espèce économique et d'espèce modèle lui a valu d'être la première espèce d'élevage à avoir son génome séquencé. L'intérêt de séquencer le génome du poulet était aussi de constituer une référence pour l'assemblage du génome d'autres oiseaux, comme par exemple la pintade (voir Vignal et al, communication courte soumise aux JRA 2019).

La principale espèce ancêtre sauvage du poulet, *Gallus gallus*, vit toujours dans son habitat naturel en Asie (Inde, au Vietnam, en Thaïlande...) de même que les espèces proches que sont *Gallus sonneratii*, *Gallus lafayetii* et *Gallus varius*. L'étude du génome de ces espèces permet donc d'aborder le processus de domestication du poulet par la génomique. Le génome du poulet est plus compact que celui des mammifères avec 1,26 Gigabases contre environ 3 Gigabases, son séquençage est donc moins coûteux. La première version du génome de référence a été produite à partir de l'ADN d'une population consanguine de poule de jungle (Hillier et al., 2004). La dernière version disponible est la version 6.

1.2. Révolution de l'approche génome entier

Une fois qu'un génome de référence a été produit pour une espèce, le réséquençage des génomes individuels permet d'obtenir une masse d'informations sur l'histoire d'une espèce ou d'une race, mais aussi sur la diversité génétique accessible pour la sélection (Vignal et Besbes, 2005) comme pour la gestion des populations en conservation. Le génome est parfois décrit comme une archive de l'histoire d'une population car son analyse permet de reconstituer les variations d'effectif, de détecter des événements d'introgression entre espèces ou entre races au sein d'une espèce, de détecter des signatures de sélection ou encore d'identifier un individu ancêtre commun aux animaux porteurs d'une anomalie génétique.

Le génome constitue ainsi un même référentiel pour tous les travaux de recherche sur la génétique d'une espèce. Le séquençage a ainsi conduit à remplacer les marqueurs microsatellites des années 1990-2010 par les marqueurs SNP, défini comme la variation par rapport au génome de référence d'une base

précisément localisée dans le génome. De ce fait, les résultats de différentes études deviennent beaucoup plus facilement comparables entre eux qu'auparavant.

Sur le plan technique, le séquençage a constamment gagné en efficacité et son coût a régulièrement diminué. En 2018, le coût du séquençage d'un génome entier de poulet, avec une profondeur de 10X (c'est-à-dire qu'une position du génome est en moyenne lue 10 fois afin d'être connue de façon certaine) variait entre 300 et 400 euros selon les prestataires de séquençage. Le résultat de ce séquençage comprend toutes les variations présentes chez l'individu par comparaison au génome de référence, et notamment celles qui lui sont propres. A titre d'exemple, le génotypage d'un poulet pour le seul outil public à haute densité de 580000 marqueurs SNP (jeu standardisé dénommé 'puce de génotypage) coûtait environ 150 euros en 2018 pour produire une information certes très utile mais ne permettant pas de détecter les variants nouveaux propres à l'individu étudié.

Toutefois, l'extraction des informations pertinentes à partir des données de réséquençage suppose un travail d'analyse beaucoup plus important que celui nécessaire pour exploiter les résultats d'une puce de génotypage. Les principales étapes sont résumées dans la figure 1. Identifier un variant génétique par sa position en nombre de bases sur un chromosome permet aussi de rechercher dans les bases de données si d'autres variants ont déjà été décrits à cette position ou à son voisinage.

L'analyse de données devient le facteur limitant, mutualiser l'analyse permet de mieux exploiter les données de réséquençage et de génotypage, et d'augmenter la puissance des comparaisons entre individus ou entre groupes, c'est la principale motivation d'un projet « Mille Génomes ».

1.3 Contexte de l'open data et de l'open science

Récemment, les agences de financement de la recherche et notamment l'Union Européenne ont souhaité que les données produites grâce à des financements publics soient accessibles à tous les acteurs de la recherche, publics ou privés, et plus généralement à tous les citoyens, qui financent indirectement ces recherches. Cette stratégie dite 'open data' est considérée comme nécessaire au développement de l'innovation à partir de la recherche. Elle s'inscrit dans une volonté plus large de science ouverte à la société. Elle est particulièrement relayée par la 'Research Data Alliance' (<https://www.rd-alliance.org/>) dont le nœud français est coordonné par le CNRS dans le cadre du projet européen RDA Europe 4.0, qui a démarré le 1er mars 2018. Les institutions de recherche telles que l'INRA ont adhéré à cette stratégie et accompagnent leurs chercheurs dans cette démarche

non intuitive, chaque chercheur ayant naturellement le réflexe de conserver ses données pour la question qu'il elle veut traiter. Il ne s'agit pas de diffuser n'importe comment des données mais de les produire et de les conserver de manière standardisée et de les documenter avec des 'méta-données' utilisant un vocabulaire standardisé afin de faciliter la réutilisation des données. L'idée est bien qu'une seule personne ou qu'un seul groupe de chercheurs ne peut exploiter totalement un jeu de données et que le partage de ces données améliore leur utilisation et la connaissance qui en est extraite. L'enjeu est aussi de ne pas être le dernier à exploiter ses données et la stratégie *open data* pousse à une certaine accélération de la recherche. Ainsi, chaque projet de recherche doit comporter un plan de gestion des données produites, décrivant le type de données produites, leurs conditions d'accès et d'utilisation futures (stockage, période d'embargo, reconnaissance du producteur initial des données).

La génomique se prête particulièrement bien à cette stratégie de partage des données, grâce à la standardisation de la description des données de séquence et à l'existence d'entrepôts internationaux qui rassemblent les données du génome de référence et des génomes individuels, tels que NCBI développé aux Etats-Unis (<https://www.ncbi.nlm.nih.gov/>), Ensembl développé en Europe (<http://www.ensembl.org>) ou encore UCSC développé par une université américaine (<https://genome.ucsc.edu/>).

2. LE CONCEPT 'Mille GENOMES'

Le concept Mille Génomes a été inventé pour l'espèce humaine (www.internationalgenome.org) en 2010 avec une publication de référence en 2012 décrivant la variabilité génétique de 1092 génomes humains (The Mille Genomes Project Consortium, 2012). L'espèce bovine a été la première espèce domestique à bénéficier de cette approche et constitue l'exemple de référence pour mettre en place un projet Mille Génomes pour le poulet.

2.1 Le projet Mille Génomes bovins :

L'idée est née en Australie en 2011 et un premier consortium a été créé en 2012 avec 7 partenaires dont l'INRA. L'objectif est d'avoir accès à un grand nombre de séquences individuelles sans augmenter les coûts par partenaire, au sein d'un consortium dans le cadre d'un contrat signé par chaque membre. Pour un membre, l'intérêt est double : il accède à une quantité d'information beaucoup plus importante ; il bénéficie

d'un traitement des données normalisé. Les lectures une fois alignées sur le génome de référence (étape 1-figure 1) sont rassemblées dans un même site où sont réalisées toutes les étapes de traitement avec une même méthode, afin de redistribuer à tous les membres du consortium les résultats de l'étape 3 de la figure 1. La première publication de référence issue de ce projet date de 2014 et montre que l'analyse de la séquence de 234 taureaux facilite la cartographie génétique de caractères monogéniques mais aussi complexes (Daetwyler et al., 2014).

Pour garantir l'homogénéité du traitement, seules les lectures de séquences obtenues par la technologie Illumina sont partagées et caractérisées par des descripteurs minimaux obligatoires : race, identifiant de l'animal (potentiellement anonymisé mais unique), sexe, profondeur de séquence, plusieurs critères techniques de réalisation du séquençage. Les membres sont obligatoirement apporteurs de données de séquences : à la création du consortium, le minimum d'apport était d'au moins 10 animaux avec une profondeur individuelle au moins égale à 4X pour un total supérieur à 100X (par exemple, au moins 10 animaux à 10X ou 25 animaux à 4X). Les données ont été apportées par fournées successives, dénommées 'run' : de 125 animaux au run 1, à 2433 au dernier run (n° 6). L'apport minimal demandé augmente à chaque run afin de renforcer l'engagement des partenaires t. Le consortium comptait plus de 30 membres en 2017. Environ la moitié des données de séquences brutes sont déposées dans les sites publics au fur et à mesure des publications réalisées. Le choix du dépôt en base publique relève de l'apporteur de données et n'est pas une obligation du consortium. On voit donc ici un exemple où l'open data est contrôlé : les données sont d'abord partagées au sein d'un consortium avant d'être progressivement rendues publiques.

Dans ce consortium, chaque membre finance ses coûts, y compris de séquençage. Le coordinateur australien (DEPI, Melbourne, Australie) finance le stockage des données et les analyses communes (étapes 1 à 3, figure 1) alors que le travail d'annotation (étape 4, figure 1) est partagé entre différents membres du consortium. A titre d'exemple, le résultat du run 4 a permis d'identifier plus de 35 millions de NP et 1,7 millions d'InDels¹ à partir des données de séquence fournies par 20 partenaires sur 1147 animaux issus de 27 races. Une particularité de ce consortium est aussi d'améliorer la qualité des données apportées en utilisant l'information de déséquilibre de liaison à l'échelle de l'échantillon total pour fournir la probabilité a posteriori du génotype de chaque individu. Ce traitement conduit à inférer les génotypes absents ou à corriger les

¹ Petite suppression (délétion) ou addition (insertion) d'un faible nombre de bases (1 à 50)

génotypes peu fiables et à faire disparaître les pseudo-allèles issus d'erreurs de séquençage. Cette étape importante pour la qualité des résultats n'est possible qu'en traitant l'ensemble des données, elle constitue donc une plus-value du consortium.

Il est capital de noter que les phénotypes et les génotypes utilisés en sélection ne sont pas partagés. Chaque partenaire peut réaliser pour son compte des analyses avec les données mutualisées. Ainsi, les associations entre variants de séquence (SNP + Indels) et phénotypes sont réalisées par les partenaires sur leurs propres données de phénotypes et ces données sont partagées lorsque l'ensemble des membres du consortium le jugent utile. Cela a permis au consortium d'identifier l'ensemble des QTLs déterminant la taille des bovins et d'en faire une publication commune (Bouwman et al., 2018).

Une autre valorisation importante pour la sélection est de pouvoir reconstituer la séquence d'un animal qui n'a été que génotypé. Ainsi, la connaissance du génotype à un certain nombre de positions dans le génome (en général au moins 50000) permet de déduire la séquence la plus probable entre chaque position génotypée grâce au référentiel de SNP établi par le projet Mille Génomes. De ce fait, des animaux génotypés sont presque aussi bien connus que des animaux séquencés. Cette étape s'appelle 'imputation' et peut être réalisée par un partenaire sans avoir à partager les résultats obtenus. Elle améliore grandement la précision de la sélection conduite par ce partenaire et permet aussi de construire de grandes populations de cartographie pour identifier directement les variants candidats sous-jacents aux QTL. Elle est d'autant plus utile que le nombre d'animaux génotypés et apparentés aux animaux séquencés est grand. Ainsi, en France et à ce jour, près de 100 000 animaux avec phénotypes et génotypes SNP de puces ont été imputés à l'échelle de la séquence. Cet argument d'imputation a été décisif pour constituer le consortium bovin à l'époque où la sélection génomique se mettait en place dans tous les pays. La majorité des animaux séquencés correspondent à des taureaux ancêtres ayant une forte contribution génétique à leur population. Ils sont très informatifs en termes de variants présents dans les populations mais n'ont aucun intérêt stratégique en termes de phénotypes ou de sélection.

2.2 Objectifs du projet Mille Génomes Gallus

L'objectif général est d'augmenter la puissance des analyses génomiques en augmentant le nombre de génomes séquencés accessibles pour une même analyse. Quatre types d'applications sont proposées.

2.2.1. Analyse de la structure du génome

Rassembler le plus grand nombre de séquences permet non seulement une identification plus complète des variants de type SNP et InDel, mais aussi d'étudier leur répartition dans le génome, notamment en distinguant les macrochromosomes des microchromosomes qui sont une particularité des oiseaux (à l'exclusion des rapaces). Une étape supplémentaire par rapport au projet bovin serait de répertorier l'ensemble des variants structuraux, par exemple les insertions rétrovirales qui sont connues pour entraîner certaines modifications morphologiques majeures chez le poulet (coquille bleue, plumage blanc, emplumement lent...).

2.2.2. Caractérisation de la diversité de l'espèce

Le poulet est une espèce à intervalle de génération court, qui a vu un grand nombre de races se différencier en quelques siècles, qui a accumulé le plus grand nombre de générations de sélection pour des populations commerciales, et qui a aussi permis le développement d'un grand nombre de lignées expérimentales présentant des phénotypes extrêmes. S'il est possible de découper la filière en poule de ponte, poulet de chair et races locales, il n'y a pas une seule race ultra dominante comme l'est la race bovine Holstein, et l'analyse de la diversité génétique est donc un objectif sans doute plus important que pour le projet Mille Génomes bovin. De plus, les projets de diversité génétique ont jusqu'à maintenant été réalisés avec différentes puces de génotypage ayant très peu de marqueurs en commun ce qui empêche la mutualisation de ces données, alors que le séquençage permettrait de connecter ces projets.

Les applications envisagées concernent l'identification d'allèles spécifiques de races ou de lignées, l'étude de scénarios de domestication et la détection de flux de gènes entre populations sauvages et domestiques ou entre populations domestiques, par exemple entre races locales et lignées commerciales. La condition est évidemment que des données de séquence soient apportées pour ces différents types de populations.

2.2.3. Identification de mutations causales

Plusieurs cas de figure peuvent se présenter :

- Répertorier toutes les variations possibles pour un gène candidat d'intérêt général : on peut penser à l'exemple du gène de coloration *MC1R*, pour lequel différents variants de la séquence codante ont déjà été publiés, sans que les variants des régions régulatrices soient encore décrits ; les partenaires disposant de données de coloration sur les animaux séquencés pourraient alors progresser dans l'analyse entre phénotype et génotype ;
- Identifier la base génétique d'un phénotype à déterminisme mendélien en comparant des

sous-ensembles de populations définis sur la base de ce phénotype, ce qui suppose l'accord des fournisseurs de données ;

- Rechercher les signatures de sélection entre deux lignées divergentes : cette analyse sera généralement prise en charge par le partenaire ayant produit les lignées mais ce partenaire aura la possibilité de savoir si les variants candidats trouvés dans ses lignées sont uniques ou plus ou moins fréquents dans d'autres populations.

2.2.4. Imputation de la séquence à partir des génotypes

Comme dans le cas du projet Mille Génomes bovin, mutualiser les données des animaux séquencés permettra de mieux tirer parti des données privées de génotypage des candidats à la sélection. Le partage des séquences améliorera l'analyse des dispositifs privés de chaque membre du consortium de séquençage, et notamment ceux des partenaires privés, sans qu'il leur soit besoin de partager leurs données de génotypage ou de phénotypage. Cette application s'est révélée particulièrement utile pour la détection d'anomalies létales à l'état homozygote chez les bovins, en enrichissant les nombreuses données de génotypage par la prédiction de la séquence.

3. LE PROJET PILOTE FRANÇAIS Mille GENOMES GALLUS

3.1. Origine des données

Ce pilote s'appuie sur la mutualisation des données de séquence de 207 animaux issus de 9 projets de recherche financés par l'ANR ou l'Europe et conduits par des équipes françaises (Tableau 1). Chaque projet devient en quelque sorte un témoin pour les autres. Les animaux séquencés proviennent des 4 espèces du genre *Gallus*, si bien qu'on peut parler d'un projet Mille Génomes *Gallus*. Chaque jeu de données est décrit par un ensemble de 18 métadonnées (Tableau 2). L'ensemble de ces données est regroupé sur le serveur de la plateforme SIGENAE du département de Génétique Animale de l'INRA à Toulouse.

L'exploration des bases de données publiques a permis d'identifier 88 jeux de données supplémentaires de séquences individuelles disponibles pour le poulet, qui représentent principalement des races locales asiatiques (67 individus), quelques lignées commerciales (6) ou expérimentales (9) et 6 individus sauvages. On peut donc déjà disposer de 295 séquences individuelles ; toutefois la profondeur des séquences varie de 5 à 90X. En considérant une profondeur minimale de 10X pour garantir une information assez fiable, on doit retirer 20 jeux de données, dont 4 issus

du projet pilote français. Il reste alors un jeu de données de bonne qualité pour 275 animaux, soit du même ordre de grandeur que le run 2 du projet bovin.

3.2. Méthodes d'analyse

L'étape 1 (figure 1) a été réalisée pour les 207 jeux de données individuels des projets français en utilisant la version 5 du génome de référence de *Gallus gallus*, mais la version 6 étant disponible depuis la fin 2018, il est probable qu'un nouvel alignement soit nécessaire dans le futur. Les étapes 2 à 4 ont été réalisées comme décrit dans le figure 1.

Nous avons commencé par des analyses de diversité génétique pour illustrer l'intérêt de la mutualisation des données. Pour cela, nous avons conservé uniquement les SNP pour lesquels un identifiant était disponible dans la base dbSNP (environ 5 millions de SNP concernés). Le calcul des distances génétiques 'identité-par-état' a été conduit avec le logiciel plink (version 1.9) ainsi que le regroupement des distances pour obtenir le graphe multidimensionnel (MDS). Les analyses de structure de population ont été réalisées avec le logiciel ADMIXTURE (Alexander et al, 2009). Le nombre de groupes K (c'est-à-dire le nombre de populations ancestrales) a été estimé en utilisant la procédure de validation croisée disponible avec ADMIXTURE (option --cv) en faisant varier K de 1 à 30 et en choisissant la valeur de K permettant de minimiser l'erreur de validation croisée.

Le gène *MC1R* a été choisi comme exemple de gène candidat dont les variants vont être inventoriés. Pour définir les positions sur la séquence génomique, une comparaison (blastn) entre la séquence codante du gène (NM_001031462) et la séquence génomique du chromosome 11 (NC_006098.4) le contenant a été réalisée. Pour chaque population, la région génomique contenant la séquence codante (un seul exon) correspondant au gène *MC1R* sur le chromosome 11 (assemblage GalGal5 : NC_006098.4 - 11:19084582..19085526) a été extraite des fichiers issus de l'étape 4. Les fréquences alléliques ont été calculées avec le logiciel vcftools. La position des variants identifiés a été comparée aux positions des mutations déjà connues (Ling et al., 2003).

3.3. Résultats

3.3.1. Analyse globale de la diversité

Le nombre total de SNP détectés atteint 42 millions. Les jeux de données montrant les valeurs les plus faibles (moins de 10 millions de SNP) correspondent aux projets qui comprennent peu d'animaux (Fr-AgEncode, Epibird) ou des animaux de lignées

fermées ou sélectionnées depuis un grand nombre de générations (ChickSeq, UtOpIGe)). Les poulets de chair et les poulets labels montrent un peu plus de variation et les individus sauvages *Gallus gallus* montrent la plus grande variation (14 millions). Les 3 autres espèces sauvages montrent aussi un haut niveau de variation par rapport au génome de référence *Gallus gallus*, avec 12 à 14 millions de SNP.

3.3.2. Structuration de la diversité génétique

L'analyse des distances génétiques d'identité-par-état (IBS : Identity By State) et leur représentation par MDS plot (figure 2) permettent de séparer les populations selon leur niveau de variabilité sur le 1er axe : à gauche les poules sauvages et traditionnelles, puis les poulets de chair et enfin les pondeuses à droite. L'axe 2 sépare les pondeuses à œufs bruns commerciales (UtOpIGe) en haut à droite des pondeuses à œufs bruns expérimentales en bas à droite, avec un gradient pour les autres populations. Les données du projet UtOpIGe montrent des distances génétiques faibles entre individus. Le projet QTLDJ ajouté sur le graphique correspond à des séquences de mélanges pour des lignées expérimentales Leghorn.

L'analyse de structure de population fait apparaître 7 regroupements ou clusters (Figure 3). Le cluster 4 est défini par le projet ACRIGEN (poulet de chair moderne) et le cluster 5 par le projet UtOpIGe (pondeuses œufs bruns) alors que le cluster 2 est caractéristique du projet ChickSeq (lignée ancienne de poulet de chair) mais on détecte un apparentement faible avec d'autres clusters à l'intérieur de ce groupe. Les animaux sauvages *Gallus gallus* définissent le cluster 7 et ceux des 3 autres espèces le cluster 6. Les races locales ont un fort apparentement au cluster 7 et montrent un apparentement faible à modéré avec tous les autres clusters, y compris le cluster 6, ce qui traduit l'absence de sélection forte chez ces populations qui sont ainsi plus proches de l'ancêtre sauvage non sélectionné. Le cluster 3 correspond probablement à l'origine White Leghorn présente dans les projets Epibird et Fr-AgEncode. Enfin, le projet SABRE, constitué d'animaux croisés F1 entre des lignées de poulet label, montre logiquement une composition diversifiée du génome avec un apparentement à 6 clusters, assez proche de la situation observée pour les races locales.

Il est possible d'identifier quelques régions du génome qui sont encore partagées entre des races locales et des races très sélectionnées alors que les races très sélectionnées ont définitivement perdu une fraction de la diversité génomique de l'espèce *Gallus* et sont les plus éloignées de l'ancêtre sauvage.

3.3.3. Diversité du gène candidat *MC1R*

La séquence codante de *MC1R* comporte 1 exon pour lequel 10 SNPs ont été identifiés entre les populations étudiées (Tableau 3). On observe nettement la fixation de l'allèle en position 427 chez les pondeuses à œufs bruns qui sont les seules à avoir un plumage coloré rouge alors que les variants en position 636 et 637 sont fixés chez toutes les pondeuses et en ségrégation chez les poulets de chair. Le SNP en position 427 correspond au remplacement d'une thréonine par une alanine en position 143 du récepteur et a déjà été associé à la couleur rouge (Ling et al., 2003). Le variant en position 376 est le plus fréquent en Leghorn blanche et présent chez quelques poulets du chair du projet ChickSeq. La variante 644 est la plus fréquente chez les poulets labels du projet SABRE. Les poulets de chair ne sont jamais fixés pour aucun variant de *MC1R* ce qui est cohérent avec le fait qu'ils ont été sélectionnés pour une absence de couleur de plumage alors que les pondeuses à œufs bruns ont été sélectionnées pour une couleur de plumage rouge. La situation des Leghorn blanche est intermédiaire car les variants 636 et 637 sont présents chez les 4 animaux étudiés, un plus grand nombre d'individus devraient être génotypés pour confirmer la fixation de ces variants dans cette population.

4. EXTENSION DU PARTENARIAT

4.1. Gouvernance et modèle économique

Comme pour le Mille Génomes bovin, un accord de consortium est nécessaire pour pérenniser cette initiative et l'étendre à des jeux de données internationaux. Dans le cas du projet bovin, l'accord de consortium est signé par toutes les institutions apportant des données mais le comité de pilotage est restreint aux membres fondateurs qui ont constitué le consortium initial et apporté les données du premier run. L'objectif de l'accord est de définir les droits et devoirs des partenaires, de fixer les conditions d'entrée, d'organiser la coordination des études afin de définir celles qui sont réalisées en commun et celles qui peuvent être réalisées en interne par chaque partenaire, et de préciser les conditions de valorisation des résultats. En particulier, toutes les études doivent être connues du comité de pilotage, doivent citer l'origine des données et, en cas de propriété intellectuelle, une licence gratuite est attribuée à tous les membres et leurs partenaires ayant fourni des données, ce qui constitue aussi une forte motivation pour intégrer le consortium. Le consortium est limité à des organismes académiques mais la plupart sont associés à des sélectionneurs privés ou coopératifs dont les intérêts sont pris en compte car ils possèdent les animaux séquencés et ont parfois contribué au financement des recherches. Des

sélectionneurs concurrents entre eux ont donc accepté de mutualiser leurs données de séquence afin d'en extraire plus d'informations pour leur usage interne.

Le projet pilote français implique des équipes (Tableau 1) qui se réunissent régulièrement dans le cadre du réseau de génétique avicole de l'INRA et actuellement coordonné par l'unité BOA à Tours. Seules les données du projet UtOpIGe ont été obtenues sur une lignée commerciale de l'entreprise NOVOGEN dans le cadre d'un projet financé par l'ANR qui a fait l'objet d'un accord de consortium. Les conditions d'utilisation de ces données sont donc déjà définies par cet accord initial qui doit continuer à être respecté pour la participation au projet Mille Génomes. Il en est de même pour tous les projets apporteurs de données sur d'autres populations pour lesquels un accord contractuel aurait déjà été établi.

Les conditions d'entrée actuellement utilisées pour le projet pilote sont les suivantes :

- Données de séquence individuelles ayant une profondeur minimale de 10X ;
- Apport minimum de 300X (30 animaux à 10X), considérant que le seuil initial de 100X choisi pour le projet bovin correspond environ à 300 Gb et donc à 300 génomes de poulet ;

Les données de séquençage de mélange sont acceptées en supplément, elles sont plus difficiles à mutualiser car elles sont soumises à un aléas supplémentaire, avec une profondeur moyenne du génome généralement faible et hétérogène entre les individus d'un mélange.

Le modèle économique est calqué sur le projet bovin : les membres financent l'acquisition des données qu'ils apportent et le coordinateur finance le stockage des données et la constitution du fichier global des variants, avec sa mise à jour en cas d'évolution de la séquence de référence. Le coût de stockage annuel de 500 séquences (10X) varie de 6000 à 12000 euros selon que les données restent immédiatement accessibles (stockage 'chaud') ou non (stockage 'froid'), ce qui représente moins de 1% du coût total d'obtention de 500 séquences 10X, estimé à 1,75 Millions d'euros. Le coût le plus important est celui des ressources humaines en charge des analyses.

4.2 Prospection de nouveaux partenaires

Plusieurs partenaires internationaux potentiels, publics ou privés, ont déjà été approchés. Il s'agit de :

- l'Université de Wageningen aux Pays-Bas et de l'entreprise Hendrix Genetics avec un total d'au moins 250 séquences individuelles ayant une profondeur au moins égale à 10X ; l'entreprise Hendrix Genetics a financé l'acquisition de la plus grande partie des données proposées mais n'exige pas d'exclusivité et le

consortium pourra rester ouvert à d'autres partenaires privés qui financeraient l'apport de données ;

- le Friedrich Loeffler Institute, Mariensee, et l'Université de Göttingen en Allemagne avec 50 séquences individuelles de profondeur au moins égale à 10X;

Ces institutions ont confirmé leur intérêt pour participer et une première version d'accord de consortium inspiré de celui du projet bovin leur a été soumise avec peu de corrections demandées. De plus, ces institutions sont partenaires du projet européen IMAGE, coordonné par l'INRA, qui finance le séquençage de 300 poulets (profondeur $\geq 10X$) représentant principalement des races locales des 3 pays, France, Allemagne, Pays-Bas. L'intégration de ces institutions et des données de IMAGE porterait le total de séquences individuelles dépassent le seuil de profondeur de 10X de 203 à 803, auxquelles s'ajoutent 72 jeux de données publiques. Des données pourraient également être apportées par le Roslin Institute (100) et l'université de Nottingham (nombre à préciser) mais leur accord de principe n'a pas été confirmé.

Au niveau international, les institutions les plus intéressées sont en Chine et en Corée. Les équipes américaines contactées n'ont pas souhaité s'engager ou bien ont proposé de commencer par stocker les données rassemblées, ce qui n'était pas notre demande.

CONCLUSIONS

La cible de 1000 séquences de poulet en profondeur 10X est atteignable à court terme. L'exemple du modèle bovin a montré une dynamique d'attractivité croissante du projet encouragée par les premiers résultats obtenus sur l'identification d'anomalies génétiques. Dans le cas du projet pilote poulet, une telle analyse n'est pas réalisable actuellement par insuffisance d'animaux génotypés dans les populations représentées par des animaux séquencés. En revanche, deux types d'analyse seraient dès maintenant réalisables : (1) identifier les régions du génome et les SNP candidats déterminant des caractères discriminant les populations, tels que la couleur de coquille (2) produire un inventaire des variants structuraux dans le génome du poulet, qui n'a jamais encore été réalisé sur un effectif aussi important représentatif des principaux types de populations, sauvages et domestiques. La première action à engager est la signature de consortium avec les différentes institutions.

REMERCIEMENTS

Les auteurs remercient les porteurs de projets nommés dans le Tableau 1 qui ont apporté leurs données à ce projet pilote.

RÉFÉRENCES BIBLIOGRAPHIQUES

Alexander D., Novembre J., Lange K., 2009. *Genome Res.*, (19), 1655-1664.

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., Ponce C.H., Sargolzaei M., et al., 2018. *Nature Genet.*, (50), 362-367.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., et al., 2014. *Nature Genet.*, (46), 858-865.

Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C., et al., 2004. *Nature*, (432) 695-716.

Ling M.K., Lagerström M.C., Fredriksson R., Okimoto R., Mundy N.I., Takeuchi S., Schiöth H.B., 2003. *Eur. J. Biochem.*, (270), 1441-1449.

Vignal A., Besbes B., 2005. 6èmes Journées de la Recherche Avicole, St Malo, 30-31 Mars, 2005.

Tableau 1 - Description des jeux de données individuelles rassemblés dans le projet pilote Mille Génomes Gallus

Projet	Correspondant (unité INRA ou entreprise)	Type de population	Nombre d'individus	Profondeur moyenne
FR-AgEncode	E. Giuffra, (GABI), S. Lagarrigue (PEGASE)	pondeuses Leghorn, lignée expérimentale	4	38
EpiBird	F. Pitel (GenPhySe)	pondeuses Leghorn, RIR et croisement F1, lignées expérimentales	12	15
UtOpIGe	P. Le Roy (PEGASE) T. Burlot (NOVOGEN)	pondeuses œuf bruns, lignée commerciale	90	14
FEEDaGENE	S. Lagarrigue et T. Zerjal (GABI)	pondeuses RIR, lignée expérimentale	19	46
ACRIGEN	E. Le Bihan-Duval (BOA)	poulet de chair, lignée expérimentale dérivée d'animaux d'un sélectionneur	16	18
CHICKSEQ	S. Lagarrigue	poulet de chair, lignée expérimentale	24	15
SABRE	S. Mignon-Grasteau (BOA)	poulet label croisement F1 entre 2 lignées expérimentales dérivées d'une lignée commerciale	6	29
Domestichick	M. Tixier-Boichard (GABI)	races locales (tous les continents : projet AvianDiv) individus sauvages des 4 espèces du genre <i>Gallus</i>	36	33
Données publiques		races locales asiatiques, lignées commerciales, lignées expérimentales, individus sauvages,	88	20
TOTAL			295	

Tableau 2 - Métadonnées utilisées pour décrire les jeux de données

catégorie	métadonnée
Jeu de données	Nom du projet
population	Espèce
	Statut : domestique ou sauvage
	Type : non sélectionnée, locale, commerciale, expérimentale, croisement
	Nom : standardisation à mettre en place en l'absence de référentiel
individu	Identifiant : format à standardiser
	Sexe
	Pedigree : oui/non
	Identifiant_père
	Identifiant_mère
échantillon ADN	Mélange
	Nombre d'individus/ mélange
séquençage	Technologie
	Date d'obtention des données
	Longueur des fragments en bp
	Méthode paired-end : oui / non
	Profondeur : nombre de génomes-équivalents
	Code md5

Tableau 3 - Fréquence des variants SNP trouvés dans la séquence codante de MC1R. Une valeur 0 indique la présence homozygote de l'allèle de référence.

position	Utopige	FEEDaGENE	FR-AgEncode	Acrigene	CHICKSEQ	SABRE
	Pondeuses œufs bruns		Leghorn blanche	Poulets de chair		
69	0	0	0	0,75	0,35	0,5
212	0	0	0	0,75	0,34	0,5
274	0	0	0,88	0,75	0,53	0,5
376	0	0	0,88	0	0,07	0
398	0	0	0,13	0,25	0	0,33
427	1	1	0	0	0,41	0,17
636	1	1	1	0,25	0,475	0,5
637	1	1	1	0,25	0,475	0,5
644	0	0	0	0,03	0,07	0,5
834	0	0	0,13	0	0	0

Figure 1 - Analyse des données de séquençage du génome entier utilisées pour le projet Mille Génomes Poulet. Après le séquençage, l'étape 1 identifie les fragments obtenus en les alignant sur le génome de référence. L'étape 2 compare les fragments localisés à la séquence de référence pour en identifier les variations. L'étape 3 applique différents filtres pour éliminer les faux positifs. L'étape 4 (facultative) est l'annotation fonctionnelle des variants.

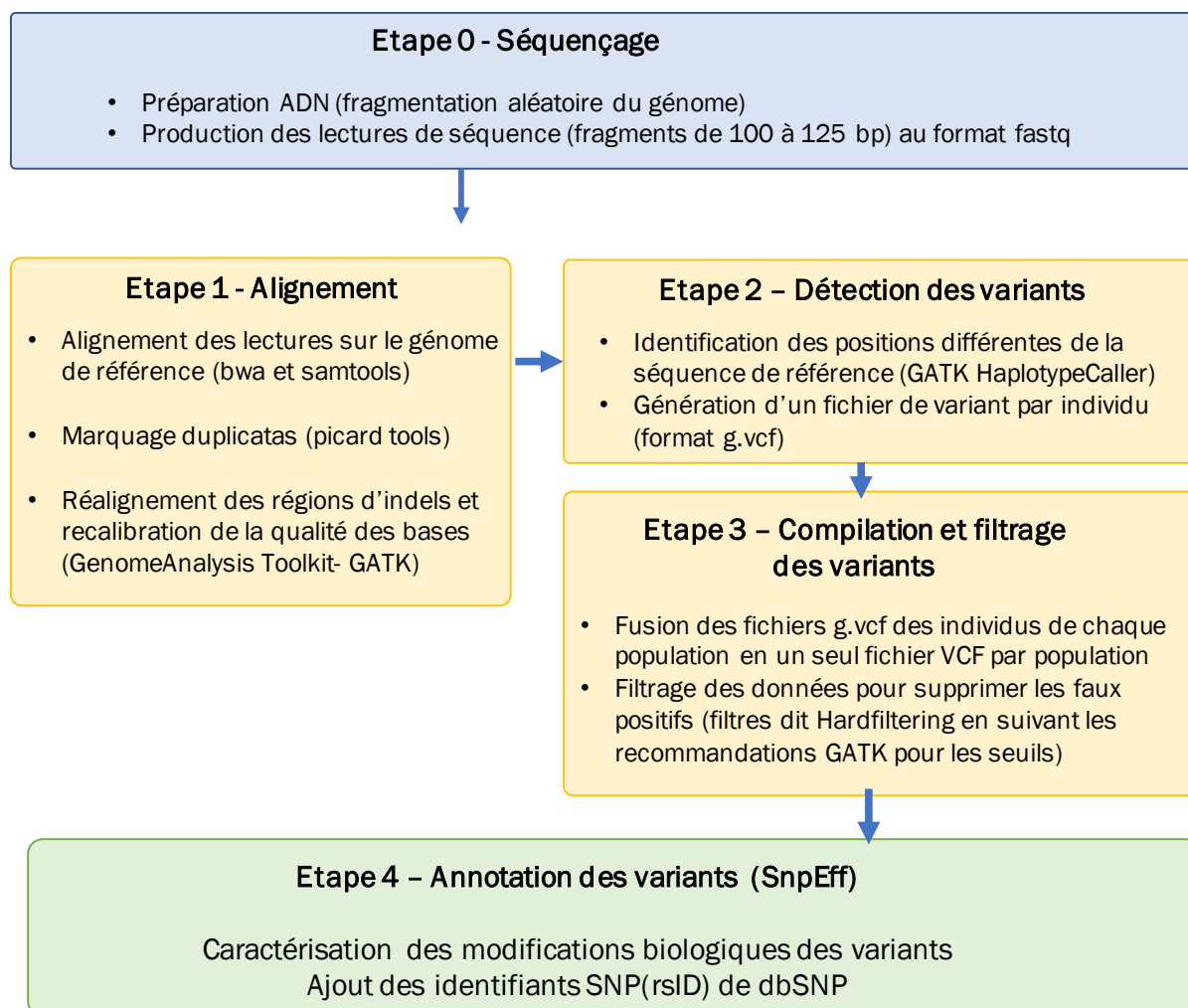


Figure 2 - Projection des individus en fonction des distances génétiques d'identité par état. Le symbole 'x' désigne les individus sauvages du projet Domestichick.

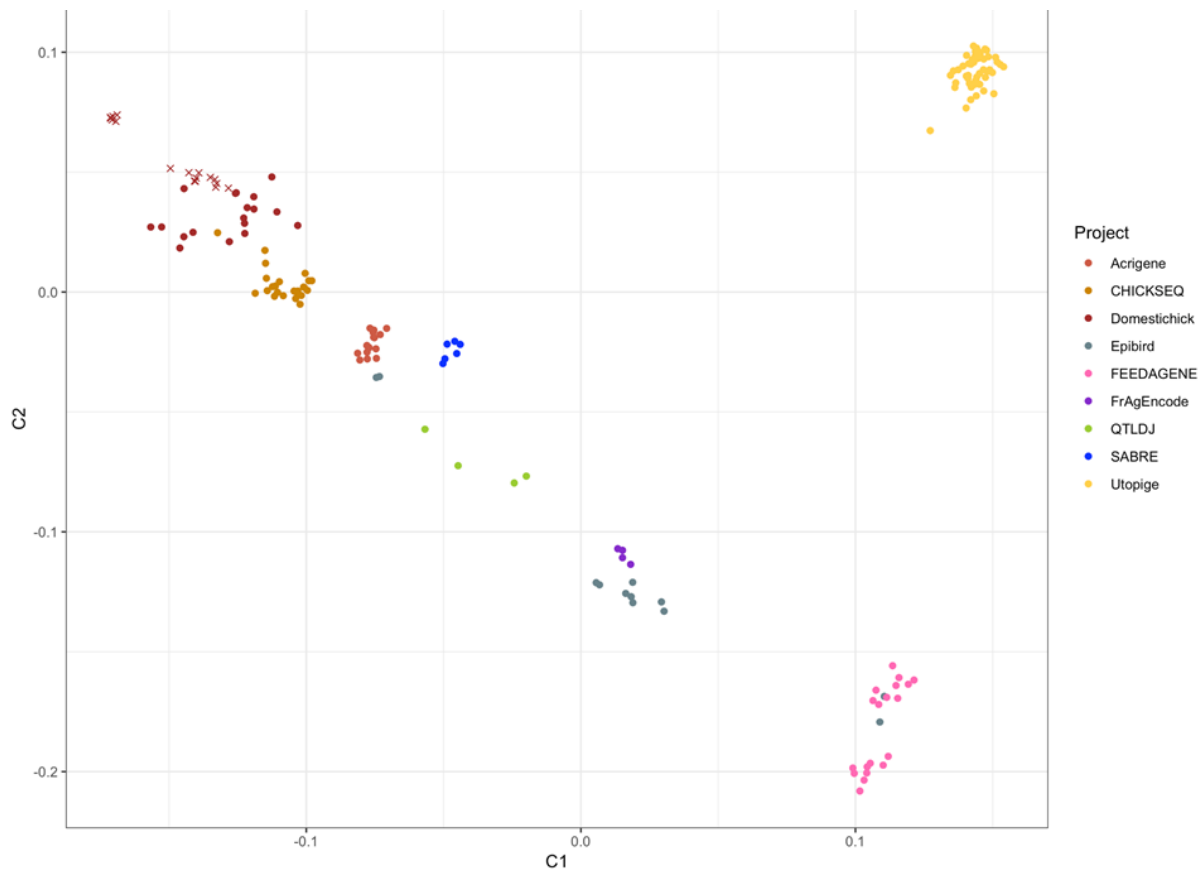


Figure 3 - Résultats de l'analyse de structure montrant l'existence de 7 groupes génétiques différents (clusters) ; un animal est représenté par une colonne, les segments de couleur différente dans une même colonne représentent la fraction d'appartenance du génome de l'individu aux différents groupes génétiques

