Utilisation d'une approche de génotypage RADSeq pour le clonage positionnel de deux gènes à effet majeur chez la caille, responsables des phénotypes " diabète insipide " et " couleur de la coquille céladon "

Anna Marissal

AGROCAMPUS OUEST

☐ **CFR Angers**

☒ **CFR Rennes**

Année universitaire : 2015 – 2016

Spécialité : SCMV

Sciences Cellulaire et Moléculaire du Vivant

Spécialisation (et option éventuelle) :

Génétique

Mémoire de Fin d'Etudes

☒ 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)

Utilisation d'une approche de génotypage RADSeq pour le clonage positionnel de deux gènes à effet majeur chez la caille, responsables des phénotypes « diabète insipide » et « couleur de la coquille céladon »

Par : Anna MARISSAL

Soutenu à Rennes, le 13/06/2016

Devant le jury composé de :

- Grégory EOT-HOULLIER
- Sandrine LAGARRIGUE
- Frédéric LECERF
- Laurent RICHARD-PARPAILLON

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].

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).

Date et signature du maître de stage[2] :

26/05/16

Institut National de la Recherche Agronomique
UMR GenPhySE - Génétique Physiologie et Systèmes d'Elevage
24 chemin de Borde Rouge - Auzeville Tolosane
CS 52627
31326 CASTANET TOLOSAN CEDEX
FRANCE

Droits d'auteur :

L'auteur[3] autorise la diffusion de son travail

☒ 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)

Date et signature de l'auteur :

26/05/2016

Autorisation de diffusion par le responsable de spécialisation ou son représentant :

L'enseignant juge le mémoire de qualité suffisante pour être diffusé

☐ 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

Résumé long

The goal of the present study is to locate two major genes in Japanese quail *Coturnix japonica*, responsible for the traits "celadon color shell" and "diabetes insipidus", knowing that previous studies have shown that these traits were determined by autosomal recessive mutations.

The "celadon" trait in quail is characterized by the production of an egg devoid of the brown pigment normally present on the surface, which lets appear the blue-green underlying color of the shell. The discovery of this mutation would allow a first approach to understand the complex mechanism of eggs camouflage in wild birds. The "diabetes insipidus" character is a disease characterized by excessive thirst (polydipsia) and the excretion of abnormally large volumes of diluted urine (polyuria). It remains at low frequency in poultry farms and is very pronounced in females. A line of diabetic quails (di/di) has been developed for several years at the National Institute of Agricultural Research (INRA), and represents a real opportunity to discover the origin of this disease in birds. Although this character does not appear to be associated with a decrease of zootechnical performances, it would allow a better understanding of the relationship between drinking and excretion mechanisms for a possible improvement of litter quality, which is an important aspect for the environmental performance of poultry farms.

The population of quails studied here is composed of 192 individuals, distributed over 3 generations: 10 F0 (5 grandfathers of a line "celadon" and 5 grandmothers of a line "diabetes insipidus"), 24 parents, all of the wild phenotype for these two characters, and 158 female descendants for whom 25% have the mutant phenotype for each traits. As the new high-throughput genotyping tools, such as SNP (Single Nucleotide Polymorphism) micro-arrays, are not yet available for less studied agronomic species such as quail, the genotyping of all these individuals has been carried out with a recent method of genotyping by sequencing, the RADSeq (Restriction site-Associated DNA sequencing). This technique allows an access to the same sub-representation of the genome of each individual by carrying out an enzymatic digestion (here with the enzymes EcoRI and TaqI) and a precise selection on the size of the resulting fragments. The short regions adjacent to the restriction sites are then sequenced in order to detect polymorphisms and to genotype at the same time a group of individuals for these markers. The detection of SNP from the sequenced regions was achieved using the software *Stacks*, first *de novo*, then with the reference genome of the quail (CoJa2), published during the internship. A panel of 121,157 SNP was obtained *de novo*, and 233,339 SNP with the reference genome. The RADSeq technique, implemented for the first time on

this species, has managed to detect a large number of markers without prior knowledge on the genome studied. From the genotypes obtained from the 192 individuals, Genome-Wide Association Studies (GWAS) were performed using the statistical software *R*, to locate the regions of *ce* and *di* mutations. A set of 36,244 SNP evenly distributed on all assembled chromosomes of the quail reference genome were used for the association analyses.

Regarding the "diabetes insipidus" trait, no peak significantly associated with this trait appeared using a polygenic linear model, but a peak appeared on chromosome 17 with a recessive linear model which does not take into account the bias due to kinship. Linkage analyses, using the software *Merlin*, will be conducted to try to get a better localization of the region of interest. However, new segregation analyses should be conducted to test the hypothesis of a major gene for that trait if no result appears with sufficient significance.

Concerning the "celadon" trait, some SNP significantly associated have been detected on the assembled part of chromosome 16 of CoJa2 and on two scaffolds (assembled fragments of the genome) whose places are still unknown on the reference genome. Linkage analyses carried out using the software *CriMap* have shown that these two scaffolds were indeed related to the currently assembled portion of chromosome 16 in quail. Using the available markers, we tried to reassemble this chromosome 16 and to insert these scaffolds by creating a genetic map of the region. Then we could compare the haplotypes of recombinant children to their phenotype using the software *Yapp*. This allowed us to refine the region of interest for the "celadon" trait in a broad area of 18 cM located between the two scaffolds for which no sequence is currently available. Additional analyses have to be conducted (*de novo* sequencing, comparative mapping between species...), waiting for a better assembly of the genome of *Coturnix japonica*, to locate more precisely the mutation.

Remerciements

Je tiens tout d'abord à remercier vivement ma maître de stage, Frédérique Pitel, pour son accueil, son fameux optimisme, et surtout la très grande disponibilité qu'elle m'a accordée tout au long de ce stage.

Je remercie chaleureusement Maria Bernard, bioinformaticienne à l'INRA de Jouy-en-Josas, qui a développé les scripts nécessaires à l'utilisation du logiciel Stacks, et sans qui ce stage ne m'aurait pas été accessible. Je la remercie aussi pour toutes les réponses et solutions qu'elle a bien voulu m'apporter lors de l'utilisation du logiciel, ainsi que pour sa grande réactivité malgré notre éloignement.

Je remercie également toute l'équipe GenEpi de l'INRA de Toulouse pour m'avoir chaleureusement accueillie pour ce long stage, ainsi que tous les autres stagiaires et thésards du bâtiment C pour les discussions animées du midi.

Enfin, je tiens à remercier Sandrine Lagarrigue, professeure de génétique à AgroCampus Ouest, pour m'avoir proposé ce stage en totale adéquation avec mes attentes.

Table des matières

<u>Liste des abréviations</u>

ADH = Antidiuretic Hormone

ADN = Acide Désoxyribonucléique

AQP = Aquaporine

AVP = Arginine Vasopressin

AVPR2 = Arginine Vasopressin Receptor 2

AVT = Arginine Vasotensin

CDI = Central Diabetes Insipidus

CE = Céladon

ddRADSeq = double digest Restriction-site Associated DNA Sequencing

DI = Diabète Insipide

FASTA = FAmily-based Score Test for Association

GBS = Genotyping By Sequencing

GWAS = Genome-Wide Association Study

Illumina®

NDI = Nephrogenic Diabetes Insipidus

NGS = Next Generation Sequencing

RADSeq = Restriction-site Associated DNA Sequencing

RADtags = Restriction-site Associated DNA tags

SNP = Single Nucleotide Polymorphism

Introduction

Les nouveaux outils de génotypage à haut débit qui ont émergé il y a quelques années, comme les puces à SNP (Single Nucleotide Polymorphism), ont largement permis d'améliorer les capacités d'analyse des génomes des espèces agronomiques d'intérêt. Cependant, les espèces moins étudiées, comme la caille, ne disposent pas encore de puce de génotypage. L'étude menée ici a pour but de localiser deux gènes majeurs chez la caille japonaise *Coturnix japonica*, ce qui nécessite une bonne couverture du génome par des marqueurs polymorphes. Il a donc été choisi de mettre en œuvre une méthode récente de génotypage par séquençage, le RADSeq (Restriction-site Associated DNA Sequencing), qui permet de détecter efficacement un grand nombre de marqueurs sans connaissance préalable du génome étudié.

Ce projet s'inscrit en continuité d'un précédent programme développé par l'équipe GenEpi (Génétique et Epigénétique moléculaires des espèces animales utilisées en croisement) de l'INRA de Toulouse, dont l'objectif est de caractériser le déterminisme moléculaire de la variabilité des caractères chez le porc, le lapin et les volailles. Le précédent projet visait à localiser chez la caille quatre mutations particulièrement intéressantes car sans homologues connus chez la poule: deux mutations du plumage, « rusty » et « curly », et deux mutations de la couleur de la coquille des œufs, « white » et « celadon » (Leroux et al., 2013). Cependant, la localisation du gène responsable du caractère « céladon », sur l'homologue chez la caille du chromosome 16 de la poule, demeurait incertaine du fait d'une significativité insuffisante à l'issue des premières analyses d'association. En effet, l'assemblage de ce micro-chromosome riche en séquences répétées était encore très incomplet. Ainsi, un nouveau programme a été initié afin de localiser, ou du moins de confirmer la localisation, du gène majeur responsable de ce premier phénotype « céladon », en plus de la localisation d'un second gène majeur, responsable du phénotype « diabète insipide ».

Le phénotype « céladon » (CE) correspond à la production d'un œuf bleu-vert chez la caille japonaise. La mutation à l'origine de ce caractère supprime complètement le dépôt de la pigmentation brune de surface normalement présente sur les œufs de caille, et laisse ainsi apparaître la couleur verte sous-jacente de la coquille. Cette pigmentation étant très répandue chez les oiseaux pour le camouflage des œufs, la découverte de cette mutation chez la caille serait une première approche pour la compréhension du déterminisme génétique de ce caractère complexe de pigmentation des œufs.

Le « diabète insipide » (DI) est une maladie caractérisée par une soif excessive (polydipsie) et l'excrétion de quantités anormales d'urine très diluée (polyurie), qui aboutit notamment à un taux de sodium dans le sang trop important, causé par des pertes d'eau trop

importantes (Babey et al., 2011). C'est un caractère qui se maintient à faible fréquence dans les élevages commerciaux de volailles et qui est très marqué chez les femelles. Des travaux menés à l'INRA ont permis de développer et de caractériser une lignée de cailles diabétiques (di/di) (Minvielle et al., 2007), qui constitue une ressource unique pour découvrir la mutation responsable du diabète insipide, encore inconnue chez les oiseaux. En effet, la proximité phylogénétique de *Gallus* et *Coturnix* et les recherches récentes menées sur diverses mutations communes aux deux taxons laissent penser qu'un même gène pourrait être impliqué dans les deux espèces. Chez l'Homme, le diabète insipide est le plus fréquemment dû à une production insuffisante d'ADH (hormone antidiurétique) au niveau de l'hypothalamus, qui aboutit à un diabète insipide dit « neurogène » ou « central » (CDI) (Babey et al., 2011). Mais il existe aussi un diabète insipide dit « néphrogénique » ou « rénal » (NDI), caractérisé par une incapacité à concentrer l'urine malgré une concentration normale d'ADH circulante. Cette perte de sensibilité des reins à l'AVP (ADH humaine) est causée dans 90% des cas par une mutation liée à l'X dans le gène codant pour son récepteur rénal, et dans 10% des cas par une mutation autosomale dans le gène codant pour l'aquaporine 2 (AQP2), qui permet la réabsorption d'eau et la concentration de l'urine au niveau des reins (Bichet, 2009). Une étude menée sur des cailles victimes de diabète insipide (Yang et al., 2008) a montré que la vasotensine (AVT), hormone antidiurétique chez les oiseaux, était correctement produite, mais que les reins des cailles malades avaient une zone médullaire moins développée et que l'AQP2 y était significativement moins exprimée, malgré le fait qu'aucune anomalie n'ait été repérée dans sa séquence codante. Aucune étude n'a encore été menée sur la potentielle implication de récepteurs à l'AVT dans cette maladie. Par conséquent, il s'agirait certainement de diabète insipide rénal, mais aucun gène candidat particulier n'est privilégié. C'est pourquoi il a été décidé pour ce nouveau projet sur la caille de mener une recherche de la mutation *di* à travers tout le génome grâce à la nouvelle technologie de RADSeq.

L'objectif de cette étude est donc de localiser chez *Coturnix japonica* les deux gènes majeurs responsables des caractères « céladon » et « diabète insipide », sachant que des travaux menés au préalable ont montré que ces caractères étaient déterminés par des mutations autosomales récessives (Ito et al., 1993; Minvielle et al., 2007). Cela devrait permettre de localiser la région de la mutation *ce*, clé d'entrée à la compréhension du mécanisme de camouflage des œufs chez les oiseaux sauvages. Dans le même temps, la mutation *di* devrait également être localisée. Bien que ce caractère ne semble pas être associé à une diminution des performances zootechniques, cela permettrait de mieux comprendre la relation entre les mécanismes d'abreuvement et d'excrétion pour une possible amélioration de la qualité des litières, aspect important pour la performance environnementale des élevages de volailles.

I. Matériel et méthodes

1.1. Matériel biologique

Les cailles utilisées dans ce programme appartiennent à l'espèce *Coturnix japonica* et ont été élevées au Pôle Expérimental Avicole de Tours (PEAT). Cinq mâles F0 de la lignée céladon (homozygotes ce/ce) et cinq femelles F0 de la lignée diabète insipide (homozygotes di/di) ont été croisés (1 mâle pour 1 femelle), pour obtenir une génération F1 de cailles toutes hétérozygotes pour les deux mutations (ce+/ce et di+/di), donc toutes porteuses du phénotype sauvage. Puis, 5 mâles F1 ont été croisés avec 20 femelles F1 (1 mâle pour 4 femelles), en évitant les croisements entre apparentés. Une des 20 mères est morte durant l'expérience. Un total de 158 femelles F2, réparties en 5 familles paternelles, ont ainsi été obtenues entre juillet et septembre 2013. Ces F2 présentaient bien les 4 types de phénotypes attendus: 10 [CE, DI] (ce/ce et di/di), 33 [CE, WT] (ce/ce, et di+/di ou di+/di+), 33 [WT, DI] (ce+/ce ou ce/ce, et di/di), et 82 [WT, WT] (ce+/ce ou ce+/ce+, et di+/di ou di+/di+), sachant que des proportions 10/30/30/88 étaient attendues selon la 2ème loi de Mendel. Ainsi, nous disposons au sein de la même population de 43 femelles de phénotype mutant et 115 femelles de phénotype sauvage pour chacun des deux caractères CE et DI.

La couleur de l'œuf et la polyurie (caractérisée par des fientes liquides) ont été évaluées pour ces 158 F2. Des prélèvements de tissus (reins en particulier) ont été réalisés pour des analyses fonctionnelles ultérieures, et des échantillons de sang ont été envoyés à l'INRA de Toulouse, où l'équipe GenEpi a effectué les extractions ADN afin de réaliser ensuite un RADSequencing.

1.2. RADSequencing

Il est vrai que l'obtention des librairies de séquençage a été réalisée avant le stage. Cependant, comme la technique de ddRADSeq est un aspect majeur du projet, et que la compréhension de cette technique est nécessaire à la compréhension des analyses effectuées par la suite, une explication de cette méthode a été développée ici.

1.2.1. Point sur la technique de ddRADSeq

Les méthodes de génotypage par séquençage d'un grand nombre d'individus partagent toutes le fait d'utiliser des enzymes de restriction pour cibler un sous-ensemble de loci à

séquencer (Da Costa et al., 2014). Elles permettent de découvrir de nouveaux marqueurs de polymorphisme et de génotyper dans le même temps un ensemble d'individus pour ces marqueurs. Le RADSeq est une technique de génotypage par séquençage particulièrement utilisée lors de l'étude d'espèces ne possédant pas de génome de référence. Elle cible à travers tout le génome une courte séquence reconnue par une enzyme de restriction choisie, pour ensuite séquencer les courtes régions adjacentes à ces sites, le but étant ensuite d'y détecter des polymorphismes. Baird et al. (2008) ont profité de l'émergence des technologies de séquençage nouvelle génération (NGS), et notamment du séquençage massivement parallèle (technologie Illumina), pour faciliter l'utilisation du RADSeq et ainsi permettre à cette technique de devenir particulièrement intéressante pour la découverte rapide de milliers de SNP chez de nombreux individus. Puis, de nouvelles méthodes, sur la base de la technique mise au point par Baird et al., ont émergé pour élargir encore le champ d'application du RADSeq, l'objectif étant de diminuer la fraction du génome à séquencer pour chaque individu, afin d'augmenter, pour le même coût, le nombre d'individus échantillonnés (pour une meilleure puissance statistique). La technique utilisée dans ce projet correspond à la méthode de ddRADSeq (double digest RADSeq) mise au point par Peterson et al. (2012), présentée Figure 1.

En RADSeq classique (Baird et al., 2008), l'ADN est digéré avec une seule enzyme de restriction, et les fragments ainsi obtenus sont flanqués à leurs extrémités d'un même adaptateur P1. En ddRADSeq (Peterson et al., 2012), deux enzymes de restriction sont utilisées simultanément, et les fragments obtenus reçoivent deux types d'adaptateurs à leurs extrémités : un adaptateur P1 spécifique du site de restriction de l'enzyme 1, et un adaptateur P2 spécifique du site de restriction de l'enzyme 2. Les librairies issues des différents individus sont ensuite regroupées. En RADSeq classique, tous les fragments sont alors redécoupés aléatoirement en fragments plus petits avant d'être sélectionnés selon la taille désirée pour le séquençage. La méthode de ddRADSeq se dispense de cette étape de coupure aléatoire, en effectuant directement une étape de sélection plus fine. Se passer d'une étape comme celle-ci permet non seulement de diminuer les coûts de production des librairies, mais aussi d'éviter des pertes d'ADN, ce qui autorise la construction des librairies à partir de peu de matériel de départ (100 ng d'ADN ou moins selon Peterson et al.). Vient enfin l'étape d'amplification.

La librairie finale se compose donc de fragments résiduels ayant une taille adaptée pour un séquençage Illumina, et possédant à leurs extrémités les amorces Illumina permettant de démarrer le séquençage. Il est à noter que seuls les fragments possédant à la fois P1 et P2 seront séquencés. En effet, un fragment dont les extrémités ont été coupées par la même enzyme aura des adaptateurs identiques, ce qui empêchera la formation de ponts lors de l'étape

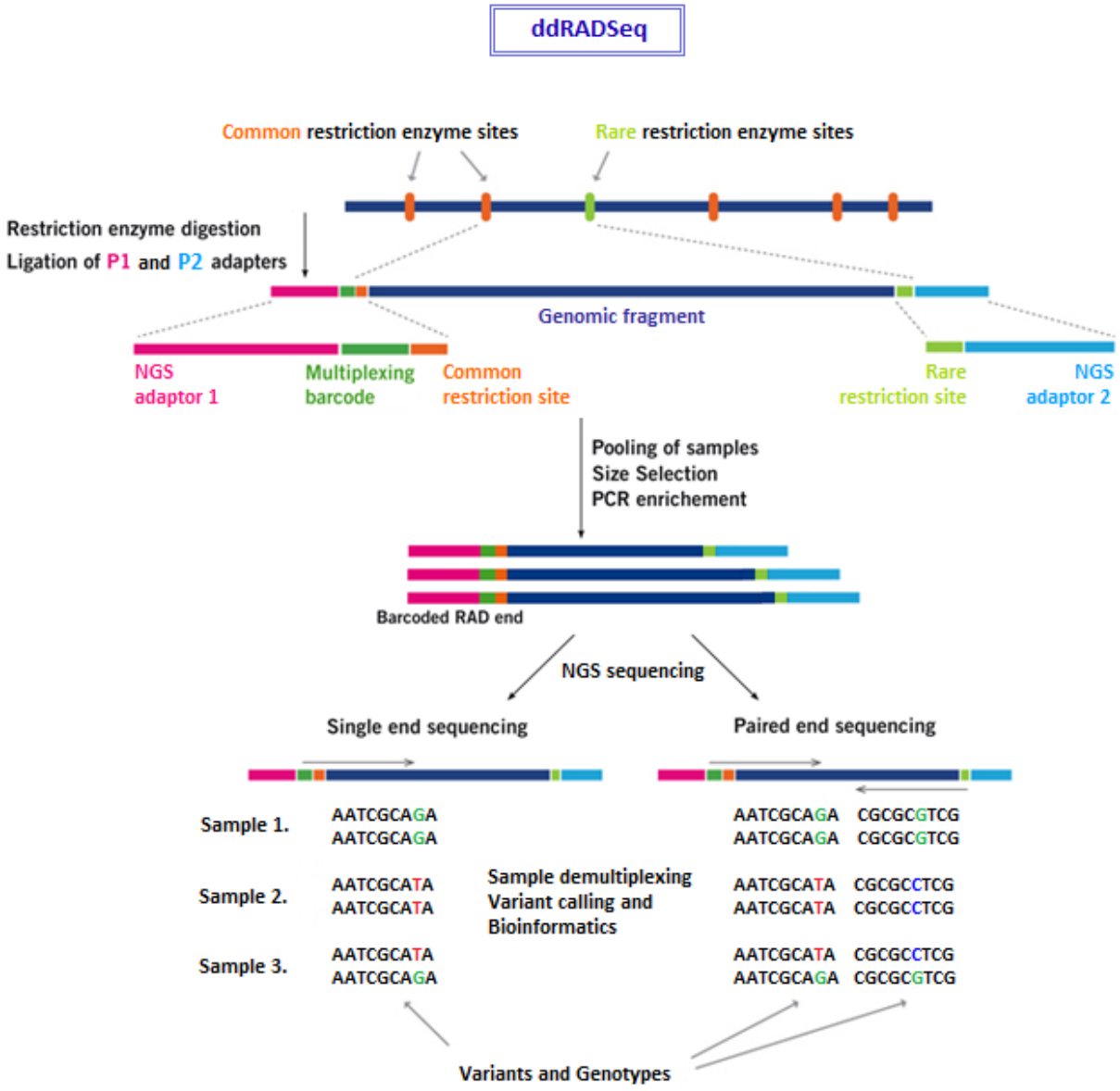


Figure 1 : Schéma illustrant le principe d'un ddRADSeq. La construction de la librairie en ddRADSeq s'opère en 5 étapes. L'ADN génomique est digéré par deux enzymes de restriction simultanément, une reconnaissant un site de restriction fréquent, l'autre un site de restriction rare. Puis des adaptateurs P1 et P2, respectivement spécifiques des sites de restriction des enzymes 1 et 2, sont liés aux extrémités des fragments issus de la digestion enzymatique. Chaque adaptateur contient des sites qui serviront à initier l'amplification PCR et le séquençage Illumina, et un des deux adaptateurs contient le code-barres nécessaire à l'identification des échantillons. Dans ce projet, l'adaptateur P2 contient le code-barres, qui fait 6 nucléotides. Pour éviter les erreurs d'assignation dues à des erreurs de séquençage, tous les codes-barres diffèrent d'au moins 2 nucléotides. Les fragments de tous les échantillons sont ensuite regroupés, sélectionnés selon une fenêtre de taille précise, puis amplifiés. Seuls les fragments possédant des sites de restriction différents à leurs extrémités seront enfin séquencés, le plus souvent sur 100 pb à partir des sites de restriction (« paired end sequencing »). (Image : www.floragenex.com/rad-seq/)

d'amplification qui a lieu sur la puce avant le séquençage. Chaque fragment est ainsi séquencé sur 100 ou 150 pb à partir des sites de restriction, ce qui permet d'obtenir de courtes lectures qui seront ensuite analysées pour révéler des SNP potentiels.

Le principe du RADSeq est donc d'isoler et de séquencer des RADtags, courtes séquences ADN flanquant les sites de restriction spécifiques d'une endonucléase, afin d'identifier à travers tout le génome des SNP, marqueurs génétiques très utilisés aujourd'hui dans diverses analyses génomiques.

1.2.2. Choix des enzymes de restriction

Différentes densités de marqueurs peuvent être atteintes de par le choix des enzymes de restriction utilisées et de la taille des fragments sélectionnés. En effet, les enzymes de restriction possédant des petites séquences de reconnaissance (4-6 pb) couperont plus fréquemment que des enzymes reconnaissant des séquences plus longues (8 pb) (Peterson et al., 2012). Ainsi, pour une fenêtre de taille donnée, le nombre de régions à séquencer sur le génome sera égal à 2 fois le nombre de fragments obtenus (2 extrémités séquencées pour chaque fragment). Il faut ensuite multiplier ce nombre de régions par la profondeur voulue et par le nombre d'individus pour déterminer le nombre de lignes de séquençage nécessaires. Il y a donc une relation inversement proportionnelle entre la fraction de génome échantillonnée et le nombre d'individus qui peuvent être séquencés en une seule ligne. Le choix des enzymes de restriction et de la taille des fragments à sélectionner est donc très important pour déterminer la fraction de génome séquencée, et donc la quantité de SNP que l'on obtiendra.

Etant donné que le génome de référence de *Coturnix japonica* n'était pas encore publié au moment des manipulations, mais que celui d'une espèce proche, *Gallus gallus*, était disponible, une estimation *in silico* du nombre de fragments obtenus suite à la digestion enzymatique a été réalisée, dans le but d'identifier un couple d'enzymes permettant d'avoir une bonne couverture du génome. Un script Python a été réalisé, admettant les principaux arguments suivants : le génome utilisé, un couple d'enzymes de restriction, la fenêtre de taille des fragments à sélectionner. Après différents tests faisant varier le choix du couple d'enzymes et de la taille des fragments à sélectionner, il a été décidé d'utiliser les enzymes EcoRI (site de restriction long: G/AATTC) et TaqI (site de restriction court: T/CGA), et de sélectionner lors de la construction des librairies des fragments dont la taille serait comprise entre 220 et 420 pb, afin d'obtenir environ 75 000 fragments (soit 150 000 RADtags). Pour le séquençage, le nouveau séquenceur Illumina HiSeq 3000, récemment installé à la plateforme de séquençage de l'INRA de Toulouse, a été utilisé, permettant ainsi de séquencer 150 pb à

partir des extrémités des fragments, et d'obtenir une couverture d'environ 2% du génome. Un total de 192 individus a été échantillonné, et une profondeur minimale de 25X est attendue pour de bonnes analyses d'association par la suite. Il a donc été estimé qu'un total de 720 millions de lectures (150 000 extrémités x 25X x 192 individus) serait à séquencer. Sachant qu'une seule ligne de séquençage Illumina peut prendre en charge 300 millions de lectures, 3 lignes ont été utilisées, dont une qui s'est révélée défaillante.

1.3. Recherche de SNP *de novo*

1.3.1. Utilisation d'un logiciel de découverte de SNP de novo *: Stacks*

Le logiciel utilisé pour la détection des SNP est *Stacks*. Il utilise des courtes séquences ADN, comme des RADtags, pour identifier et génotyper des loci chez un ensemble d'individus, soit *de novo*, soit à l'aide d'un génome de référence (Catchen et al., 2011). Il est écrit en C++ et en Perl, et est disponible sur Genotoul, le cluster de la plateforme de génomique de Toulouse.

Les fichiers d'entrée doivent être au format FASTA ou FASTQ. Les données obtenues du séquenceur Illumina HiSeq 3000 sont déjà démultiplexées : un fichier FASTQ R1, regroupant les lectures EcoRI, et un fichier FASTQ R2, regroupant les lectures TaqI, ont été obtenus pour chacun des 192 individus pour chacune des 3 lignes de séquençage.

Stacks fonctionne ensuite en 5 étapes majeures. La première étape utilise le programme `process_radtags.pl`, qui analyse chaque lecture contenue dans les fichiers FASTQ et exclut de la suite de l'analyse celles de mauvaise qualité (lectures avec moins de 90% de confiance dans la fenêtre de lecture, nucléotides inconnus, codes-barres faux, sites de restriction déficients). Tous les R1 et R2 ayant passé les filtres qualité sont ensuite regroupés, indépendamment des paires, dans un unique fichier FASTQ par individu.

La seconde étape utilise le programme `ustacks`, qui se déroule indépendamment pour chaque individu (Fig. 2A-F). Plusieurs paramètres sont disponibles à cette étape lorsque *Stacks* est lancé *de novo* (Catchen et al., 2011) :

- Le paramètre ***m***, appelé « profondeur de pile », détermine le nombre minimal de lectures identiques à partir duquel une pile pourra être formée (Fig. 2A). Cela permet d'écarter momentanément les lectures avec d'éventuelles erreurs de séquençage (lectures « secondaires »). Nous avons gardé la valeur par défaut qui est à 3.
- Le paramètre ***M***, aussi appelé « distance intra-individuelle » ou « distance nucléotidique », détermine le nombre maximal de mésappariements autorisés entre deux piles uniques pour

former un locus putatif (Fig. 2B-C), et correspond à peu près au nombre de SNP attendus dans un locus. Sa valeur par défaut est 3, mais nous l'avons diminuée à 2 car cela nous semblait suffisant pour nos lectures, qui font 149 pb.

- Le paramètre ***--max_locus_stacks*** définit le nombre maximal de piles uniques au sein d'un locus putatif (Fig. 2D). Sa valeur par défaut est 3 dans la dernière version de *Stacks*. Cependant, nos individus étant diploïdes, nous avons choisi de ne prendre au maximum que 2 piles uniques pour former un locus putatif. Si trop de piles uniques se regroupent, *Stacks* redécoupe le locus en sous-groupes de 2 piles maximum, en gardant ensemble les piles avec le moins de distance nucléotidique entre elles et avec une profondeur similaire.
- Le paramètre ***N*** détermine le nombre maximal de mésappariements autorisés pour associer une lecture secondaire à un locus déjà formé (Fig. 2E), autrement dit le nombre d'erreurs de séquençage autorisées pour utiliser une lecture secondaire. Par défaut ce paramètre est à *M*+2 (nombre attendu de SNP + 2 erreurs de séquençage acceptées), mais nous avons décidé de ne pas autoriser l'utilisation des lectures secondaires en choisissant N=0.

L'étape suivante utilise le programme `cstacks`, dont l'objectif est de recenser les loci et les génotypes présents dans la population (Fig. 2G). Le catalogue peut n'être construit qu'avec les parents (dans le cadre d'un protocole familial), ou simplement une partie de la population représentative de la diversité existante, mais nous avons choisi d'y insérer l'ensemble des individus de notre population. Le paramètre ***n***, aussi appelé « distance inter-individuelle », définit le nombre de mésappariements autorisés pour regrouper deux loci dans le catalogue, en plus des positions variantes déjà connues (SNP détectés lors d'`ustacks`). Cela permet de découvrir de nouveaux SNP putatifs qui seraient homozygotes chez chacun des individus, donc non détectés à l'étape `ustacks`, mais hétérozygotes entre individus. La valeur par défaut pour ce paramètre est 0, mais nous avons préféré prendre 1, afin de ne pas « rater » un SNP dont les allèles seraient fixés dans la population.

La quatrième étape utilise le programme `sstacks`, qui identifie les loci et SNP pour tous les individus de la population (descendants ou nouveaux individus, mais aussi ceux ayant précédemment servi à construire le catalogue) (Fig. 2H).

La dernière étape permet de construire les fichiers finaux regroupant les génotypes de tous les individus de la population, qui serviront notamment pour les analyses d'association. Le programme `genotypes.pl` est souvent utilisé dans le cadre d'un protocole familial où seuls les parents ont servi à construire le catalogue : il identifie les génotypes présents chez les descendants à partir des SNP identifiés chez les parents, et peut effectuer des corrections sur les génotypes. Nous avons préféré utiliser le programme `populations.pl`, qui génotype simplement toute la population aux marqueurs informatifs identifiés.

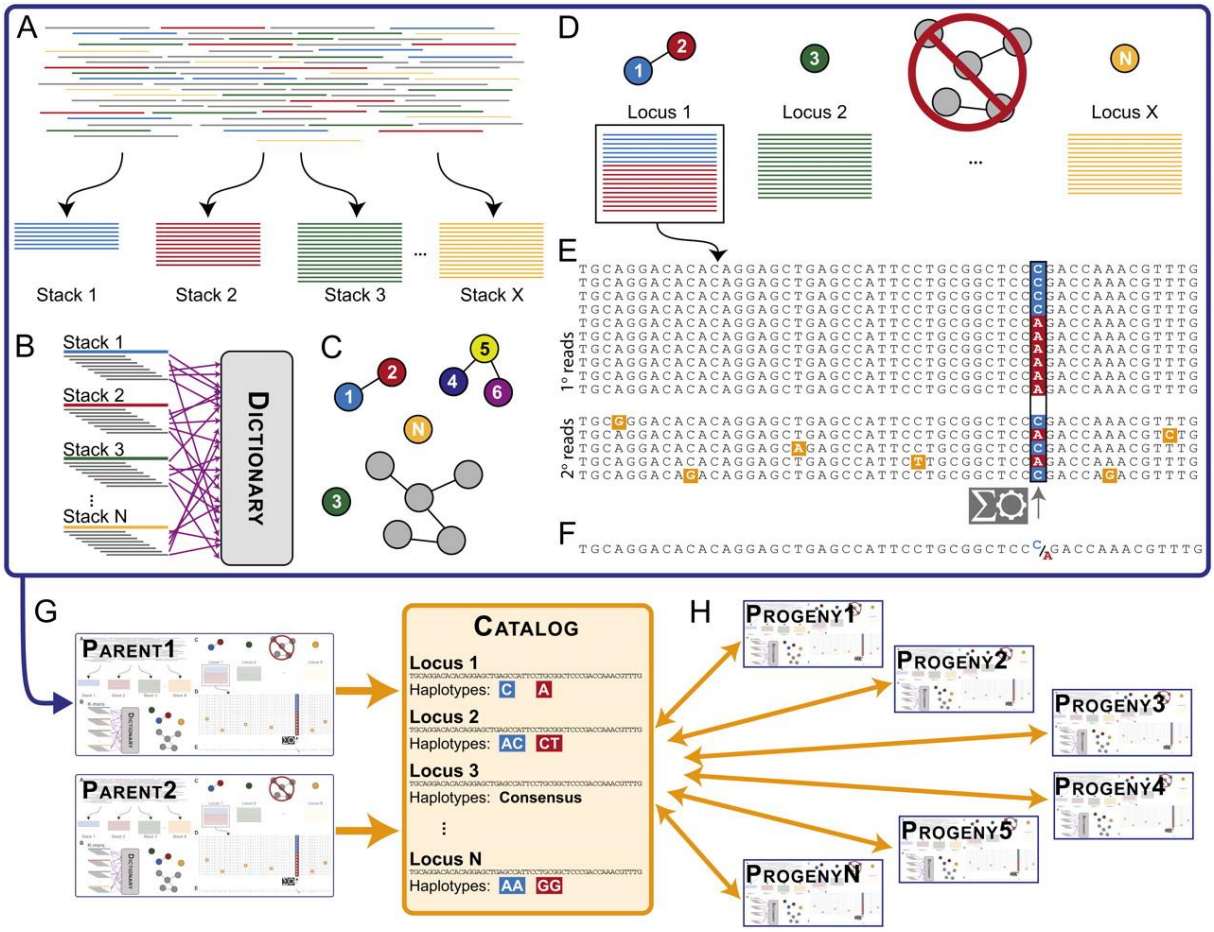


Figure 2 : Schéma récapitulatif du principe du logiciel *Stacks* lancé en *de novo.* A-F : programme *ustacks* ("unique stacks") effectué indépendamment pour chaque individu, G : programme *cstacks* ("catalog stacks"), H : programme *sstacks* ("search stacks"). (A) Les lectures R1 et R2 issues du fichier FASTQ de l'individu concerné sont groupées en piles de séquences parfaitement identiques. Deux allèles d'un même locus seront donc séparés en deux « piles uniques » différentes. Les piles de trop faible profondeur sont écartées, car elles correspondent certainement à des lectures contenant des erreurs de séquençage, ce qui les empêchent de se grouper avec d'autres lectures. (B) Chaque séquence unique est ensuite découpée en k-mères de 49 pb (pour nos lectures de 149 pb), qui sont répertoriés dans un dictionnaire. Puis, chaque pile est de nouveau découpée en k-mères qui sont comparés à ceux existant dans le dictionnaire. Si un nombre suffisant de k-mères entre deux piles s'apparient, les deux piles uniques sont regroupées pour former un locus putatif (D). (C) Cela peut s'illustrer comme des nœuds séparés par une certaine distance nucléotidique. (D) Un locus composé de deux piles uniques correspond théoriquement à un locus polymorphe (deux allèles), alors qu'un locus composé d'une seule pile correspond à un locus monomorphe. Si trop de piles se regroupent en un même locus, *Stacks* redécoupe la structure en plusieurs sous-ensembles. (E) Les lectures écartées à l'étape A du fait de potentielles erreurs de séquençage (lectures « secondaires ») sont récupérées et ajoutées aux loci putatifs pour augmenter leur profondeur. *Stacks* calcule ensuite à chaque position nucléotidique un maximum de vraisemblance afin de déterminer s'il y a un SNP ou non. (F) Enfin, pour chaque locus une séquence consensus est gardée et les données de SNP et d'haplotypes sont enregistrées. (G) Une liste exhaustive et unique de toutes les séquences consensus établies chez les parents lors d'*ustacks* est créée au sein d'un catalogue. Le 1er individu initialise le catalogue, puis pour chaque individu additionnel, ses loci sont découpés en k-mères et comparés aux loci déjà existants dans le catalogue. (H) Enfin, les loci de tous les individus de la population sont comparés au catalogue afin d'identifier les haplotypes à chaque locus. Les loci ambigus, c'est-à-dire s'appariant à plusieurs loci du catalogue, sont exclus (cas par exemple des séquences répétées). (Catchen et al., 2011)

Des scripts écrits en Bash permettant de lancer manuellement *Stacks* étape par étape ont été développés par Maria Bernard, bio-informaticienne de l'équipe SIGENAE à l'INRA de Jouy-en-Josas. Des statistiques descriptives sont ainsi disponibles à chaque étape : nombre de lectures utilisées, nombre de lectures primaires et secondaires, nombre de lectures par pile, par locus, profondeur, nombre de SNP et d'haplotypes, nombre de loci polymorphes…

1.3.2. Localisation des SNP

En sortie de *Stacks*, une liste de SNP est disponible (fichier plink.map), ainsi que les génotypes de tous les individus pour l'ensemble de ces marqueurs (fichier plink.ped). Cependant, la construction des loci ayant été réalisée *de novo*, on ne connaît pas la position de ces SNP sur le génome. Il faut donc les localiser avant de pouvoir ensuite localiser les deux gènes majeurs étudiés. Initialement, il était prévu de réaliser cette localisation à l'aide du génome de référence de la poule, mais un génome de référence pour la caille a été publié sur le site NCBI (http://www.ncbi.nlm.nih.gov/) au cours du stage, le 3 mars 2016, ce qui a permis de réaliser une localisation plus exacte des marqueurs trouvés.

Pour cela, un script en Bash, dont le principe est expliqué ici, a été réalisé. *Stacks* donne en sortie un fichier au format FASTA qui regroupe l'identifiant et la séquence consensus de tous les loci pour lesquels un SNP a été détecté lors du déroulement du pipeline. J'aligne dans un premier temps ces loci sur le génome de référence de la caille (CoJa2) à l'aide de BWA (Li et al., 2009). On obtient un fichier au format SAM dont on récupère les informations suivantes : numéro du locus, « flag » (0 si alignement forward, 16 si alignement reverse, 4 si non aligné), chromosome, position de la séquence sur le chromosome, qualité de l'alignement, ainsi qu'une information appelée « CIGAR ». De plus, *Stacks* donne un fichier plink.map qui contient uniquement la liste des identifiants des SNP trouvés. Cet identifiant est composé du numéro du locus où se situe ce SNP ainsi que de la position du SNP (comptée en base 0) dans la séquence de 149 pb. Toutes ces informations sont regroupées au sein d'un fichier unique dont chaque ligne correspond à un SNP.

Deux filtres qualité sont alors appliqués : les loci dont la qualité d'alignement est inférieure à 20 sont éliminés, ce qui permet dans le même temps d'éliminer les loci non alignés sur le génome de référence, et les SNP qui sont situés dans les 5 premières (position<5) ou les 5 dernières bases (position>144) de la séquence de 149 pb sont également éliminés, ces positions étant sources de biais de génotypage.

Je m'intéresse ensuite au CIGAR, qui se compose d'une série de chiffres et de lettres comme suit : 20S30M1I26M2D70M, avec S = nombre de bases avant la première base

alignée, M = nombre d'appariements et de mésappariements une fois la première base alignée, I = insertion, D = délétion. Le nombre de bases total est toujours de 149 (longueur de nos séquences consensus). Il s'avère donc délicat de déterminer la position exacte de chaque SNP, car il ne faut pas simplement ajouter la position du SNP dans la séquence à la position du début de la séquence consensus. Il faut tout d'abord prendre en compte le fait que la position du début de la séquence indiquée dans le fichier SAM ne compte pas les premières bases non alignées (S). Les SNP dont la position est comprise dans une région S (non alignée) sont d'ailleurs éliminés. Il faut aussi tenir compte des I et des D en fonction de la position du SNP dans la séquence (est-il situé avant ou après une insertion ou une délétion ?), car cela change la position réelle du SNP sur le génome de référence. Enfin, il faut faire attention à l'alignement forward (flag=0) ou reverse (flag=16), qui change le sens de lecture de la séquence consensus et donc le calcul. Le script développé donne un fichier récapitulatif indiquant notamment l'identifiant du SNP, l'identifiant du locus, le chromosome, la position du locus, la position du SNP dans la séquence, le nombre de bases non alignées (S), d'insertions (I), et de délétions (D) à prendre en compte, et enfin la position exacte du SNP sur le chromosome (position locus – S + position SNP – I + D pour un alignement forward). Enfin, un fichier plink.map regroupant uniquement les informations utiles (chromosome, identifiant du SNP, position exacte du SNP) est créé.

Il est à noter que le fichier plink.ped, qui regroupe les génotypes de tous les individus pour l'ensemble des marqueurs, doit également être modifié, notamment en supprimant les marqueurs éliminés par les filtres qualité lors de la construction du fichier plink.map.

1.4. Recherche de SNP avec un génome de référence

Stacks est un logiciel qui peut aussi utiliser un génome de référence, s'il en existe un, pour chercher des SNP à partir de lectures de RADSeq. Comme le génome de référence de *Coturnix japonica* a été publié durant le stage, cet aspect de *Stacks* a également été testé, car il apparaissait intéressant de comparer les résultats obtenus par les deux méthodes.

La première différence avec *Stacks de novo* est que les lectures doivent être alignées sur le génome de référence entre l'étape de filtre qualité (`process_radtags.pl`), et l'étape proprement dite de recherche des SNP, afin de générer des fichiers au format SAM. Le programme utilisé alors n'est pas `ustacks` mais `pstacks`, qui construit les loci à partir des positions d'alignement des lectures, plutôt qu'à partir de la distance nucléotidique entre des piles uniques. Le seul paramètre à entrer, aussi appelé ***m***, est le nombre minimal de lectures alignées au même endroit pour qu'une pile se forme. Cette fois, ce ne sont pas des piles

« uniques », mais directement des loci putatifs. Ce paramètre est à 1 par défaut, mais nous avons choisi de le mettre à 10 pour assurer une bonne profondeur minimale pour les loci. Il n'est d'ailleurs plus question de lectures primaires ou secondaires, car à partir du moment où l'alignement est réussi, *Stacks* conserve toutes les lectures alignées, même celles avec de potentielles erreurs de séquençage. Une fois les loci putatifs formés, la détection des SNP se déroule comme pour `ustacks`, c'est-à-dire qu'à chaque position nucléotidique un maximum de vraisemblance est calculé pour déterminer s'il y a un SNP ou non. L'étape du catalogue, qui détermine une liste unique de tous les loci présents dans la population, est bien plus rapide car il ne s'agit plus de comparer les loci deux à deux entre les différents individus mais simplement de regrouper les loci qui ont les mêmes positions d'alignement.

De même que pour *Stacks de novo*, on obtient en sortie un fichier plink.ped avec les génotypes de tous les individus pour l'ensemble des marqueurs trouvés, ainsi qu'un fichier plink.map complet, indiquant l'identifiant et la position de chaque SNP le long du génome.

1.5. Analyses d'association sur R

Les analyses d'association permettent de bien mettre en évidence et de localiser de façon précise des gènes majeurs (pour un dispositif animal adapté et un nombre de marqueurs informatifs suffisant). C'est donc ce type d'analyse qui a été effectué en premier lieu pour tenter de localiser les gènes responsables des caractères CE et DI.

Pour cela, on a utilisé le package GenABEL (Aulchenko et al., 2007), qui a été créé dans le but de faciliter les GWAS (Genome-Wide Association Study) sur *R*, logiciel puissant pour effectuer des analyses statistiques. Ce package permet notamment de mener différents contrôles qualité sur les données génotypiques avant l'analyse d'association proprement dite :

Contrôles qualité sur les individus :

- $callrate_{IND}$: correspond au taux de réussite du génotypage : nombre de SNP génotypés pour un individu / nombre total de SNP. Si un animal est génotypé pour trop peu de marqueurs (callrate < 50%), il est retiré de l'analyse. Nous avons choisi ici de mettre un seuil assez faible afin de garder l'ensemble des individus dans l'analyse ;
- *taux d'hétérozygotie* : nombre de fois qu'un animal est hétérozygote / nombre total de SNP : si le taux d'hétérozygotie est trop élevé par rapport à l'hétérozygotie moyenne, l'animal est retiré de l'analyse, car deux échantillons ADN ont peut-être été mélangés ;
- *identity by state* : taux de génotypes identiques entre deux individus : si deux animaux ont trop de génotypes identiques (IBS ≥ 0,95), ils sont retirés de l'analyse, car il y a sûrement eu une erreur dans la préparation de l'échantillon.

Contrôles qualité sur les marqueurs :

- $callrate_{SNP}$: nombre d'individus génotypés pour un SNP / nombre total d'individus. Si un marqueur est génotypé chez trop peu d'individus (callrate < 95%), il est retiré de l'analyse. Nous avons choisi ici un seuil élevé dans le but de ne garder que les SNP informatifs pour presque toute la population ;
- *minor allele frequency* : si la fréquence de l'allèle mineur est trop faible (MAF < 5%), il y a un risque d'erreur dans le calcul de l'effet du SNP, et ce marqueur est retiré de l'analyse ;
- *Hardy-Weinberg equilibrium* : si la déviation par rapport à l'équilibre d'Hardy-Weinberg est trop forte pour un SNP ($P_{HWE} < 0$), il y a peut-être une erreur, et le marqueur est retiré de l'analyse. Nous avons laissé ce seuil à 0 car cet équilibre n'est de toute façon pas respecté étant donné que nous avons un protocole familial et que les croisements n'ont pas été faits au hasard.

Le package GenABEL permet ensuite de faire une analyse d'association dans le cadre de populations ou de dispositifs familiaux. Les deux caractères étudiés, « diabète insipide » et « couleur de la coquille céladon », sont qualitatifs : les phénotypes des animaux sont binaires, c'est-à-dire codés 1 = affecté et 0 = non affecté. L'analyse d'association consiste dans ce cas à savoir si, pour chaque marqueur, un des allèles est significativement associé au phénotype des cas (individus affectés par la mutation) par rapport aux contrôles (individus non affectés). Ainsi, du fait du déséquilibre de liaison, si le phénotype observé est effectivement dépendant du génotype (p.value faible), cela signifie que le marqueur testé est proche de la mutation causale. Afin d'éviter d'éventuels biais dus à la stratification de la population, j'ai également pris en compte l'apparentement entre individus, qui est calculé à partir des génotypes. Les p.values associées à chaque marqueur sont donc calculées à l'aide du modèle linéaire mixte suivant : $Y = \mu + G\beta + A\gamma + E$, avec Y = phénotype, μ = moyenne de Y dans la population, $G\beta$ = effet du génotype/du SNP, $A\gamma$ = effet de l'apparentement (effets polygéniques), E = effets dus à l'environnement (erreur). Pour chaque SNP, l'hypothèse $\beta=0$ vs $\beta\neq0$ est testée (test de Student).

1.6. Analyses de liaison avec les logiciels CriMap et Yapp

Au vu des résultats des analyses d'association, j'ai tenté de reconstruire une carte génétique du chromosome 16 du génome de la caille à l'aide des marqueurs à ma disposition. Pour cela j'ai utilisé le logiciel *CriMap* (version 2.4), créé pour automatiser la construction de cartes génétiques multilocus (Green et al. 1990). Les loci présentant trop peu de méioses informatives ou trop d'erreurs mendéliennes on d'abord été écartés (option PREPARE). Puis,

pour chaque locus de 149 pb gardé, j'ai sélectionné le marqueur présentant le plus de méioses informatives. Une analyse de la liaison entre les marqueurs deux à deux (option TWOPOINT) a ensuite été lancée avec un seuil de LOD score de 3 (1000 fois plus de chances d'être liés que d'être indépendants génétiquement). Puis, l'option BUILD a été utilisée pour ordonner les marqueurs : deux marqueurs sont choisis pour donner un ordre de départ puis les autres sont ajoutés un par un à leur position la plus probable. Enfin, les options FLIPS et CHROMPIC ont été utilisées pour vérifier si une meilleure carte pouvait être obtenue, respectivement en intervertissant les marqueurs, et en observant la quantité de recombinaisons en résultant.

Dans un second temps, j'ai essayé de réduire la zone où pouvait se situer la mutation *ce* sur le chromosome 16, en observant les haplotypes des enfants recombinants. Pour cela j'ai utilisé le logiciel *Yapp* (développé par Bertrand SERVIN à l'INRA de Toulouse), qui reconstruit au sein de chaque famille les haplotypes les plus probables à partir d'une carte génétique déjà établie. Cette fois, tous les SNP des loci précédemment placés sur la carte génétique ont été utilisés.

II. Résultats

2.1. RADSequencing

La quantité de fragments qui devait être obtenue par ddRADSeq a été en réalité surestimée, car le premier script d'estimation *in silico* ne prenait pas en compte le fait que seuls les fragments avec des sites de restriction différents à leurs extrémités seraient correctement séquencés. Un second script tenant compte de cela a été réalisé par Céline NOIROT, bio-informaticienne au sein de la plateforme bioinformatique de l'INRA de Toulouse. Pour une taille de 220 à 420 pb, comme il avait été décidé au départ, une quantité de 25 700 fragments a été estimée. Mais, l'étape de sélection des fragments selon leur taille n'ayant pas été réalisée aussi finement que prévu, plus de fragments ont finalement été récupérés. Plus tard au cours du stage, les lectures obtenues après le séquençage ont été alignées sur le génome de la caille, ce qui a permis de voir que la fenêtre de taille effectivement sélectionnée était comprise entre 140 et 850 pb, et que par conséquent, à peu près 75 000 fragments avaient bien été obtenus en moyenne pour chaque individu.

Les données issues de la 1ère ligne de séquençage ont été disponibles le 24 décembre 2015, et un total de 351 714 110 lectures de très bonne qualité a été obtenu. Ces données ont été pré-analysées avant de valider le lancement d'une 2nde ligne. Les données de la 2nde ligne ont été obtenues le 9 février 2016. Malheureusement, le kit Illumina de préparation des

lectures était défectueux. Un total de 447 960 384 séquences a été obtenu, mais la qualité s'est révélée médiocre. Les données de cette ligne ont malgré tout été gardées dans les analyses sachant que la 1ère étape de *Stacks* permet de ne garder que les séquences de bonne qualité, et que nous avons choisi de ne pas accepter les lectures avec des erreurs de séquençage pour la détection des SNP. Une 3ème ligne a donc été lancée en remplacement de la précédente. Un total de 475 442 160 lectures de très bonne qualité a été obtenu le 23 février 2016.

Deux individus F2 se sont révélés défaillants au vu du nombre de lectures obtenues. En effet, pour l'ensemble des 3 lignes, entre 2 et 5 millions de lectures ont été obtenues par individu, alors que ces 2 échantillons n'en présentaient que quelques milliers, ce qui peut être dû à une mauvaise qualité du matériel génomique de départ.

2.2. Recherche de SNP *de novo*

2.2.1. Détection de SNP avec Stacks de novo

Pour les 192 individus et les 3 lignes réunies, un total de 1 275 116 654 lectures a été obtenu et soumis à *Stacks*. Lors de la 1ère étape de filtre qualité (`preprocessing`), 78,4% des lectures étaient de bonne qualité et ont été gardées. La 2nde étape, `ustacks`, correspond à l'assemblage des séquences identiques en piles et à l'assemblage de ces piles en loci putatifs. Cette étape se déroule pour chaque individu séparément, et les lectures secondaires ne sont pas utilisées. En moyenne, 186 394 loci par individu ont été formés (Fig. 3A). Parmi ces loci, 9,5% présentaient des SNP, à hauteur de 1,4 SNP par locus en moyenne. A l'étape `cstacks`, les loci découverts lors d'`ustacks` chez chaque individu séparément sont comparés deux à deux. Cela permet de créer une liste exhaustive et unique de tous les loci présents dans la population et obtenus par ddRADSeq. A cette étape, nous savons qu'au total 284 534 loci avec SNP sont répertoriés dans le catalogue, les 2 individus « défaillants » ayant été momentanément écartés à cette étape. L'étape `sstacks` correspond à la « soumission » de tous les individus au catalogue, afin d'identifier à quels loci du catalogue s'apparient leurs loci issus d'`ustacks`. Cette étape ne nous apprend rien ici, si ce n'est que 2 individus, déjà connus comme ayant très peu de lectures, ont un nombre très faible de correspondances au catalogue (<10 000), alors qu'en moyenne 181 661 loci par individu s'apparient au catalogue. Enfin, l'étape `populations` permet de génotyper chaque individu pour tous les marqueurs existant dans la population. Même en demandant une quantité de CPU à la limite de ce qui est possible physiquement sur le cluster de Toulouse, un problème de manque de mémoire est apparu, empêchant cette étape d'aller jusqu'au bout. J'ai donc utilisé une option proposée par

Stacks à cette étape : *–r = 0.30*, qui permet d'écarter au fur et à mesure les loci qui ne sont pas présents chez au moins 30% des individus de la population. Ce filtre permet ainsi d'exclure la plupart des loci apportant peu d'information, et donc de réduire la quantité de données. Ainsi, un total de 98 815 loci polymorphes a été trouvé.

On obtient en sortie de *Stacks* un fichier qui indique la profondeur de chaque locus polymorphe et de chaque allèle à ces loci. Cela nous a permis de calculer la profondeur moyenne sur l'ensemble des loci polymorphes obtenus. Nous sommes à 13,5X, ce qui est assez faible par rapport aux 20X conseillés (Catchen et al., 2011). Le détail des résultats obtenus au fur et à mesure du déroulement du pipeline *de novo* est récapitulé Figure 3C.

2.2.2. Sélection et localisation des SNP

Les 98 815 loci polymorphes présents à la fin de *Stacks* regroupent un total de 162 672 SNP. Seulement 564 de ces loci (0,6%) n'ont pas été alignés sur le génome de référence de la caille (CoJa2). Ils font d'ailleurs partie des loci éliminés lors du 1er filtre, qui supprime les SNP dont la qualité d'alignement est inférieure à 20. On n'obtient alors plus que 147 487 SNP. Le 2ème filtre qualité porte sur la position des SNP dans la séquence, et élimine les SNP situés aux extrémités du locus (position <5 ou >144). On descend alors à 138 488 SNP. Un 3ème filtre élimine les SNP dont la position se trouve dans une partie non alignée du locus (S), ce qui nous donne 137 075 SNP au final. Enfin, la position exacte des SNP sur les chromosomes est calculée, et le fichier plink.map est créé. Il est également possible de supprimer les duplicats. En effet, deux identifiants de SNP différents peuvent avoir exactement la même position sur le chromosome si leurs loci se chevauchent, alors qu'il s'agit en réalité du même SNP. Le nombre réel de SNP (sans les duplicats) est alors de 121 157.

2.3. Recherche de SNP avec un génome de référence

J'ai également fait tourner *Stacks* avec le génome de référence de la caille (CoJa2), étant donné que celui-ci s'est révélé disponible au cours du stage. La 1ère étape de contrôle qualité des lectures issues du séquenceur (`preprocessing`) reste la même que *de novo*, c'est-à-dire que 78,4% des lectures sont conservées. La seconde étape, `pstacks`, aligne dans un premier temps les séquences sur le génome de référence, puis liste les loci ainsi formés pour chaque individu, et enfin identifie les SNP présents sur ces loci. En moyenne, 99 469 loci par individu ont été obtenus (Fig. 3B). Parmi ces loci, 27,6% présentaient des SNP, à hauteur de 2,3 SNP par locus en moyenne. A l'étape `cstacks`, 137 951 loci avec SNP sont répertoriés

dans le catalogue. Enfin, les étapes `sstacks` et `populations` permettent d'obtenir les génotypes de chaque individu pour tous les marqueurs répertoriés dans la population. De même que lors de l'utilisation de *Stacks de novo*, l'option *–r = 0.30* a été utilisée. Ainsi, un total de 88 006 loci polymorphes pour 276 933 SNP (233 339 sans les duplicats) a été obtenu.

Sur l'ensemble de ces loci génotypés, la profondeur moyenne est de 33,1X, ce qui est satisfaisant. Le détail des résultats obtenus au fur et à mesure du déroulement du pipeline avec le génome de référence de la caille est récapitulé Figure 3D.

2.4. Analyses d'association sur R

2.4.1. Description des SNP utilisés

Les analyses d'association ont été menées sur les sorties de *Stacks de novo* et de *Stacks* avec génome de référence, d'abord avec un $\text{callrate}_{\text{SNP}}$ assez strict de 0.95, puis de 0.50 dans un second temps afin de conserver un plus grand nombre de marqueurs.

De novo, sur les 137 075 SNP localisés, 23 998 (20 896 sans les duplicats) sont gardés suite aux contrôles qualité, ainsi que 190 des 192 individus de départ. En effet, comme nous l'avons déjà précisé, 2 individus comportent très peu de séquences et ont donc d'emblée été écartés lors des contrôles qualité de GenABEL, à cause d'une trop faible proportion de génotypes couverts sur l'ensemble des marqueurs. Pour un callrate de 0.50, on conserve cette fois 75 356 SNP (66 966 sans les duplicats). Avec un génome de référence, parmi les 276 933 SNP obtenus, 44 378 (36 244 sans les duplicats) sont gardés avec un callrate de 0.95, et 184 547 (155 298 sans les duplicats) avec un callrate de 0.50, et toujours 190 individus.

La répartition des SNP obtenus pour *Stacks* avec génome de référence et pour un $\text{callrate}_{\text{SNP}}$ de 0.95 est représentée Figure 4. Concernant leur position sur le génome (Fig. 4A), on voit que les SNP sont répartis de façon assez homogène et recouvrent l'ensemble des parties assemblées de CoJa2. La distance moyenne entre marqueurs pour les chromosomes autosomaux est de 22 588 bases, avec une distance médiane de 35, et une distance maximale de 1 161 936 bases présente sur le chromosome 3. Quant à la quantité de SNP, 98,6% sont présents sur les chromosomes assemblés de CoJa2, le reste se trouvant sur des scaffolds non assignés. Mis à part les chromosomes sexuels, la quantité de SNP sur un chromosome ou un scaffold est proportionnelle à la taille de celui-ci, ce qui confirme une bonne répartition des SNP à travers tout le génome. Enfin, le nombre de SNP reste équivalent entre chromosomes avant et après les filtres qualité menés par GenABEL (Fig. 4B).

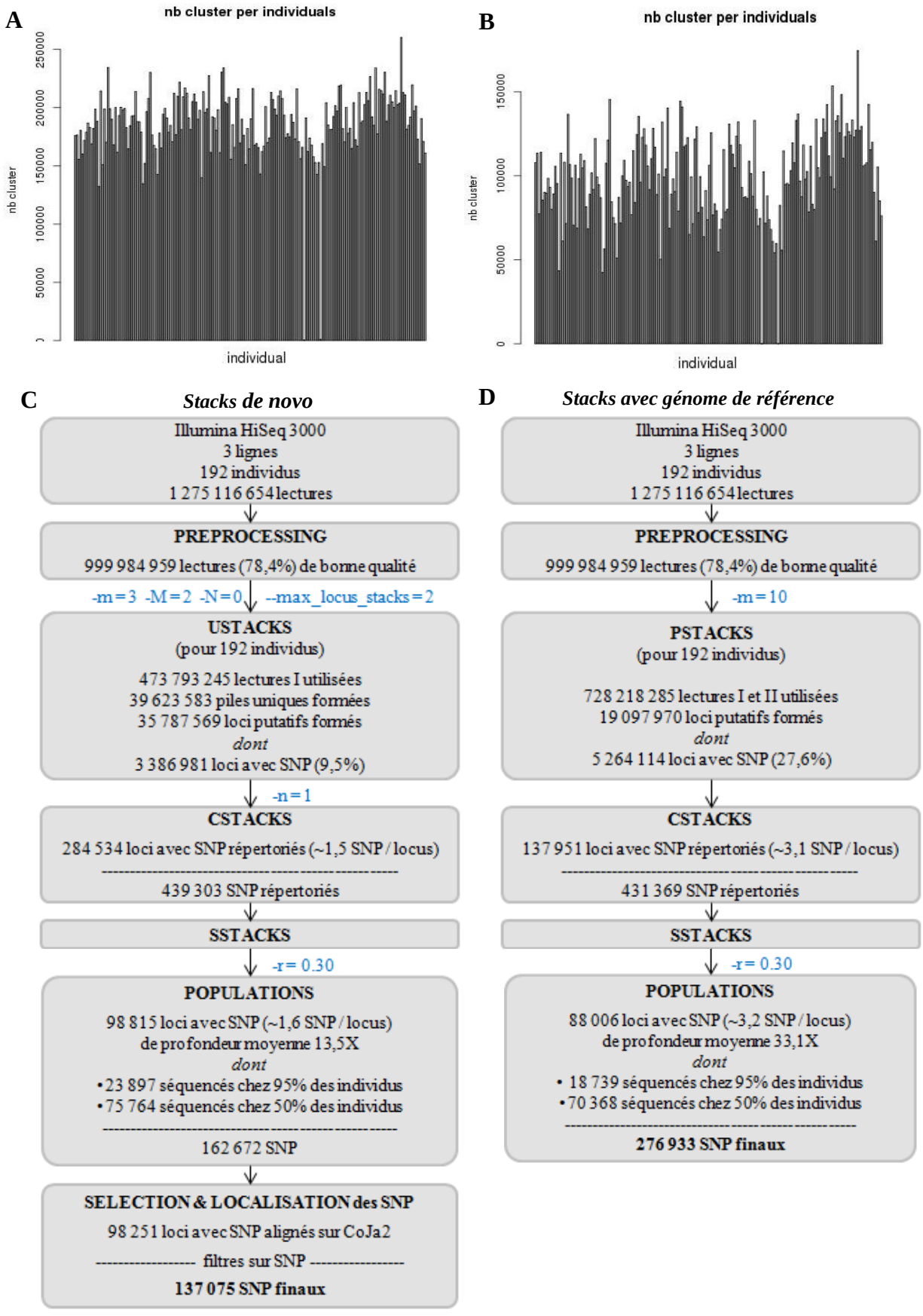
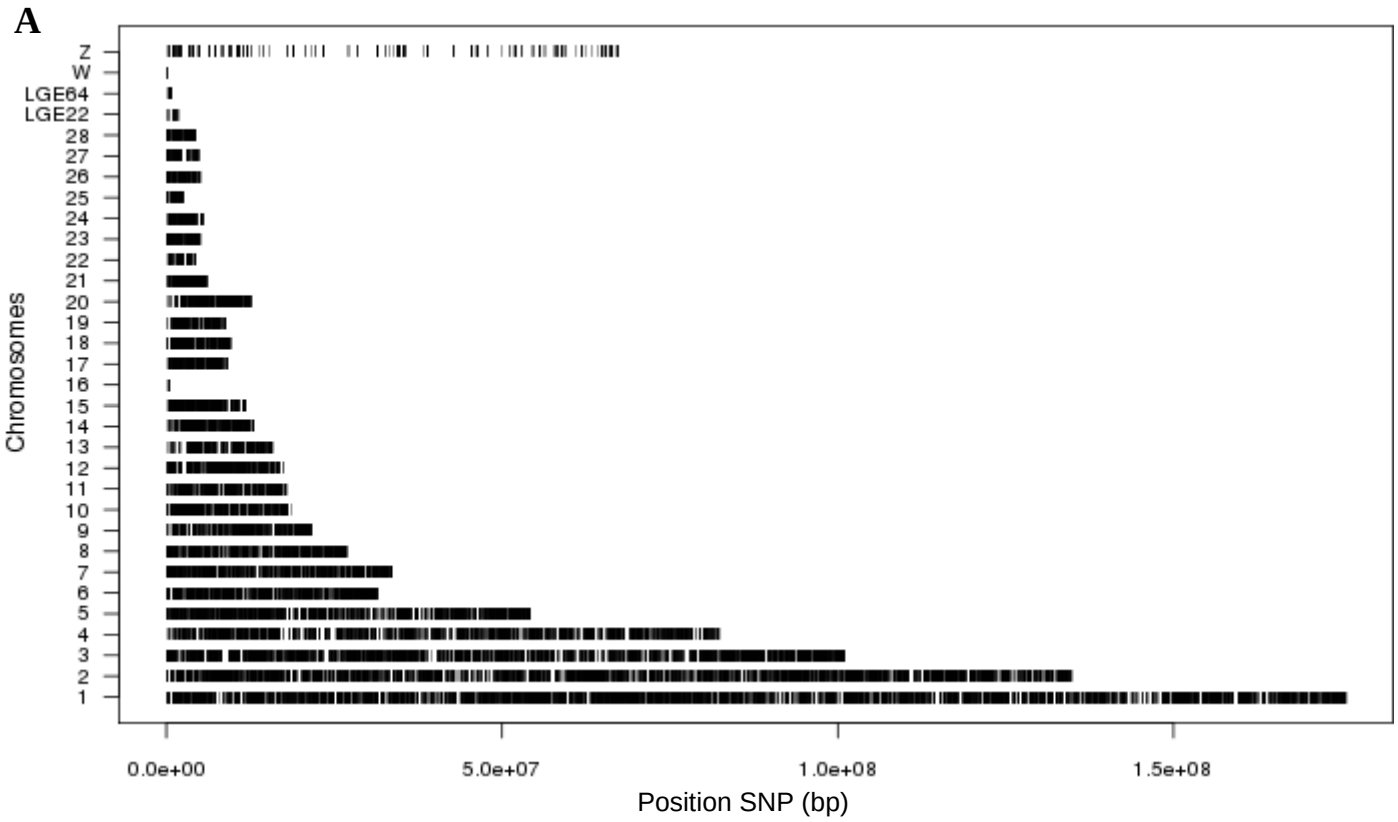


Figure 3 : Résultats récapitulatifs des données obtenues avec *Stacks*. (A) Histogramme du nombre de loci formés pour chaque individu lors de l'étape `ustacks` de *Stacks* lancé *de novo*. Les lectures secondaires ne sont pas utilisées et chaque locus est constitué de deux piles uniques maximum. (B) Histogramme du nombre de loci formés pour chaque individu lors de l'étape `pstacks` de *Stacks* lancé avec le génome de référence de la caille (CoJa2). Chaque locus se compose de toutes les lectures qui s'alignent à un même endroit sur le génome, et a une profondeur d'au moins 10X. Deux individus défaillants présentent beaucoup moins de loci, et sont d'ailleurs momentanément écartés de l'étape `cstacks` du déroulement de *Stacks* (C et D).



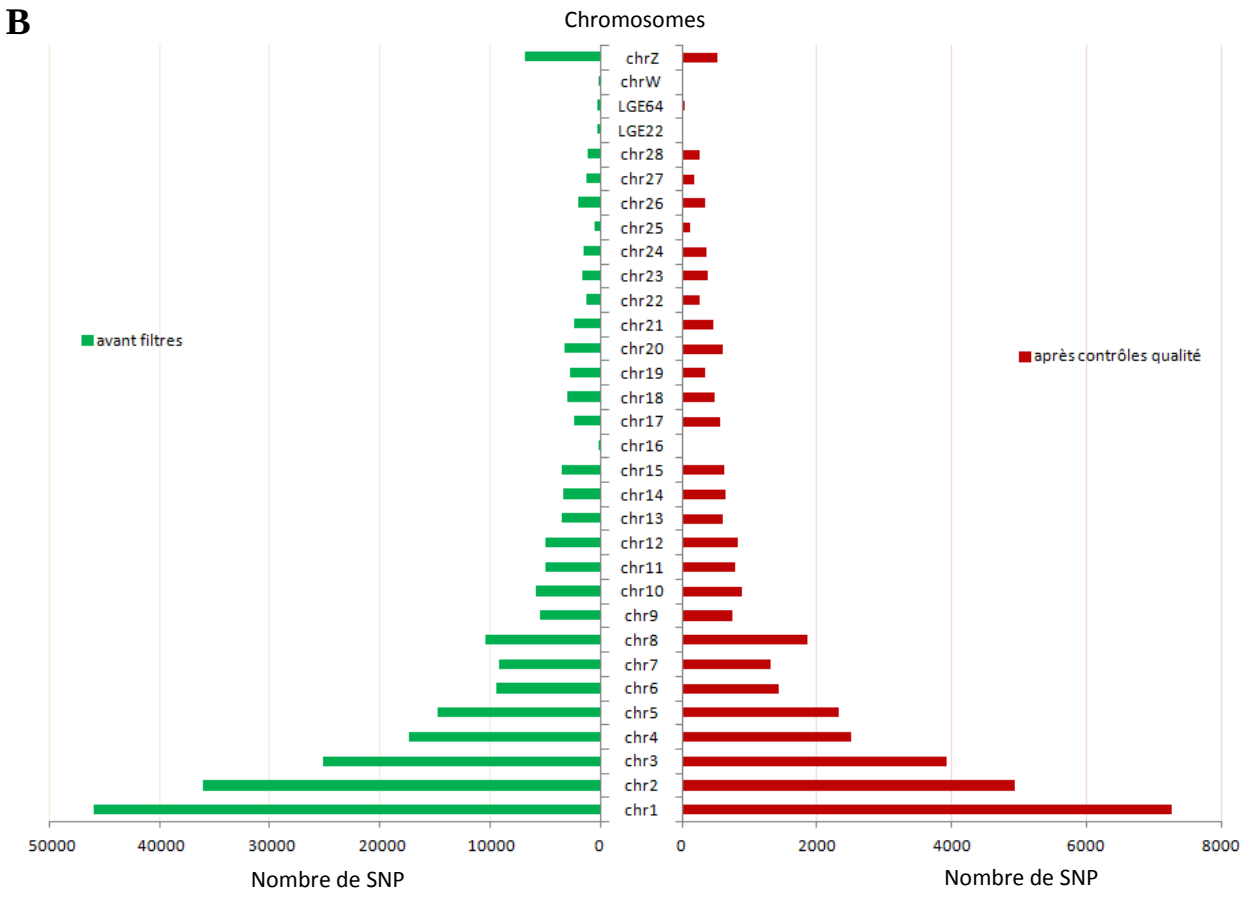


Figure 4 : Répartition sur l'assemblage de CoJa2 des SNP utilisés en GWAS pour un callrate$_{SNP}$ de 95% sur les données issues de *Stacks* avec génome de référence. (A) Répartition « géographique » des SNP (représentés par les traits verticaux) sur les chromosomes assemblés de CoJa2. Pour chaque chromosome, la dernière barre indique la longueur de celui-ci. (B) Pour chaque chromosome, nombre de SNP obtenu en sortie de *Stacks* avec génome (barres vertes) et une fois les filtres qualité de GenABEL appliqués (barres rouges).

2.4.2. Localisation des mutations causales

Avec les données issues de *Stacks de novo*, aucun pic majeur n'est apparu lors du tracé des manhattan plots pour aucun des deux caractères étudiés. Le seul résultat notable est la présence d'un unique marqueur associé au caractère « céladon » avec une p.value de $1{,}12.10^{-23}$! Ce marqueur est localisé sur le scaffold 492 (KQ966703.1), dont la place dans l'assemblage de CoJa2 reste inconnue…

L'utilisation de *Stacks* avec un génome de référence a donné de meilleurs résultats. Les manhattan plots obtenus pour chacun des caractères à un callrate de 0.95 sont présentés Figure 5. Pour le caractère « céladon », un pic majeur apparaît au niveau du chromosome 16 (CM003796.1) (Fig. 5A), où 16 SNP (assez groupés) ont une p.value inférieure à $2{,}59.10^{-6}$, jusqu'à $3{,}68.10^{-8}$ pour le meilleur d'entre eux (Tableau 1). De plus, le marqueur très associé découvert précédemment avec *Stacks de novo* sur le scaffold 492 sort également ici avec une p.value très faible de $6{,}24.10^{-23}$. Ce résultat laisse penser que ce scaffold ferait peut-être partie du chromosome 16 et correspondrait en réalité à une zone très proche de notre mutation causale sur le chromosome 16. La validation de cette hypothèse fait l'objet de la partie suivante. Quant au caractère « diabète insipide », aucun pic clair n'est apparu en GWAS (Fig. 5B). La répartition des individus sains et malades dans les 3 classes de génotypes possibles n'est pas convaincante, même pour le meilleur marqueur associé du graphique (Tableau 1). Cependant, si l'on précise dans le modèle que la mutation attendue est récessive, un pic apparaît au niveau du chromosome 17 (Fig. 6), où le meilleur SNP associé obtient une p.value de $5{,}59.10^{-7}$. La répartition des individus selon les trois génotypes possibles pour le meilleur SNP associé (Tableau 1) paraît à peu près satisfaisante, du moins dans l'hypothèse où l'on n'est pas encore très proche de la mutation causale. L'apparentement n'est cependant pas pris en compte dans un tel modèle.

Le calcul de l'inflation a d'ailleurs montré que l'apparentement était important à prendre en compte pour le caractère « diabète insipide ». La valeur de l'inflation λ correspond à la pente de la droite de régression des p.value observées en fonction des p.value attendues pour les marqueurs. Si cette pente est supérieure à 1, c'est qu'il y a un aspect important qui n'est pas pris en compte dans le modèle. Pour le caractère « céladon », on passe de $\lambda=1{,}3$ à $\lambda=1{,}0$ lorsque l'on prend en compte l'apparentement, alors que pour le caractère « diabète insipide » on passe de $\lambda=2{,}0$ à $\lambda=1{,}0$ si l'apparentement est pris en compte dans le modèle.

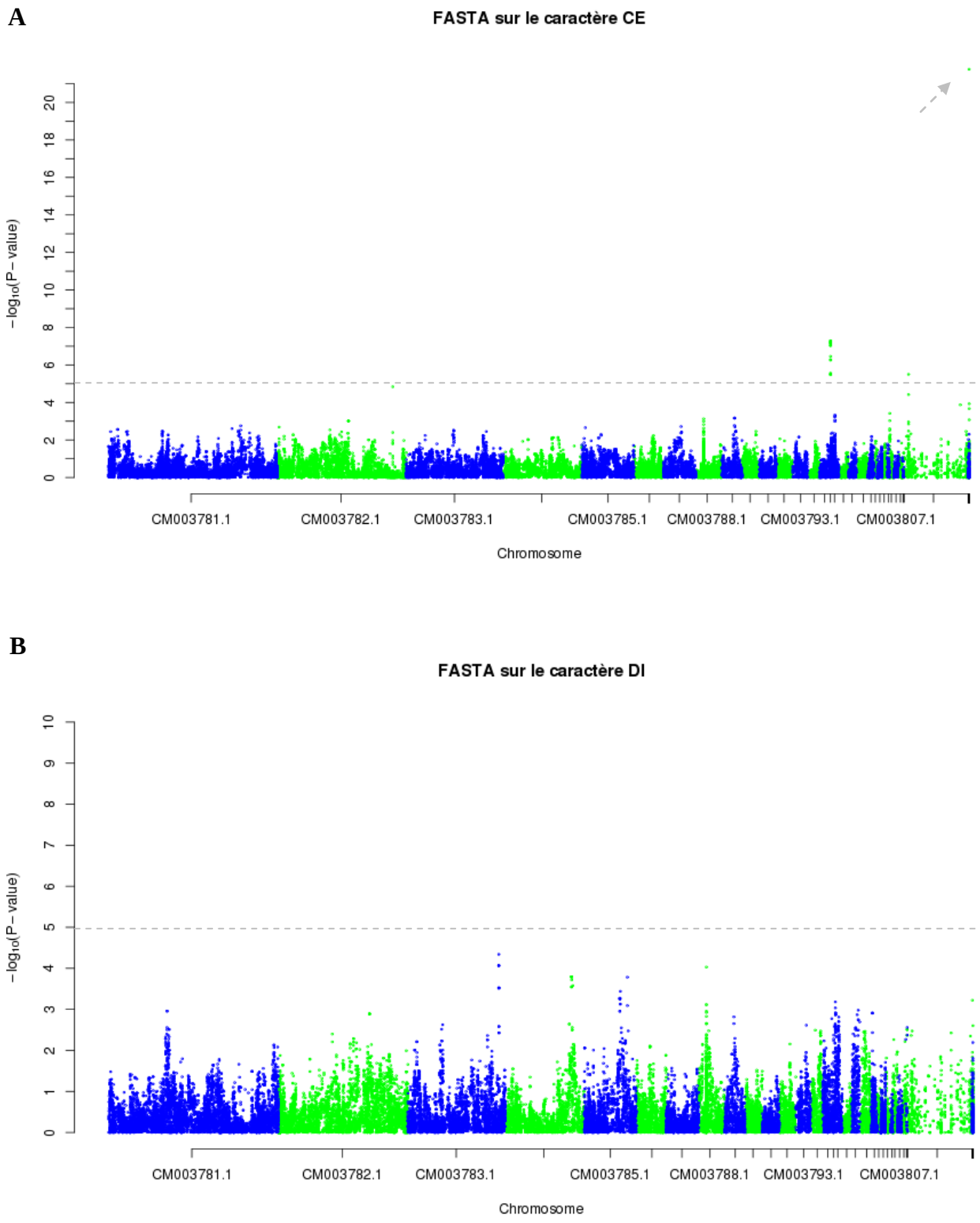


Figure 5 : **Graphes d'association des marqueurs au caractère étudié dans le cadre d'un modèle linéaire mixte.** (A) Caractère « couleur de la coquille céladon » (CE). (B) Caractère « diabète insipide » (DI). Chaque point indique le degré de significativité du test d'association d'un SNP au caractère en question. La p.value du test effectué pour chaque marqueur est calculée avec une méthode d'estimation rapide (FASTA, FAmily-based Score Test for Association) sur la base d'un modèle linéaire mixte.

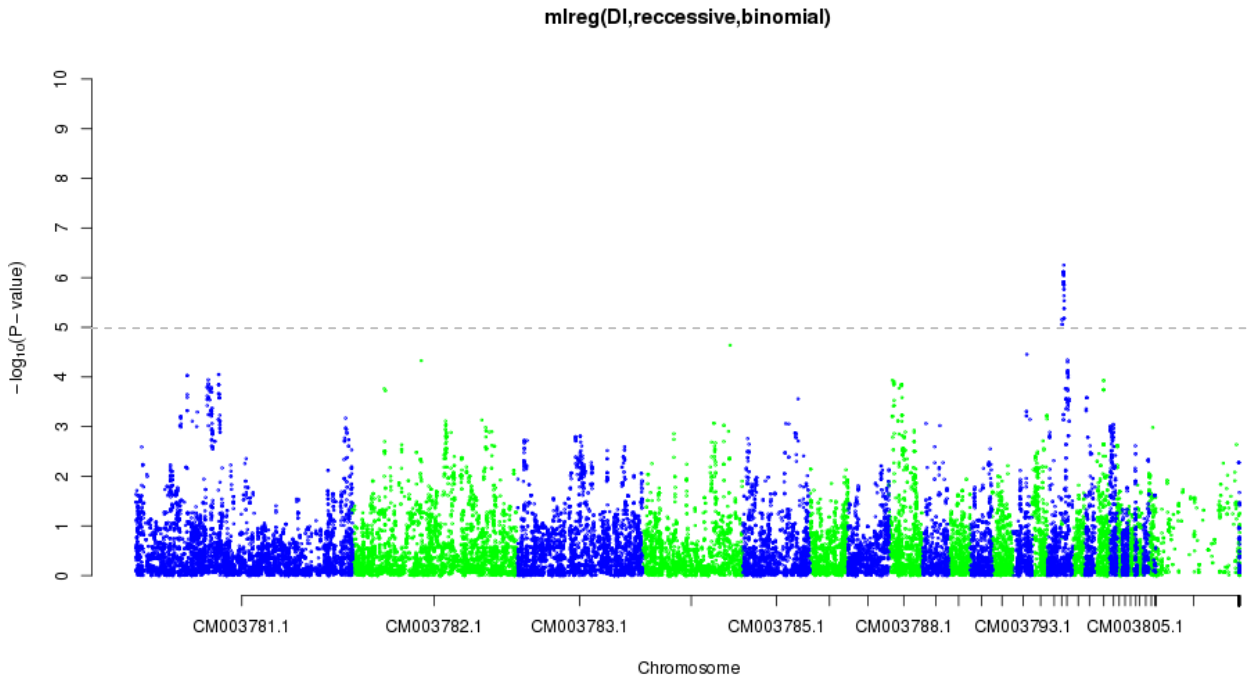


Figure 6 : **Graphe d'association des marqueurs au caractère DI (diabète insipide) dans le cadre d'un modèle linéaire récessif.** Chaque point indique le degré de significativité du test d'association d'un SNP au caractère. L'apparentement n'est pas pris en compte ici.

Tableau 1 : Tableau récapitulatif des meilleurs SNP associés trouvés lors des analyses d'association. Les tables de contingence du nombre d'individus non-porteurs (0) et porteurs (1) du caractère étudié, en fonction des 3 génotypes possibles pour chaque marqueur, sont indiquées pour les marqueurs les plus associés des graphes d'association ci-dessus (Figures 5 et 6).

					TOTAL		F0		F1		F2	
SNP	**Chromosome**	**Position**	**P.value**	**Génotype**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
CELADON	*Modèle linéaire mixte*											
107386_64	Scaffold 492	69280	6.24e-23	C/C	4	40	0	3	0	0	4	37
				T/C	89	5	0	0	23	0	66	5
				T/T	49	0	5	0	0	0	44	0
12238_41	Chr 16	296426	3.68e-08	G/G	15	24	0	4	1	0	14	20
				T/G	80	18	0	1	19	0	61	17
				T/T	43	2	5	0	2	0	36	2
DIABETE INSIPIDE	*Modèle linéaire mixte*											
70501_99	Chr 3	94260226	4.56e-05	A/A	0	2	0	0	0	0	0	2
				C/A	19	20	1	2	4	0	14	18
				C/C	123	26	4	3	20	0	99	23
DIABETE INSIPIDE	*Modèle linéaire récessif*											
12348_100	Chr 17	1873163	5.59e-07	A/A	16	23	0	5	0	0	16	18
				G/A	79	15	0	0	23	0	56	15
				G/G	45	10	5	0	0	0	40	10

2.5. Affinage de la région de la mutation « céladon »

Nous avons vu que des marqueurs présents sur le chromosome 16 et le scaffold 492 étaient très associés au caractère « céladon » (Fig. 5A). Pour un callrate$_{SNP}$ à 0.50 (résultat non montré), des SNP associés à CE apparaissent également au niveau du scaffold 1340 (KQ967262.1) avec une p.value de 1,42.10^{-9}. Les analyses de liaison menées avec le logiciel *CriMap* ont confirmé l'hypothèse que ces scaffolds, actuellement non placés dans l'assemblage du génome de référence de la caille, appartenaient certainement au chromosome 16. En effet, les SNP présents sur ces trois morceaux sont liés avec des LOD scores supérieurs à 3. Cependant, il s'est révélé très difficile de réaliser une carte génétique pour réassembler le chromosome 16 et y intégrer les deux scaffolds. Malgré tout, une carte a été choisie en concordance, d'une part avec les valeurs de liaison des SNP deux à deux (en incluant également un SNP fictif qui aurait parfaitement correspondu à la mutation *ce*), et d'autre part avec la quantité de recombinaisons obtenues et la probabilité de validité globale de la carte. Seuls 8 marqueurs sur les 10 sélectionnés (8 sur le chromosome 16 et 2 sur les scaffolds) ont pu être placés sur la carte (Fig. 7).

Dans un second temps, l'étude des génotypes des descendants présentant des haplotypes maternels ou paternels recombinants, à l'aide du logiciel *Yapp*, a permis de réduire la région de la mutation « céladon » à une zone qui se situerait entre les deux scaffolds, soit une distance d'environ 18 cM entre les marqueurs flanquant. Pour cela, l'ensemble des SNP présents sur les 8 loci précédemment placés sur la carte on été utilisés, soit un total de 46 SNP (42 sur le chromosome 16, 3 sur le scaffold 1340, et 1 sur le scaffold 492).

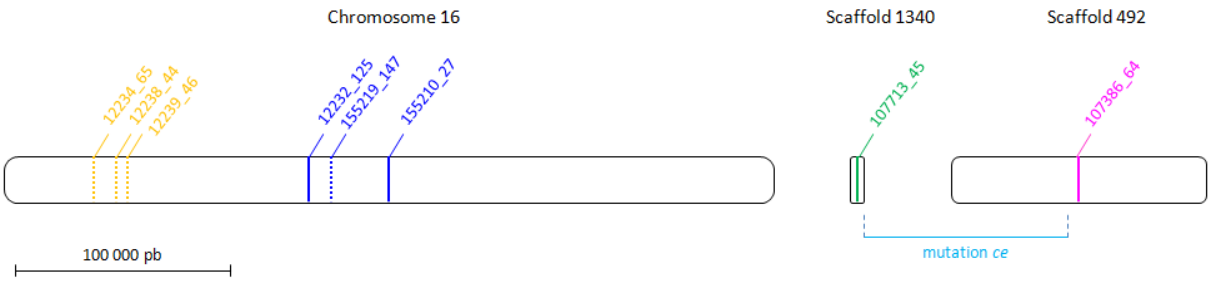


Figure 7 : Ordonnancement de la partie assemblée du chromosome 16 avec les scaffolds 1340 et 492, encore non placés dans l'assemblage actuel du génome de référence de la caille CoJa2. Les SNP dont le positionnement est indiqué en pointillé correspondent à des loci dont la position a été modifiée par rapport à leur alignement sur l'assemblage actuel du chromosome 16. Ces positions restent cependant approximatives et la carte génétique montrée ici représente le meilleur compromis que nous avons obtenu suite aux analyses de liaison entres les marqueurs à notre disposition.

III. Discussion

L'étude menée ici a permis de localiser une région d'intérêt pour le caractère « céladon », ce qui confirme que l'utilisation de la technique de RADSeq peut pallier le manque de puce SNP pour des espèces peu étudiées comme la caille. Un grand nombre de marqueurs et une bonne couverture de l'ensemble du génome ont été obtenus, même si quelques améliorations restent néanmoins possibles.

3.1. Avantages et limites du ddRADSeq

Le ddRADSeq est une technique de génotypage par séquençage mise au point par Peterson et al. (2012), qui cible à travers tout le génome des sites reconnus par deux enzymes de restriction choisies, et séquence les courtes régions adjacentes à ces sites, afin d'accéder à une partie réduite des SNP présents dans le génome. Leur protocole diffère de celui de RADSeq classique (Baird et al., 2008) sur trois points: deux enzymes au lieu d'une sont utilisées pour l'étape de digestion, il n'y a pas d'étape de redécoupage aléatoire des fragments, et l'étape de sélection des fragments selon leur taille est plus précise. L'ajout d'une seconde endonucléase et une sélection plus précise de la taille des fragments permettent une plus grande flexibilité du nombre de fragments obtenus dans la librairie finale, et donc un meilleur contrôle de la fraction de génome représentée (Peterson et al., 2012). En effet, le choix d'enzymes coupant plus ou moins fréquemment et le choix de la taille des fragments à conserver peuvent permettre de réduire le nombre de régions génomiques couvertes pour chaque individu. Ainsi, non seulement le coût de séquençage par individu est réduit par rapport à un RADSeq classique, mais plus d'individus peuvent également être génotypés sur une seule ligne de séquençage Illumina (pour une couverture moyenne identique). De plus, l'élimination d'une étape du protocole permet de diminuer les coûts (temps et argent) de préparation de la librairie, en plus de pouvoir utiliser une plus faible quantité d'ADN de départ.

Cependant, ces méthodes introduisent différents biais qui peuvent influencer de manière significative les résultats. Tout d'abord, on introduit un biais dans la représentation du génome obtenue, étant donné que les enzymes de restriction coupent à des endroits spécifiques du génome et que les fragments obtenus sont sélectionnés selon une fenêtre de taille précise. On n'obtient donc pas une représentation aléatoire du génome, mais cela permet de cibler les mêmes régions chez tous les individus, et donc d'optimiser la détection des SNP (Peterson et al., 2012). Deuxièmement, l'amplification PCR est fortement biaisée en faveur des petits fragments, qui seront donc mieux représentés (plus forte profondeur) (Da Costa et

al., 2014). Pour pallier ce biais, Da Costa et al. (2014) n'ont extrait que la moitié de la bande du gel d'agarose pour les petits fragments lors de l'étape de la sélection sur la taille. Peterson et al. (2012) ont eux insisté sur le fait que l'étape de sélection des fragments selon leur taille devait être très précise, afin de limiter le biais en faveur des petits fragments mais aussi d'obtenir une profondeur homogène entre toutes les régions couvertes. Malheureusement, la technique de Blue Pippin, comme recommandée par Peterson et al., n'était pas au point à la plateforme de génomique de l'INRA de Toulouse lors des expériences. Tout comme Da Costa et al., la taille des fragments obtenus dans la librairie finale s'est révélée assez variable par rapport à la fenêtre de taille initialement désirée. Cela pose surtout un problème pour les fragments dont la taille est inférieure au nombre de bases lues par le séquenceur (150 pb), car l'adaptateur lié au fragment est alors également en partie séquencé. Comme Da Costa et al. lors du traitement initial de leurs données, j'ai dû rogner la séquence correspondant à de l'adaptateur dans les lectures obtenues avant de pouvoir les aligner sur le génome de la caille. De plus, si l'étape de sélection sur la taille des fragments échoue, la représentation des loci ne sera pas la même entre les différents individus. Nous avons d'ailleurs obtenu lors du déroulement de *Stacks de novo* un fort bruit de fond correspondant à des loci présents chez seulement un individu (108 126 loci sur les 284 534 répertoriés dans le catalogue ne sont pas présents chez plus d'un individu). Enfin, l'homogénéité de la représentation des loci entre individus peut également être biaisée à cause d'une coupure à un endroit non spécifiquement reconnu par une enzyme de restriction, mais ce cas est rare (Da Costa et al., 2014).

En comparaison des approches de RADSeq déjà existantes, le ddRADSeq permet une plus grande flexibilité et une plus grande robustesse concernant la couverture du génome obtenue, mais aussi réduit le coût, la quantité de matériel génomique requise, et le temps d'investissement pour la construction des librairies. La vraie limite de cette technique était jusqu'à récemment l'existence d'outils informatiques pour traiter ces données.

3.2. Avantages et limites du logiciel de détection de SNP Stacks

Stacks a été créé pour pallier ce manque de logiciels capables de traiter un grand nombre de données RADSeq. Il permet de construire et de génotyper des loci chez un ensemble d'individus, soit *de novo*, soit par comparaison à un génome de référence.

3.2.1. Choix des paramètres

Stacks, lorsqu'il est lancé en *de novo*, tourne autour d'un ensemble de paramètres clefs

dont les valeurs influencent fortement la construction des loci putatifs. Les valeurs optimales pour ces paramètres dépendront à la fois du polymorphisme du génome analysé, du taux d'erreurs de séquençage et de la profondeur atteinte (Catchen et al., 2011). J'ai donc d'abord testé différentes valeurs de paramètres sur les données issues de la 1ère ligne de séquençage à ma disposition.

Le paramètre ***m*** correspond au nombre minimal de lectures identiques pour créer une pile unique. Si sa valeur est trop faible, les lectures avec des erreurs de séquençage seront retenues dans des piles. Si sa valeur est trop élevée, certains « vrais » allèles ne seront pas retenus. Ce paramètre peut donc permettre de distinguer les vrais allèles d'erreurs de PCR ou de séquençage, mais à l'inverse augmenter les erreurs en considérant comme homozygote un locus hétérozygote (Mastretta-Yanes et al., 2015). Le paramètre ***M*** détermine le nombre de mésappariements autorisés entre loci (pour un individu). Si sa valeur est trop faible, des loci allant ensemble ne seront pas regroupés, et des SNP seront « perdus ». Si sa valeur est trop forte, des loci issus en réalité de régions différentes (séquences répétées ou paralogues) seront joints, et de faux SNP seront créés (Mastretta-Yanes et al., 2015). J'ai donc gardé les valeurs par défaut de ces deux paramètres pour l'étape `ustacks`. Cependant, j'ai choisi d'autres valeurs pour les deux autres paramètres disponibles à cette étape. Le paramètre ***max_locus_stacks***, qui correspond au nombre maximal de piles dans un locus putatif, permet d'éviter les cas de confusion où des centaines de piles se regrouperaient au sein d'un même locus putatif alors qu'il s'agit d'erreurs de séquençage ou de séquences répétées (Mastretta-Yanes et al., 2015). J'ai choisi une valeur de 2 au lieu de sa valeur par défaut à 3 (allèles portés par des individus diploïdes, plus une autre pile avec une erreur de séquençage), car cette dernière amenait à de nombreux loci avec plus de 2 haplotypes, ce qui ne convenait pas pour des individus diploïdes. De plus, nous avons remarqué que l'utilisation de lectures secondaires, qui ont pour avantage d'augmenter la profondeur des loci afin de confirmer des SNP potentiels, a également comme désavantage de créer de faux SNP si les lectures secondaires sont trop nombreuses. C'est pourquoi je n'ai pas autorisé les lectures secondaires, en choisissant la valeur 0 pour le paramètre ***N***, qui correspond au nombre d'erreurs de séquençage autorisées au sein d'un locus putatif. En prenant *N=0* et *max_locus_stacks=2* au lieu des valeurs par défaut, la profondeur des loci polymorphes et le nombre de SNP détectés sont réduits, mais nous obtenons plus de 150 000 SNP génotypés chez 50% de la population, ce qui semblait largement suffisant pour réaliser ensuite des analyses d'association, c'est pourquoi j'ai été assez stricte sur ces 2 derniers paramètres en *de novo*.

Stacks lancé avec un génome de référence accepte quant à lui comme seul paramètre le nombre minimal de lectures alignées au même endroit pour qu'un locus putatif soit créé

(lectures secondaires comprises). Ce paramètre, également appelé ***m***, permet de décider de la profondeur minimale des loci putatifs, ce qui n'est pas possible en *de novo*. En effet, le paramètre *m* en *de novo* a pour principal objectif d'écarter les lectures avec erreurs de séquençage de l'analyse, en imposant une profondeur minimale pour former une pile unique. Cela revient à demander une profondeur minimale de 2*m* pour les loci polymorphes (constitués de deux piles), mais de seulement *m* pour les loci monomorphes (constitués d'une seule pile). Ainsi, si l'on veut une profondeur d'au moins 20X pour nos loci, on peut choisir *m*=10, qui permet d'atteindre au moins 20X pour les loci polymorphes, mais autorise des loci monomorphes à seulement 10X, ce qui n'est pas suffisant pour décider qu'un individu est homozygote pour ce locus. Mais en prenant *m*=20 pour atteindre 20X pour des loci monomorphes, on risque de perdre des loci polymorphes en étant trop exigeant sur la profondeur d'un seul allèle. Je pense qu'il serait donc bénéfique d'ajouter un paramètre au pipeline *de novo*, qui permettrait, indépendamment de la profondeur des piles uniques, de décider de la profondeur minimale des loci putatifs, afin de ne garder que des loci fiables.

Catchen et al. (2012) ont d'ailleurs montré que *Stacks* identifiait très bien les loci lorsque la profondeur de séquençage se trouve entre 20 et 40X, mais qu'une profondeur de 10X était insuffisante, excepté pour un taux d'erreur de séquençage très faible (0,5%). Pour des études avec des profondeurs faibles, Da Costa et al. (2014) conseillent eux d'utiliser des réplicats dans la préparation des librairies, afin d'éviter au mieux les erreurs de génotypage.

3.2.2. Comparaison des résultats de novo *et avec génome de référence*

Avec les paramètres que nous avons choisis, *Stacks* lancé en *de novo* donne presque deux fois plus de loci putatifs que s'il est lancé avec le génome de référence de la caille (`ustacks` Fig. 3). Cela peut s'expliquer d'une part par le fait que l'on a décidé d'une profondeur minimale de 10X pour les loci alignés sur CoJa2, alors que l'on a pris seulement 3X pour la profondeur d'une pile unique, ce qui revient à 6X au minimum pour un locus polymorphe et 3X pour un locus monomorphe. Ainsi, de nombreux loci ont dû être écartés lors du lancement de *Stacks* avec CoJa2. D'autre part, nous n'avons autorisé que 2 mésappariements pour former un locus putatif en *de novo*, alors que *Stacks* ne tient pas compte de cela avec un génome. Des piles uniques allant certainement ensemble ont donc dû être séparées en deux loci putatifs au lieu d'un en *de novo*, si elles différaient de plus de 2 SNP. En effet, la quantité moyenne de SNP par locus putatif avec CoJa2 (3,1) est le double de celle en *de novo* (1,5) (`cstacks`). Le nombre de loci polymorphes est néanmoins assez équivalent à la fin du pipeline, avec 12% de loci polymorphes de plus en *de novo* qu'avec un

génome (`populations`). La profondeur est cependant très différente, avec une moyenne de 13,5X, pour une médiane à 10,6X, en *de novo*, et une moyenne à 33,1X, pour une médiane à 20,0X avec génome. La profondeur plus faible en *de novo* peut s'expliquer d'une part par le fait que l'on n'a pas autorisé l'utilisation des lectures secondaires, contrairement à *Stacks* avec génome où toutes les lectures alignées sont gardées. D'autre part, la profondeur minimale autorisée pour les loci putatifs est dès le départ plus élevée lorsque l'on a lancé *Stacks* avec CoJa2, ce qui fait augmenter la profondeur moyenne finale. Enfin, même avec un peu moins de loci, la quantité de SNP finale est bien plus élevée (x1,7) lorsque *Stacks* utilise un génome de référence que *de novo*. En effet, plus de SNP peuvent être détectés du fait que plus de piles sont regroupées au sein d'un même locus putatif, que les lectures secondaires sont utilisées, mais aussi que les lectures sont comparées au génome de l'individu de référence.

Dans tous les cas, il n'aurait certainement jamais été possible de trouver un pic pour le caractère « céladon » en faisant tourner *Stacks* en *de novo*, car les SNP apparaissant au niveau du chromosome 16 en GWAS (Fig. 5A) appartiennent à des loci possédant chacun entre 6 et 10 SNP. Pour retenir de tels loci en *de novo*, il aurait été nécessaire de choisir une grande valeur pour ***M*** (nombre de mésappariements autorisés entre deux piles uniques pour former un locus), ce qui aurait créé de nombreux loci putatifs faux car issus du groupement de piles ne correspondant pas en réalité à deux allèles d'un même locus. Le *de novo* n'est donc pas très adapté pour détecter des SNP dans des régions très polymorphes.

Il existe aujourd'hui plusieurs logiciels permettant de traiter des données RADSeq, mais *Stacks* est celui qui a été le plus largement adopté du fait de sa faible demande en mémoire et de son efficacité à détecter un grand nombre de loci (Sovic et al., 2015).

3.3. Le caractère « couleur de la coquille céladon »

Le précédent projet visant à localiser la mutation « céladon » (Leroux et al., 2013) n'avait obtenu que deux SNP associés à *ce* avec une faible significativité et localisés sur l'homologue chez la caille du chromosome 16 de la poule. L'étude menée ici montre que la mutation causale de ce caractère se situe effectivement sur le chromosome 16. Malheureusement, l'étude de ce chromosome chez les oiseaux reste difficile, car c'est un micro-chromosome riche en C+G et en séquences répétées, dont l'assemblage est encore aujourd'hui incomplet. Chez la poule, la taille de ce chromosome est estimée à environ 10Mb, mais le dernier assemblage en date (Galgal5) ne présente que 1Mb de séquence assignée à ce chromosome, dont seulement 650Kb sont assemblés (NCBI). Nous avons néanmoins

progressé pour ce qui est de la caille japonaise, dans le sens où nous savons désormais que les scaffolds 294 et 1340 font certainement partie de ce chromosome 16. De plus, il est clair que la région de la mutation « céladon » se limite certainement à la zone située entre ces deux scaffolds, bien que cela représente une distance assez grande de 18 cM, pour laquelle aucune séquence n'est actuellement disponible. Aucun gène candidat évident pour ce caractère n'a d'ailleurs été trouvé dans l'assemblage déjà disponible. La prochaine étape pour identifier la mutation causale serait donc de séquencer des individus pour cette région, mais aucun assemblage complet de ce chromosome n'est encore disponible. Je n'ai d'ailleurs moi-même pas obtenu en RADSeq beaucoup de séquences qui ne s'alignaient pas sur l'assemblage de référence CoJa2 (564 loci non alignés, soit 0,6% des 98 815 loci polymorphes obtenus en *de novo*), et qui correspondent à des zones encore non séquencées.

3.4. Le caractère « diabète insipide »

Sur la base d'une hypothèse de mutation autosomale récessive (Minvielle et al., 2007), l'étude menée ici avait également pour but de localiser la mutation causale du diabète insipide observé dans les élevages de cailles. Malheureusement, les analyses d'association n'ont révélé aucun pic majeur associé à ce caractère (Fig. 5B). Plusieurs hypothèses peuvent tenter d'expliquer cela. Tout d'abord, il est possible que la région d'intérêt n'ait pas été « accessible », soit parce qu'elle ne contenait pas de sites de coupure spécifiques des enzymes de restriction choisies, soit par manque de SNP couvrant cette zone. J'ai cependant obtenu pour cette étude un grand nombre de marqueurs, qui couvraient de façon assez complète et homogène l'ensemble du génome (Fig. 4A), du moins pour les chromosomes autosomaux. Une seconde explication possible serait l'existence d'erreurs de phénotypage, faussant ainsi nos analyses statistiques. Cependant, les animaux choisis étaient ceux dont le phénotype DI était certain. Ces animaux ont en effet été phénotypés comme porteurs de la maladie au moins trois fois dans leur vie, et notre protocole ne contient que des descendants femelles, le caractère étant plus marqué chez elles que chez les individus mâles (Minvielle et al., 2007). Il aurait également été possible d'avoir de la pénétrance incomplète (individus porteurs de la mutation mais non malades), mais cela aurait été visible par un écart par rapport aux proportions mendéliennes attendues pour ce caractère dans la population, ce qui n'était pas le cas. Enfin, peut-être qu'il ne s'agit en réalité simplement pas d'un gène majeur, comme il a été supposé par les analyses de ségrégation précédemment menées (Minvielle et al., 2007), mais de plusieurs gènes en interaction. Dans ce cas, il serait intéressant de refaire des analyses de ségrégation sur un plus grand nombre d'individus.

Conclusion

Cette étude, lors de laquelle la méthode de ddRADSeq a été utilisée pour la première fois à l'INRA de Toulouse, montre l'importance des méthodes de génotypage par séquençage pour accéder, chez les espèces encore peu étudiées, à des données semblables à celles obtenues grâce aux puces SNP commerciales. En effet, nous avons ainsi pu localiser la région de la mutation responsable du caractère de dépigmentation des œufs chez la caille, ce qui permettra certainement, lorsqu'un assemblage plus complet du génome de *Coturnix japonica* sera disponible, de mieux comprendre le caractère complexe de camouflage des œufs chez les oiseaux sauvages. Aucune région n'a en revanche été détectée pour le caractère diabète insipide. Il faudra certainement réaliser d'autres analyses, prenant par exemple en compte l'hypothèse d'un déterminisme génétique plus complexe.

Références bibliographiques

1. Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL : an R library for genome-wide association analysis. *Bioinformatics* 23 : 1294–1296.

2. Babey, M., Kopp, P., and Robertson, G.L. (2011). Familial forms of diabetes insipidus : clinical and molecular characteristics. *Nature Reviews Endocrinology* 7 : 701–714.

3. Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3 : e3376.

4. Bichet, D.G. (2009). V2R Mutations and Nephrogenic Diabetes Insipidus. *Progress in Molecular Biology and Translational Science* 89 : 15–29.

5. Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H., and De Koning, D.J. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3 *Genes|Genomes|Genetics* 1 : 171–182.

6. Da Costa, JM., Sorenson, MD. (2014). Amplification Biases and Consistent Recovery of Loci in a Double-Digest RAD-seq Protocol. *PLoS ONE* 9 : e106713.

7. Green, P., Falls, K., Crooks, S. (1990). CRIMAP Documentation. Available at: http://saf.bio.caltech.edu/saf_manuals/crimap-doc.html. (Accessed: 22nd May 2016)

8. Ito, S., Tsudzuki, M., Komori, M., Mizutani, M. (1993). Celadon: An Eggshell Color Mutation in Japanese Quail. *The Journal of Heredity* 84 : 145-147.

9. Leroux, S., Dehais, P., Vignal, A., Bouchez, O., Faraut, T., Rossignol, MN., Arnould, C., Leterrier, C., Pitel, F., Minvielle, F. (2013). Détection de SNP chez la caille : production d'un outil commun pour des programmes scientifiques multiples. *Dixièmes Journées de la Recherche Avicole et Palmipèdes à Foie Gras*, La Rochelle, France.

10. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 : 1754–1760.

11. Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D., and Emerson, B.C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources* 15 : 28–41.

12. Minvielle, F., Grossmann, R., and Gourichon, D. (2007). Development and Performances of a Japanese Quail Line Homozygous for the Diabetes Insipidus (di) Mutation. *Poultry Science* 86 : 249–254.

13. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq : An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* 7 : e37135.

14. Sovic, M.G., Fries, A.C., and Gibbs, H.L. (2015). AftrRAD : a pipeline for accurate and efficient *de novo* assembly of RADseq data. *Molecular Ecology Resources* 15 : 1163–1171.

15. Yang, Y., Minvielle, F., Kuykindoll, RJ., Gasc, JM., Yamamoto, T., Nishimura, H. (2008). "Diabetes Insipidus" strain quail show poorly developed medullary cones and low aquaporin 2 water channel expression. *Experimental Biology*, San Diego, Californie. FASEB J. 2008 22:934.24.

	Diplôme : Ingénieur Agronome et Master 2 Recherche Spécialité : SCMV (Sciences Cellulaire et Moléculaire du Vivant) Spécialisation / option : Génétique Enseignant référent : Frédéric LECERF

Auteur(s) : Anna MARISSAL Date de naissance* : 13/02/1995	Organisme d'accueil : INRA Adresse : 24 Chemin de Borde Rouge 32 326 CASTANET-TOLOSAN Maître de stage : Frédérique PITEL
Nb pages : 30 Annexe(s) : /	
Année de soutenance : 2016	

Titre français : Utilisation d'une approche de génotypage RADSeq pour le clonage positionnel de deux gènes à effet majeur chez la caille, responsables des phénotypes « diabète insipide » et « couleur de la coquille céladon ».

Titre anglais : Use of a RADSeq genotyping approach for positional cloning of two major effect genes in quail, responsible for « diabetes insipidus » and « shell color celadon » phenotypes.

Résumé : L'étude menée ici vise à localiser deux gènes majeurs chez la caille japonaise *Coturnix japonica*, responsables des caractères « diabète insipide » et « couleur de la coquille céladon ». Pour cela, 158 femelles F2 présentant les phénotypes mutants et sauvages pour les deux caractères ont été obtenues par croisement de lignées homozygotes pour l'une ou l'autre des mutations. Les nouveaux outils de génotypage à haut débit, comme les puces SNP, ont largement permis d'améliorer les capacités d'analyse des génomes d'espèces modèles, mais restent encore indisponibles pour des espèces agronomiques de plus faible intérêt comme la caille. Une technique récente de génotypage par séquençage, le ddRADSeq, a donc été utilisée, et une détection de SNP a ensuite été réalisée à l'aide du logiciel *Stacks*, d'abord *de novo*, puis avec le génome de référence de la caille, publié au moment du stage. Plus de 230 000 SNP ont été repérés, et environ 36 000 ont été utilisés lors d'analyses d'association, qui ont permis de localiser la région de la mutation « céladon » au niveau du chromosome 16, ainsi qu'une région potentielle pour le caractère « diabète insipide ». Enfin, des analyses de liaison ont également permis d'insérer deux scaffolds non assignés dans l'assemblage du chromosome 16.

Abstract : The goal of the present study is to locate two major genes in Japanese quail *Coturnix japonica*, responsible for the traits "diabetes insipidus" and "celadon color shell". For this purpose, 158 F2 females holding the wild and mutant phenotypes for the two traits have been created by crossing two homozygous lines for one or the other mutation. The new tools for high-throughput genotyping, such as SNP microarrays, have broadly improved the capacities to analyze the genome of model species, but are not yet available for agronomic species of lower interest such as quail. Thus, a genotyping by sequencing technique, the ddRADSeq, has been used, and a detection of SNP has been then realized with the software *Stacks*, first *de novo*, then with the reference genome of the quail, which has been published during the internship. More than 230,000 SNP have been spotted, and about 36,000 of them have been used in genome-wide association studies. This allowed us to locate the region of the "celadon" mutation on chromosome 16, as well as a potential region for the trait "diabetes insipidus". Linkage analyses also allowed us to insert two unplaced scaffolds in the assembly of the chromosome 16.

Mots-clés : céladon – diabète insipide – gène majeur – SNP – RADSeq – GWAS – analyse de liaison

Key Words: celadon – diabetes insipidus – major gene –SNP – RADSeq – GWAS – linkage analysis

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