Potentiels du séquençage des ARN pour explorer les micro-variations du génome

Fabien Degalez

AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2019-2020 Spécialité : Biologie moléculaire et cellulaire (BMC) Spécialisation (et option éventuelle) :	**Mémoire de fin d'études** X d'ingénieur de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ de master de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ d'un autre établissement (étudiant arrivé en M2)

Potentiels du séquençage des ARN pour explorer les micro-variations du génome

Par : Fabien DEGALEZ

*Soutenu à **Rennes** le **17 juin 2020***

Devant le jury composé de :

- Frédéric Lecerf — Maîtres de stage : Sandrine Lagarrigue et Frédéric JEHL
- Emmanuel Giudice — Enseignant référent : Frédéric Lecerf

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

ABSTRACT

RNA-seq (RNA sequencing) is now mainly used to quantify transcriptome expression (all transcripts expressed in a sample), as a replacement for gene chips. For example, it can be used to study genetic regulations, processes involved in diseases, responses to environmental variations or to understand differences between lineages with contrasting phenotypes. It also permits the modelling of genes and transcripts, the analysis of allele-specific expression or the study of RNA editing. Indeed, thanks to its sequencing step, the RNA-seq provides access to the sequence and thus allows the detection of variants in the transcribed regions of the genome. This resource, which is still under-exploited for detecting variants in the genome, is interesting for three reasons : (*i*) the access to SNP of coding regions, opening the way to the study of their effects on protein function and eventually character or disease phenotypes, thus offering an alternative to "exome" technologies that are less or not developed in non-model species such as farmed species ; (*ii*) assuming a sufficient number of SNP are available, markers can be used for studies of diversity and genotype-phenotype association, particularly in poorly studied populations, thus offering an alternative to genotyping chips limited to the major populations of the species and whose development cost is considerable ; (*iii*) the accumulation over the years of these RNA-seq data on a relatively large number of individuals compared to the DNA-seq, and belonging to populations that may be very varied, thus offering an original resource of data that is already available and as yet unexploited.

In this context, the general objective of the work undertaken during this internship is the detection of variants from RNA-seq data in hens. The host team has more than 700 RNA-seq data from 10 commercial or experimental populations of laying hens or broilers.

The first objective (of the internship) was to finalize a pipeline for reliable SNP detection from RNA-seq data, using the reference tools STAR for alignment and GATK for SNP detection. A filter resulting from the "good practices" proposed by the GATK developers was commented. Using a population of 15 laying hens for which DNA and RNA sequencing data were available for the same tissue, we showed that RNA-seq was able to detect more than 85% of the SNP captured in DNA-seq and even more if several tissues were studied ; the match of SNP between RNA-seq and DNA-seq data exceeded 90%, thus confirming the reliability of these SNP. We replicated this work on another independent population of 8 broilers and obtained similar results. We also showed that an increasing number of tissues can significantly increase the number of SNP, which is a good indication because transcriptome studies are increasingly performed on several tissues. This pipeline used on the 10 hen populations allowed the extraction of ~9.9M SNP in total with an average of 1.8M SNP per population. By then focusing

only on SNP with well-informed genotypes, it was possible to extract an average of ~0.5M SNP per population and 250,000 SNP common to all 10 populations. This set reveals a large number of SNP compatible with genotype-phenotype association or population genetics studies. We have therefore performed a first analysis of the genetic links between the 10 populations of interest, the results of which are consistent with their phylogenetic history.

The second objective was to analyze the functional impact of the 9.9M SNP on transcripts and associated proteins; in particular, we have identified 1590 *stop_gained* SNP that remain to be further analyzed. We have also developed a program to predict the functional impact of double or triple phased SNP within the same codon. We focused on this type of proximate SNP for two reasons: (*i*) these SNP are expected in significant numbers given the number of individuals (more than 300) and populations studied (*ii*) reference tools such as VEP (*Variant Ensembl Predicto*r) predict the consequences of variants on associated proteins by considering only the SNP independently of each other. However, the GATK suite now provides in its output files the phase for short-range SNP and to our knowledge, none of the annotation tools use this information. We show that these phased SNP in codons are still rare (0.2% of variants) but not negligible in number here (~20 000), and that about 70% of the functional predictions on these double or triple phased SNP codons are erroneous.

In conclusion, this work shows that RNA sequencing data can be used to access variants of the genome in the expressed regions. These variants are in sufficient number to be used in genetic diversity or genotype-phenotype association studies. This study also highlights the importance of the phase between close SNP in predicting consequences in protein coding regions.

REMERCIEMENTS

Je tiens tout d'abord à remercier Sandrine Lagarrigue ainsi que Frédéric Jehl pour l'encadrement que j'ai reçu durant ces cinq mois de stage. Merci à vous deux pour la pédagogie dont vous avez fait preuve mais également pour la patience et la bonne humeur qui vous caractérisent. Merci pour l'écoute bienveillante que vous avez su faire perdurer malgré les conditions pas toujours optimales imposées par cette période de confinement.

Merci à toi, Frédéric J., qui tel le bruit blanc de l'eau a su, tous les jours, supporter mes moments de doutes et d'incompréhension. J'espère que bientôt, tu monteras une amicale des *duck debugging* ? Un club ? Une association peut-être ? Merci pour l'accompagnement tant professionnel qu'amical.

Sandrine, merci pour l'autonomie que tu m'as laissé tout en conservant ce regard bienveillant et critique permettant d'avancer rapidement et toujours avec motivation. Merci pour les réponses à mes questions que ce soit à toutes heures du jour ou de la nuit. La passion que tu dégages se révèle être une réelle source d'inspiration.

Je tiens à remercier également Sophie, Pauline et Laeticia qui, à coup de blagues et gags savaient animer les pauses et repas pour repartir travailler dans la joie et la bonne humeur. Merci pour votre gentillesse et votre bienveillance notamment lors de la période de confinement, votre soutien sans faille a été une vraie source de gaieté.

Merci également à Frédéric L., qui à plusieurs reprises a su m'apporter de l'aide que ce soit dans le cadre du travail ou pour développer mon humour.

Je tiens également à remercier Colette, Morgane et Marie-Emmanuelle qui ont toujours été là pour savoir si tout allais bien et pour leurs conversations très enrichissantes.

Pour finir, je tiens à remercier l'ensemble des personnes que j'ai pu côtoyer et qui plus d'une fois ont vu leurs programmes chamboulés par mon arrivée impromptue dans leur bureau et cela bien souvent sans rapport avec le travail.

TABLE DES MATIERES

Introduction **1**

Matériels et méthodes **4**

• Matériels de départ 4

→ *Provenance des données de RNA-seq et de DNA-seq*

→ *Génération des données de RNA-seq et de DNA-seq*

• Création des fichiers VCF 4

• Filtration des variants 4

• Quantification de l'expression génique 5

• Quantification de l'expression exonique 5

• Détection des homopolymères et jonctions exon-exon 5

• Obtention des fréquences génotypiques et alléliques 5

• Analyse exploratoire des liens génétiques entre populations à partir de leurs fréquences génotypiques 6

• Prédiction de l'impact fonctionnel des mutations sur les protéines 7

Résultats **7**

• Comparaison du séquençage ARN et ADN pour la détection de variants 7

• L'ajout de données de tissus différents augmente le nombre de SNP détectés par RNA-seq 9

• Application : Caractérisation génétique de 10 populations de poulets à partir de SNP détectés par RNA-seq 10

• Prédictions des conséquences fonctionnelles par l'outil *Variant Effect Predictor* (VEP) des SNP détectés par RNA-seq 12

Discussions **13**

• Détection de SNP et génotypes par RNA-seq – comparaison avec le DNA-seq 13

• Caractérisation génétique de 10 populations de poules à partir des SNP de RNA-seq 16

• Prédictions des conséquences fonctionnelles associées aux SNP détectés 17

Références

ABREVIATIONS

ACP : Analyse en composantes principales

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

ASE : Allele Specific Expression / Expression allèle spécifique

CAH : Classification ascendante hiérarchique

CDS : Coding DNA Sequence / Séquence codante

CT : Contrôle

DNA-seq : DNA-sequencing / séquençage de l'ADN

ENCODE : Encylcopedia of DNA element

FS : Fisher Strand

GATK : Genome Analysis ToolKit

GT : Génotype

HS : Heat Stress / Stress Thermique

INDEL : Insertion-Deletion

MAF : Minor Allele Frequency / Fréquence de l'allèle mineur

NMD : Nonsense-Mediated mRNA Decay

Pb : Paire de bases

QD : Qual by Depth

RIR : Rhode Island Red

RNA-seq : RNA-sequencing / Séquençage de l'ARN

RpKb : Read per Kilobase / Lectures par kilobases

RSEM : RNA-Seq by Expectation-Maximization

SNP : Single nucleotide polymorphism / Polymorphisme mononucléotidique

STAR : Spliced Transcripts Alignment to a Reference

TPM : Transcripts Per Million / Transcrit par million

UTR : UnTranslated Region / Région non traduite

VEP : Variant Effect Predictor

WES : Whole Exome Sequencing / Séquençage de l'exome entier

WGS : Whole Genome Sequencing / Séquençage du génome entier

INTRODUCTION

Le RNA-seq (*RNA sequencing*, séquençage de l'ARN) est aujourd'hui surtout utilisé pour quantifier l'expression du transcriptome (ensemble des transcrits exprimés dans un échantillon), en remplacement des puces à gènes (Mortazavi et al., 2008). Il permet par exemple d'étudier les régulations génétiques, les processus impliqués dans des maladies (Savary et al., 2020), les réponses à des variations d'environnement (Jehl et al., 2019a) ou de comprendre des différences entre lignées aux phénotypes contrastés (Gondret et al 2017). Il permet également de modéliser des gènes et transcrits (Muret et al., 2017), d'analyser l'expression allèle-spécifique (Lagarrigue et al., 2013b) ou d'étudier l'édition de l'ARN, (Roux et al., 2015). Cependant les mécanismes d'édition sont décrits comme des processus rares quelles que soient les espèces (souris : Lagarrigue et al., 2013a, poule : Roux et al., 2015, Frésard et al., 2015). Ainsi, par son étape de séquençage, le RNA-seq permet d'accéder à la séquence des transcrits et donc aux micro-variations du génome que l'on retrouve sur les transcrits, comme proposé par Piskol et al. en 2013, ces micro-variations étant en majorité des substitutions d'un nucléotide (SNP, *Single Nucleotide Polymorphism*) et en minorité des insertions-délétions (INDEL). Ces données de RNA-seq, encore peu exploitées actuellement à des fins de détection de variants génomiques, sont pourtant intéressantes à plusieurs titres. D'abord, elles permettraient d'étudier les variations dans les régions codantes des zones transcrites (*Coding DNA Sequence*, CDS). Les variations dans ces régions, quoique plus rares car contre-sélectionnées, participent dans une certaine mesure à la variation des caractères et des maladies (Pickrell, 2014). La possibilité de faire un lien entre un gène rendu dysfonctionnel par un variant délétère et un phénotype est une manière d'inférer un rôle ou une fonction pour le gène en question (Karczewski et al., 2020). Bien que chez l'être humain, ces régions soient accessibles par séquençage de l'exome (*Whole Exome Sequencing*, WES), les banques d'oligonucléotides nécessaires à la capture de l'exome sont peu ou pas disponibles pour les espèces non-modèles comme les espèces d'élevage. Ensuite, puisque les régions transcrites sont globalement bien réparties sur le génome, accéder aux polymorphismes qu'elles portent donnerait accès à un jeu de SNP bien répartis eux aussi, permettant d'étudier la diversité génétique des populations d'une espèce ou de réaliser des études d'association entre génotypes et phénotypes, pour identifier des régions impliquées dans des caractères complexes. Ces études sont aujourd'hui réalisées grâce à des puces de génotypages. Cependant, pour sélectionner les SNP utilisés dans ces puces, qui doivent être polymorphes sur un maximum de populations et assez nombreux (les puces de génotypages de moyenne densité portant environ 50K SNP ; Groenen et al., 2011), il est nécessaire de séquencer le génome entier (*Whole Genome Sequencing*, WGS) de plusieurs individus pour chaque population d'intérêt par DNA-seq, ce qui est très coûteux. Cela limite donc

l'identification de ces SNP aux populations les plus étudiées de l'espèce d'intérêt, et réduit donc d'autant le potentiel d'utilisation de ces puces dans des espèces non-modèles, comme les espèces d'élevage, pouvant être caractérisées par un grand nombre de populations.

Ainsi, le RNA-seq peut permettre (*i*) d'accéder aux SNP des régions codantes, ouvrant la voie à l'étude de leurs effets sur la fonction des protéines, et si leur nombre est suffisant de (*ii*) disposer de marqueurs pour des études de diversité et d'association, notamment dans des populations peu étudiées. Les données de *WES* ou *WGS* sont limitées par leur coût, alors qu'une grande quantité de données de RNA-seq se sont accumulées dans les espèces d'élevage depuis plusieurs d'années, issus de dispositifs de quelques dizaines d'animaux de fonds génétiques différents et placés dans des conditions expérimentales variées.

Cependant, cette ressource est encore peu exploitée car le transcriptome est plus complexe à étudier que le génome. En effet, il est notamment composé de transcrits matures (c-à-d épissés), avec des exons issus de régions génomiques distantes (car séparées par les introns) rendant plus difficile leur alignement sur le génome par rapport aux séquences de DNA-seq (Pan et al., 2008). De plus, les transcrits ont des niveaux d'expression très variables, entraînant des profondeurs de lectures (*reads)* variées d'une position génomique à une autre (d'une dizaine de *reads* à plusieurs milliers), contrairement au DNA-seq qui offre une profondeur homogène sur tout le génome. Ce point est à prendre en compte pour la détection par RNA-seq du polymorphisme et surtout des génotypes individuels (Sims et al., 2014).

Alors que des méthodes d'alignement de données de RNA-seq sont bien maitrisées, les procédures de détection des SNP sont encore à l'étude. A notre connaissance, depuis Piskol et al., 2013, une dizaine de travaux (Quinn et al., 2013, Tang et al., 2014, Wang et al., 2014, Wolfien et al., 2016, Guo et al., 2017, Oikkonen et al., 2017, Cornwell et al., 2018, Adetunji et al., 2019) ont proposé des outils de détection de SNP à grande échelle à partir de données RNA-seq. Parmi eux, seul Adetunji et al., 2019 utilise les outils de référence en RNA-seq proposés par ENCODE : STAR pour l'alignement (Dobin et al., 2013) et GATK pour la détection des variants (Auwera et al., 2013). Trois de ces études se sont intéressées à la concordance entre RNA-seq et DNA-seq des variants et génotypes, le DNA-seq constituant la référence. Cependant, ces études se limitaient à quelques échantillons (< 4 pour Piskol et al., 2013, Adetunji et al., 2019), voire à un seul (Guo et al., 2017) et ne disposaient jamais des deux types de données sur les mêmes échantillons.

Dans ce contexte, les travaux entrepris durant mon stage ont pour objectif de détecter les SNP à partir de données RNA-seq chez la poule. La poule est une espèce d'importance économique dont les produits (œufs et viande) sont consommés dans le monde entier (FAO, 2020). Elle constitue aussi un bon modèle d'étude du développement car l'embryon est

facilement accessible et manipulable (Brown et al., 2003). Elle fait donc l'objet de nombreuses études transcriptomiques visant à mieux comprendre sa physiologie et sa génétique. Ainsi, dans le cadre de projets financés par l'Agence nationale de la recherche ou la commission européenne, mon équipe d'accueil a acquis des données RNA-seq sur plus de 700 échantillons issus de populations commerciales et expérimentales, de chair ou de ponte.

Le premier objectif du stage a été de finaliser un *pipeline* (suite de programmes informatiques) de détection de SNP fiables à partir de données de RNA-seq, en utilisant les outils de référence STAR et GATK. La fiabilité des SNP détectés par RNA-seq (notés SNP de RNA-seq) a été évaluée en prenant pour référence les SNP détectés par DNA-seq (SNP de DNA-seq). Nous avons également commenté un filtre proposé par GATK pour les SNP de RNA-seq. L'étude a été réalisée sur deux populations indépendantes, l'une de ponte l'autre de chair, de 15 et 8 individus respectivement et pour lesquelles les données de RNA-seq et DNA-seq étaient disponibles sur les mêmes échantillons. Nous avons ensuite étudié l'impact de l'ajout de nouveaux tissus sur le nombre de SNP détectés. Enfin, à partir de ces SNP, nous avons calculé les fréquences génotypiques et alléliques à l'échelle des populations.
Nous avons appliqué ce *pipeline* à 10 populations différentes de poules avec deux objectifs : (*i*) donner des ordres de grandeur du nombre de SNP et génotypes détectables par RNA-seq pour chacune des populations et sur l'ensemble des population (union et intersection) et (*ii*) caractériser les liens génétiques entre ces 10 populations à partir de ces SNP de RNA-seq.

Le second objectif a été de caractériser les conséquences des SNP dans les régions codantes détectés par RNA-seq dans les 10 populations sur les protéines. En particulier, nous nous sommes consacrés à l'étude de l'impact fonctionnel de doublets ou triplets de SNP situés dans un même codon et avons développé un programme permettant de les repérer et de prédire leur impact. Nous avons étudié ces SNP car (*i*) quoiqu'ils soient attendus comme rare, le fait de travailler sur 10 populations génétiquement différentes devrait nous permettre d'en détecter un certain nombre, et (*ii*) les outils de référence pour la prédiction de l'impact fonctionnel des variants tel que VEP (*Variant Effect Predictor*, McLaren et al., 2016), ANNOVAR (Wang et al., 2010) ou SnpEff (Cingolani et al., 2012), prédisent les conséquences fonctionnelles en considérant les SNP indépendamment les uns des autres. Or, il est aujourd'hui possible d'accéder à la phase pour les SNP à courte distance, ce qui permet de savoir s'ils affectent bien le même codon. A notre connaissance, aucun des outils d'annotation n'utilise pour l'instant cette information.

MATERIELS ET METHODES

Matériels de départ

• *Provenance des données de RNA-seq et de DNA-seq* : Les données de RNA-seq sont issues de 10 populations de poules : (*i*) la race ancêtre des races actuelles la Red junglefowl (RJFh) ; (*ii*) deux lignées de poulets de chair, une commerciale, la Cobb 500 (Cobb Vantress), une expérimentale, la "Gras/Maigres" (FLLL) ; (*iii*) quatre lignées de pondeuses : deux commerciales à oeufs bruns de souche Rhode Island Red (RIR), A3A3 et N4A3 (NOVOGEN), deux expérimentales, RpRm également à oeufs bruns et de souche RIR et FrAg à oeuf blanc de souche Leghorn ; (*iv*) deux populations non sélectionnées : la Cou Nu (LSnu), composée d'individus nains sans plume au niveau de son cou (Mou et al., 2011) et la Fayoumi (FAyo), race égyptienne en conservation ; (*v*) la Rmx6, population atypique, à oeufs blancs et très divergentes génétiquement de nos autres populations. Dans le détail, les populations FLLL et RpRm sont composées chacune de 2 lignées divergentes respectivement pour le pourcentage de gras corporel (lignées FL et LL) et l'efficacité alimentaire (lignées Rp et Rm). Elles comprennent des individus spécifiques de chacune des 4 lignées et des individus hybrides (Rp x Rm ou FL x LL). Au final, nous disposons de 337 individus pour ces 10 populations dont 264 sans les individus hybrides. L'ARN étant extrait de plusieurs tissus (entre 1 et 5) pour un même individu, un total de 744 échantillons RNA-seq sont disponibles. Les données DNA-seq sont issus du foie de 15 individus RpRm et 8 individus FLLL. Ces deux groupes ont été utilisées pour le premier objectif de l'étude (comparaison des résultats de détection de SNP et de génotypes entre RNA-seq et DNA-seq).

• *Génération des données RNA-seq et DNA-seq* : Les données de DNA-seq et RNA-seq ont été produites sur séquenceurs Illumina de type HiSeq avec entre 30 à 40 millions de fragments séquencés par échantillons pour le RNA-seq et en 20X moyen pour le DNA-seq. Ces fragments ont été en général séquencés à leurs deux extrémités sur 150 paires de bases (appelés *reads*) et de façon orientée pour le RNA-seq. Les *reads* des fichiers *.fastq* ont été nettoyés des adaptateurs par TrimGalore v0.4.5 (Krueger, 2020), puis alignés sur le génome de référence *Gallus_gallus_5* de *Ensembl* à l'aide de STAR v.2.5.2. La détection de variants à partir des fichiers *.bam* issus de l'alignement a été réalisée avec la fonction "HaplotypeCaller" de GATK v3.7.0. afin de générer un fichier *.gvcf* par individu. Ces étapes ont été réalisées en amont par l'équipe d'accueil.

Création des fichiers VCF. A partir des *.gvcf* individuels et avec la fonction "GenotypeGVCFs" de GATK v3.7.0, nous avons généré pour chaque population et tissu un fichier global *.vcf* regroupant tous les SNP polymorphes pour cet ensemble (population/tissu).

Filtration des variants. A partir de ces fichiers *.vcf*, les SNP bi-alléliques ont été sélectionnés avec GATK v3.7.0 avec l'option "--restrictAllelesTo BIALLELIC" et l'option

“--excludeNonVariants“ de l'outil "SelectVariants" et ont ensuite été filtrés à l'aide de l'outil "VariantFiltration" comme suit. Pour les SNP de RNA-seq, nous avons en partie utilisé les filtres proposés (mais non validés) par GATK (GATK, 2017) : (1) le filtre "QD < 2” - qualité du variant normalisé par le nombre de *reads* associés - pour supprimer les SNP dont la qualité totale est biaisée par un nombre de *reads* important ; (2) le filtre "FS > 30" pour supprimer les SNP avec un biais de brin, c’est-à-dire pour lesquels l'allèle alternatif est détecté de manière disproportionnée dans des *reads* alignés sur un brin plutôt que l’autre. En revanche, nous n’avons pas retenu le filtre *SnpCluster* évinçant les SNP regroupés par 3 ou plus dans une fenêtre glissante de 35 nucléotides, comme expliqué dans la partie *Résultats*. Pour les SNP issus du DNA-seq, nous avons appliqué les filtres recommandés par GATK (GATK, 2020) et utilisés par de nombreux auteurs.

Quantification de l’expression génique. L'expression des gènes a été quantifiée avec RSEM v.1.3.0 (Li et al., 2011) avec le fichier GTF étendu de Jehl et al. 2019b, avec la métrique TPM (*Transcript Per Million*) classiquement utilisée en RNA-seq. Un gène est ainsi considéré comme exprimé si son TPM est supérieur à 0,1 et avait au moins 5 *reads* - comme dans le projet GTEx (Goede et al., 2019) - dans au moins 80 % des échantillons d’une condition de notre population comme fait au laboratoire en routine. (Jehl et al., 2019b)

Quantification de l’expression exonique. L'expression des exons a été estimée avec FeatureCount v1.6.2 (Liao et al., 2014) avec les options -t “exon” et -g “exon_id”. L’expression à ce niveau étant peu étudiée, nous avons défini une métrique *RpKb* (*Read per Kilobase*) calculée à l'aide de la formule : $RpKb = \frac{\sum_i^n R_i}{l \times n} \times 10^3$ où R_i correspond au nombre de *reads* totaux, l la taille de l’exon et n au nombre d’individus. Pour sélectionner les exons exprimés, c’est à dire ayant une expression supérieure au bruit de fond, nous avons évalué ce dernier en quantifiant l'expression de régions aléatoires sur les chromosomes 1 à 33 ayant la même distribution de taille que les exons connus, mais situées à au moins 1kb des régions annotées dans le GTF (comme dans Jehl et al., 2019b). Comme attendu la distribution des régions aléatoires est bien en deçà de celle correspondant aux régions exoniques ; le 3ème quartile de ces régions aléatoires et le 1er quartile de l’expression des exons étant égales à 0.5 RpKb, cette valeur nous semble être un bon seuil pour définir un exon exprimé.

Détection des homopolymères et jonctions exon-exon. Les zones où plusieurs nucléotides identiques se succèdent (appelées homopolymères), et les zones de jonctions composées de 5 pb de part et d’autre de l’exon, ont été analysées à l’aide de scripts *ad hoc*.

Obtention des fréquences génotypiques et alléliques. Les fréquences génotypiques et alléliques ont été calculées à partir des fichiers *.vcf* propres à chaque population et contenant les SNP détectés à l’échelle de la population ainsi que les génotypes associés à chaque individu

(notés GT). Les calculs n'ont été réalisés qu'avec les SNP pour lesquels les génotypes satisfaisaient des critères définis dans le laboratoire par Frédéric Jehl et al., (thèse en cours), à savoir un pourcentage de GT renseignés dans le *.vcf* pour le tissu et la population étudiée supérieur à 50% parmi les N individus de la population et un pourcentage de GT supportés par au moins 5 *reads* supérieur à 20% parmi les N individus de la population. Le premier critère a un fort impact sur le nombre de génotypes retenu ce qui est essentiel lorsqu'on veut calculer des fréquences génotypiques et alléliques. Le second critère apporte de la fiabilité dans les génotypes retenus. En effet, des SNP fiables peuvent être détectés au niveau de la population en cumulant les *reads* de tous les individus, sans pour autant qu'il y ait suffisamment de *reads* par individu pour avoir des génotypes avec une bonne confiance, comme illustré dans la Fig. 1 par le "SNP2" du "tissu2". Par ailleurs, les individus analysés en transcriptomique peuvent être différents par leur origine génétique ou encore par les conditions environnantes auxquels ils sont soumis (contrôle CT *vs* un stress HS). Ils peuvent donc avoir des niveaux d'expressions variables pouvant conduire à ce que certains individus aient un génotype renseigné alors que d'autres non (cf SNP1, tissu1, CT *vs* HS). C'est pourquoi les filtres évoqués plus haut sont nécessaires pour obtenir des génotypes fiables et en nombre suffisant afin de calculer des fréquences génotypiques et alléliques. Pour finir, pour les populations avec plusieurs tissus, si le SNP était détecté dans plusieurs d'entre eux, seules les informations provenant du tissu avec le nombre de *reads* maximal au sein de la population étaient conservées, ceci afin de maximiser la fiabilité des informations génotypiques comme visible dans la Fig 1 "pop. final".

pop1_tissu1_SNP.tsv

	DPtot	ind.1	ind.2	ind.3	ind.4
		CT	CT	HS	HS
SNP1	108	55	45	5	3
		0/1	0/1	./.	./.
SNP2	159	41	37	39	42
		1/1	1/1	1/1	0/1

pop1_tissu2_SNP.tsv

	DPtot	ind.1	ind.2	ind.3	ind.4
		CT	CT	HS	HS
SNP1	658	164	154	140	200
		0/1	0/1	1/1	0/0
SNP2	12	2	1	4	5
		./.	./.	./.	./.

pop. final

	DPtot	ind.1	ind.2	ind.3	ind.4
		CT	CT	HS	HS
SNP1	658	164	154	140	200
		0/1	0/1	1/1	0/0
SNP2	159	41	37	39	42
		1/1	1/1	1/1	0/1

Fig. 1. Données issues de deux fichiers *.vcf* correspondant à une même population mais à deux tissus différents (tissu 1 à gauche et tissu 2 au milieu). En colonne, 4 individus (ind. n°j) soumis à deux conditions CT et HS; en ligne, le nombre de *reads* observé pour le SNP i et l'ind. j et les génotypes associés avec 0/0 et 1/1 génotypes homozygotes pour respectivement l'allèle de référence et l'allèle alternatif et 0/1 génotype hétérozygote; ./. génotype manquant (non analysable); DPtot pour un SNP et un tissu : somme des *reads* cumulés au sein des individus de la population pour ce SNP et ce tissu. A droite, illustration de la gestion de l'information "multi-tissu" avec pour chaque SNP, sélection du tissu portant le plus d'information au regard de la DPtot.

Analyse exploratoire des liens génétiques entre populations à partir de leurs fréquences génotypiques. L'analyse exploratoire a été réalisée avec les 264 individus caractérisant les 9 populations pour lesquelles les données de foie étaient disponibles soit 269 239 SNP communs et polymorphes. Une classification ascendante hiérarchique (CAH) ainsi qu'une analyse en

composantes principales (ACP) ont été effectuées en utilisant respectivement les fonctions "snpgdsHCluster" et "snpgdsPCA" du package SNPRelate v1.8.0. (Zheng et al., 2012) prenant en entrée les fichiers *.vcf* et utilisant le génotype de chaque individu pour chaque SNP. Les génotypes sont au préalable convertis sous format numérique (0 = homozygote référence, 1 = hétérozygote, 2 = homozygote alternatif, 3 = non renseigné). La distance entre deux individus repose sur les corrélations de ces vecteurs (GT) pour les 269 239 SNP. Pour la CAH, le critère d'agrégation utilisé est la moyenne des distances.

Prédiction de l'impact fonctionnel des mutations sur les protéines. La prédiction des conséquences fonctionnelles a été réalisée avec l'outil *Variant Effet Predictor* (VEP) de *Ensembl* avec l'annotation génique v92 sur le génome de référence *Gallus_gallus_5*. Un programme sous R v3.3.3 a été développé pour prendre en compte la présence de plusieurs variants au sein du même codon. Cette re-prédiction n'était effectuée que si les phases étaient renseignées dans les fichiers *.vcf* pour les doubles ou triples SNP au sein d'un même codon.

RÉSULTATS

Comparaison du séquençage ARN et ADN pour la détection de variants

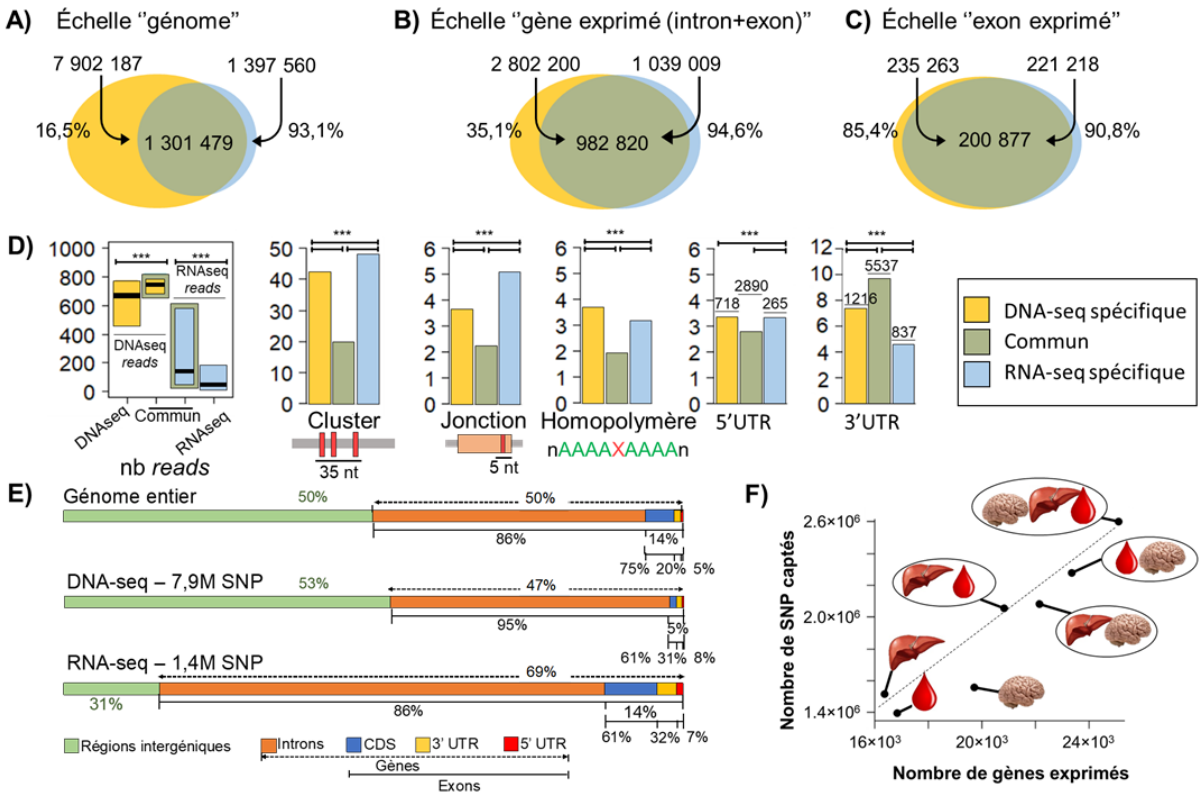

Fig. 2. Etude comparative des SNP issus des données RNA-seq et DNA-seq. Ensemble des SNP détectés au niveau du génome (A), des gènes exprimés (exons + introns) (B) et des exons exprimés (C). (D) Caractérisation des SNP détectés à l'échelle des exons exprimés dans chaque ensemble. (E) Distribution des SNP au sein des différentes régions génomiques par rapport à leur surface couverte dans le génome entier. (F) Nombre de SNP captés dans 1, 2 ou 3 tissus en fonction du nombre de gènes exprimés pour un ou plusieurs tissus {foie, sang, hypothalamus}.

Nous avons détecté 10 167 286 variants en utilisant les données de DNA-seq provenant du foie de 15 individus RpRm. Parmi eux, 8,8M (87%) étaient des SNP, les 13% restant des INDEL. L'exclusion des SNP non-biallélіques et l'application des filtres classiques de GATK (voir *Matériels et Méthodes*) a permis d'extraire 7 902 187 SNP, soit 78% des SNP initiaux. Pour les données issues du RNA-seq, à partir des 1 563 917 variants initiaux, 1,4M (90%) étaient des SNP, les 10% restant des INDEL. L'exclusion des SNP non-biallélіques et l'application des filtres suggérés par GATK (*SnpCluster* exclus, comme discuté plus loin) a réduit ce nombre à 1 397 560 SNP, soit 89% des SNP initiaux. Nous avons observé que 93% des SNP de RNA-seq l'ont été également par DNA-seq (Fig. 2-A). Comme le RNA-seq ne permet de capter que les variants dans les régions exprimées, nous avons comparé les deux méthodes en nous concentrant sur les gènes exprimés (introns et exons) (Fig. 2-B). Le RNA-seq a ainsi permis de détecter 1 039 009 SNP. Enfin, nous nous sommes focalisés sur les exons exprimés pour être à région analysée équivalente entre les 2 technologies (Fig. 2-C). Les exons ont été sélectionnés sur leur expression comme décrit en *Matériels et Méthodes*. Dans ces régions, 221 218 SNP ont été captés en RNA-seq contre 235 263 SNP en DNA-seq avec une intersection de 200 877 SNP correspondant à 85% des SNP détectés par DNA-seq et à 91% des SNP captés par RNA-seq. Des résultats semblables ont été observés pour la lignée de chair FLLL composée de 8 individus, l'intersection des deux technologies (250 071 SNP) correspondait à 72% des SNP captés par DNA-seq et 91% des SNP détectés par RNA-seq.

Afin d'étudier les SNP non détectés par l'une ou l'autre méthode, nous avons analysé différents éléments (Fig. 2-D). D'abord, 20% des SNP détectés par les deux méthodes correspondaient à des *SnpCluster*, ce pourcentage était significativement plus élevé (48% et 43%, χ^2;$p<10^{-16}$) pour les SNP spécifiques au RNA-seq et au DNA-seq respectivement. Au vu de ces résultats, nous n'avons donc pas filtré les SNP sur ce critère (voir *Discussions*). Nous avons également observé que les SNP spécifiques d'une seule méthode étaient couverts par un nombre de *reads* de cette méthode significativement inférieur (t-test;$p<10^{-16}$) à celui des SNP détectés par les deux méthodes. De plus 5,07% des SNP spécifiques au RNA-seq se trouvaient à 5pb ou moins d'une jonction exon-exon, contre 3,66% et 2,22% pour ceux détectés respectivement par DNA-seq et par les deux techniques (χ^2;$p<10^{-16}$). De plus, 1,93% des SNP détectés par les deux méthodes étaient situés dans des régions dites homopolymères de degré 5 - zones où au moins 5 nucléotides identiques se succèdent - contre 3,20% et 3,69% pour ceux respectivement détectés par RNA-seq ou DNA-seq (χ^2;$p<10^{-16}$). De même, nous avons observé que les SNP spécifiques au RNA-seq étaient significativement moins présents (χ^2;$p<10^{-16}$) dans les régions 3'UTR des gènes que ceux des deux autres ensembles.

Nous avons également comparé la répartition des SNP détectés par DNA-seq et RNA-seq dans le génome (Fig. 2-E). Tout d'abord, nous montrons que le génome de poule est composé de 50% de séquences géniques (intron + exon), parmi lesquelles 86% sont de la séquence intronique et 14% de la séquence exonique. Cette dernière est composée à 75% de séquences codantes, 20% de 3'UTR et 5% de 5'UTR. Dans les régions géniques (exon + intron), les SNP détectés par DNA-seq sont en proportion plus élevée dans les introns que la part de ces régions dans le génome (95% *vs* 86%) alors qu'on observe un déficit de ces SNP dans les régions exoniques par rapport à la proportion de ces régions dans le génome (5% *vs* 14%). Avec les données RNA-seq, nous avons observé sans surprise une majorité de SNP dans les régions géniques, et notamment dans les introns même si la proportion est moindre que celle observée en DNA-seq, tous les introns étant loin d'être exprimés (69% *vs* 95%). De plus, 31% des SNP détectés en RNA-seq appartiennent à des zones intergéniques correspondant donc à des zones du génome non annotées. Enfin, si on analyse plus finement les régions exoniques, on observe un excès de SNP dans les régions 5'UTR et 3'UTR (+2-3% et +11-12% respectivement par rapport à la proportion qu'occupent ces deux types de régions dans les exons) ; en revanche on observe un déficit de SNP dans les régions codantes (CDS) de -14% avec les 2 méthodes. En résumé, ces résultats montrent donc une répartition non aléatoire des SNP dans les différentes régions du génome, qui sera discutée dans le chapitre *Discussion*.

L'ajout de données de tissus différents augmente le nombre de SNP détectés par RNA-seq

En utilisant des échantillons de sang et d'hypothalamus des 15 animaux du paragraphe précédent, nous avons étudié l'effet d'un nombre de tissu croissant sur le nombre de SNP détectés par RNA-seq. Comme vu précédemment 1,4 M de SNP ont été détectés dans le foie. Dans l'hypothalamus 1,56 M de SNP ont été détectés et 1,51 M dans le sang. Entre 2,05 M et 2,28 M de SNP étaient détectés pour l'union de deux tissus et 2,6M pour trois tissus. Nous avons comptabilisé 16 814, 16 346 et 19 733 gènes exprimés respectivement dans le foie, le sang et l'hypothalamus, 20 855 à 23 381 gènes étaient exprimées pour l'union de deux tissus et 25 126 pour trois tissus. Le nombre de SNP obtenus avec un, deux ou trois tissus est fortement corrélé au nombre de gènes exprimés dans ces mêmes ensembles de tissus (Fig. 2-F), (corrélation de Spearman = 0,96). Fort de ces résultats indiquant une bonne détection des SNP par RNA-seq dans les régions exprimées, un fichier *.vcf* par population a été créé et ceci pour chaque tissu disponible ou pour l'ensemble des tissus (voir *Matériels et Méthodes*). Leur contenu en SNP et génotype a été analysé comme indiqué dans la section suivante.

Application : Caractérisation génétique de 10 populations de poulets à partir de SNP détectés par RNA-seq.

Comme indiqué dans la table 1, le nombre de SNP détectés dans chacune des 10 populations dans le foie uniquement va de 1,2M à 4,1M environ. Après analyse de tous les tissus à notre disposition (soit 1 à 5 tissus selon la population), ces nombres varient de ~1,5M à ~5,9M pour la population Cobb. En accord avec les résultats précédents, une hausse substantielle du nombre de SNP est observable lorsque le nombre de tissus analysés augmente (cf table 1 $\Delta\%_1$ allant de 39% à 119%). Pour l'analyse multi-tissu, l'union des SNP polymorphes dans au moins une des 10 populations est de 9 949 072 SNP. L'intersection, correspondant à l'ensemble des SNP ayant au moins un allèle alternatif (noté ALT), i.e. un allèle différent de celui présent dans le génome de référence dans les 10 populations, est de 288 484 SNP.

Table. 1. Nombre de SNP totaux (1), de SNP avec génotypes fiables (2), et de SNP avec génotypes fiables et une MAF ≥ 10% (3), détectés par population à partir des données RNA-seq disponibles ("foie" et "multi-tissu").

Abrv.	Type	Nb ind.	Nb ech.	Nb tiss.	(1) SNP totaux			(2) SNP - GT fiables et en nombre suffisant			3) SNP - GT fiables et en nombre suffisant - MAF ≥10 %			
					Foie	Multi-tissus	Δ%1	Foie	Multi-tissus	Δ%1	Foie	Δ%2	Multi-tissus	Δ%2
RJFh	ancestral	18	72	3	.	2 646 463	.	277 194	583 914	111	151 204	55	324 447	56
Cobb	chair	48	96	2	4 122 798	5 867 458	42	952 757	1 686 055	77	557 528	59	951 059	56
FLLL	chair	32	64	2	1 833 854	3 416 944	86	537 688	1 114 302	107	368 109	68	714 065	64
A3A3	ponte	32	32	1	1 490 773	1 490 773	.	449 768	449 768	.	264 698	59	264 698	59
N4A3	ponte	64	104	2	1 308 551	2 172 690	66	391 955	740 806	89	243 790	62	449 249	61
RpRm	ponte	88	286	5	1 883 978	4 130 952	119	557 563	1 288 405	131	306 928	55	631 358	49
RMx6	ponte	19	19	1	.	2 223 852	.	.	718 531	.	.	.	483 236	67
FrAg	ponte	4	7	2	1 267 812	1 764 922	39	789 382	986 728	25	525 301	67	427 566	43
Lsnu	ponte	16	32	2	1 526 260	2 355 898	54	593 152	840 429	42	384 682	65	534 812	64
Fayo	atypique	16	32	2	1 349 391	2 089 712	55	498 272	701 654	41	288 432	58	396 381	56
MOYENNE					1 847 927	2 815 966	66	560 859	911 059	78	343 408	61	517 687	58
UNION		337	744		5 682 906	9 949 072		1 678 971	3 423 310		1 244 012		2 112 626	
INTERSECTION					294 528	288 484		269 239	85 127		2 717		2 017	

$\Delta\%_1$: gain (en %) de SNP entre les analyses multi-tissus et les analyses "foie". $\Delta\%_2$ gain (en %) de SNP avec GT fiables et MAF ≥ 10% par rapport aux SNP avec GT fiables et en nombre suffisant quelle que soit la fréquence de l'allèle mineur, et ce pour chaque catégorie ("foie" et "multis-tissus").

Si maintenant on considère le nombre de SNP avec des génotypes renseignés (voir critères en *Matériels et Méthodes*), l'analyse "multi-tissu" montre un nombre de SNP variant de ~450 000 à ~1,7M pour la population Cobb. L'union, i.e. l'ensemble des SNP polymorphes dans au moins une des 10 populations et avec des génotypes suffisamment renseignés, est alors de 3,4M de SNP, nettement moins que les 9,9M SNP précédemment observés. Comme expliqué dans le *Matériels et Méthodes*, certains SNP ont pu être détectés à l'échelle de la population grâce à l'accumulation de *reads* dans les individus qui la composent, alors qu'au sein de chaque individu, le nombre de *reads* est insuffisant pour détecter de façon fiable un génotype (Fig. 1, *Matériels et Méthodes*). L'intersection, i.e. l'ensemble des SNP polymorphes et avec des génotypes suffisamment renseignés dans chacune des 10 populations est quant à elle composée de 85 127 SNP. Notons qu'à l'échelle d'un seul tissu, le foie, disponible dans 9 des 10

populations (Rmx6 étant exclu ici), le nombre de SNP communs à ces 9 populations est de 269 239 SNP et atteint curieusement 2 717 SNP lors de l'application d'une MAF (fréquence de l'allèle minoritaire) ≥ 10%, soit 4 observations de cet allèle dans les petites populations de 16 individus, jusqu'à 18 observations pour la plus grande à 88 individus. Une analyse plus fine des fréquences alléliques (résultats non présentés) montre que cette réduction a deux origines : en moyenne 63% des SNP perdus après application du filtre MAF ont un des deux allèles en très faible fréquence dans la population (fréquence allélique < à 10%), les 37% restant correspondent à des SNP pour lesquels l'allèle alternatif (ALT) est fixé (freq. ALT = 100%). Ce dernier pourcentage varie de 7% pour la Cobb à 50% pour la RpRm. Nous avons ensuite étudié les liens génétiques entre les populations par ACP et CAH (voir *Matériels et Méthodes*) à partir des génotypes REF/REF, ALT/REF et ALT/ALT des 269 239 SNP polymorphes, i.e. ayant au moins un allèle ALT par rapport au génome de référence, dans les 9 populations. Les résultats sont présentés en figure 3.

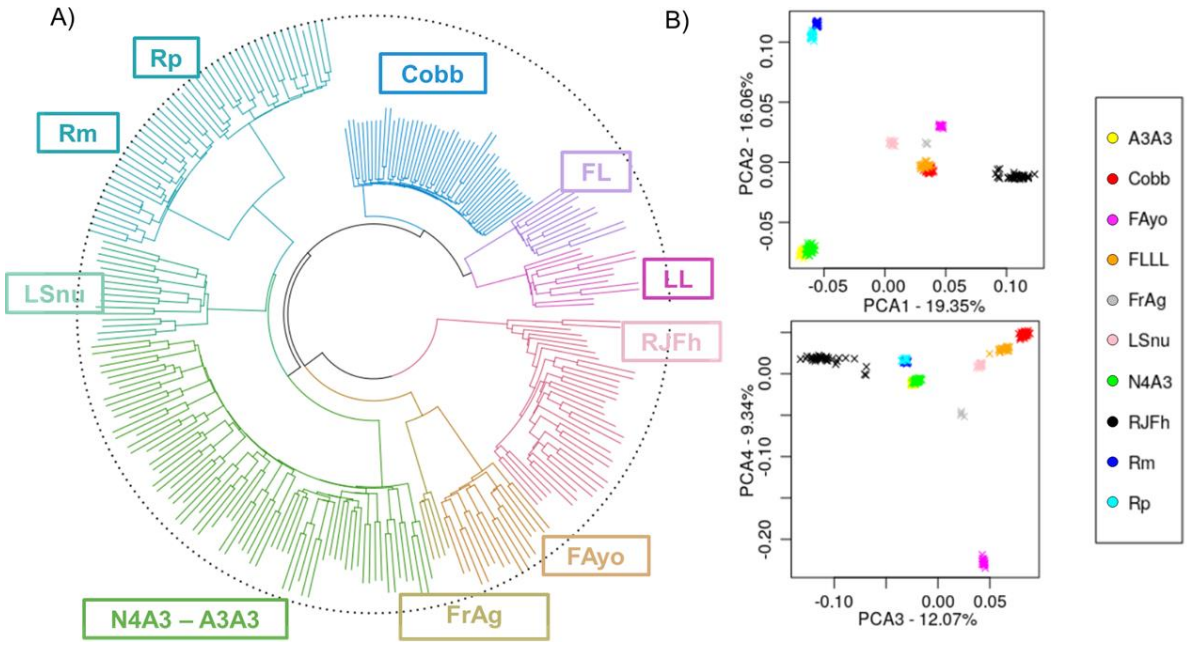

Fig. 3. Observation des liens génétiques entre populations par classification ascendante hiérarchique (A) et analyse en composantes principales (B). Ces analyses ont été réalisées à partir des 269 239 SNP communs aux 9 populations.

On observe que la CAH permet de bien séparer nos populations (Fig. 3-A) comme discuté dans la partie *Discussion*. Concernant l'ACP (Fig. 3-B), les 4 premiers axes ont été utilisés car ils résument à eux seuls 56% de la variance entre nos populations, chacun d'eux ayant une contribution explicative de la variance assez élevée de 9% à 19% et contribuant à la séparation de populations différentes selon des combinaisons de SNP différentes. Nous envisageons de réitérer l'analyse avec la liste très réduite de 2 717 SNP caractérisés par une MAF ≥ 10% pour voir si les séparations entre populations sont similaires et évaluer ainsi l'influence des allèles rares et/ou fixés.

Prédictions des conséquences fonctionnelles par l'outil *Variant Effect Predictor* (VEP) des SNP détectés par RNA-seq

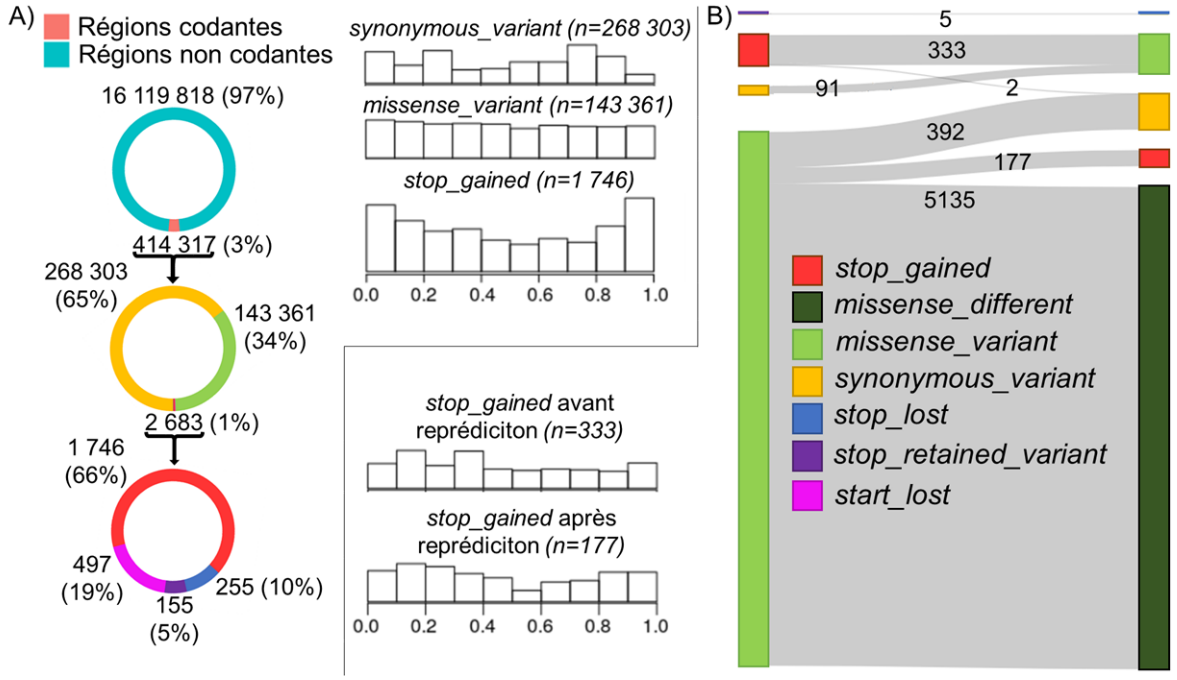

Fig. 4. Analyse des prédictions des conséquences fonctionnelles des 9,9M de SNP (A) et re-estimation de ces prédictions dans le cas de 2 SNP phasés présents dans un même codon (B). Au milieu, distribution par décile de la position des SNP dans la séquence codante pour (en haut - A) 3 catégories de conséquences avant ré-évaluation et pour (en bas - B) la catégorie *stop_gained* avant et après ré-évaluation. La légende présente dans la figure (B) est commune à la figure (A).

L'utilisation de VEP sur les 9,9M SNP constituant l'ensemble des SNP de RNA-seq présents au sein de nos 10 populations a permis de générer 16,5M prédictions de conséquences de variant pour chaque transcrit et protéine associée, référencés dans le fichier d'annotation *Ensembl* v92. Comme vue précédemment, la majorité des SNP sont dans les régions non codantes (Fig. 4-A). Dans les régions codantes, 65% des conséquences sont de types silencieuses (noté *synonymous_variant)* correspondant à un changement de nucléotide sans changement de l'acide aminé, 34% sont des faux-sens (noté *missense_variant)* correspondant à un changement de nucléotide et d'acide aminé dans la protéine associée et 1% (2683) provoque l'apparition ou la disparition de codons *start* ou *stop* ou la modification d'un codon *stop (*noté *stop_retained_variant)*. Dans ce dernier ensemble, 1 746 variants provoquent l'apparition d'un codon *stop* (noté *stop_gained*). Les prédictions associées à la perte d'un codon *start* (noté *start_lost)* présente un taux de faux positif de l'ordre de 20% (100 parmi 509) dans le cas de notre étude. En effet, en absence de codon *start*, annoté dans le fichier *.gtf*, VEP considère que le premier codon du premier CDS est censé être un codon *start* ATG ainsi si un variant est présent dans ce codon, il sera indiqué comme provoquant la disparition du codon *start* même si il n'est pas localisé dans un ATG.

La figure 4-B présente la distribution des conséquences par déciles de la séquence codante associée. Les *missense_variant* et *synonymous_variant* se répartissent uniformément dans la séquence codante alors que les *stop_gained* sont plus fréquents en début et fin de séquence codante. Cependant, sachant que les prédictions sont faites en considérant uniquement les SNP indépendamment les uns des autres, certaines prédictions peuvent être erronées. Ainsi, nous avons reconsidéré les conséquences prédites pour les variants présents au sein d'un même codon et phasés, c'est à dire portés par la même molécule d'ADN et pour lesquels les haplotypes sont indiqués dans le fichier *.vcf*. 9 178 codons possèdent 2 SNP phasés soit environ 18 000 SNP sur 9,9M de SNP (0,2%) et ont donc été ré-évalué en termes d'impact (Fig. 4-B). Par souci de clarté, lorsque deux conséquences ou plus étaient prédites dans un seul codon, nous ne retenons qu'une conséquence selon la priorisation suivante : *synonymous_variant* puis *stop_retained_variant*, *missense_variant*, *stop_lost*, *start_lost*, *stop_gained*. Ainsi, pour les 9 178 codons possédant deux SNP phasés, 96% (333/346) des conséquences de type *stop_gained* correspondraient en réalité à des *missense_variant*. A l'inverse, 177 codons indiqués comme *missense_variant* (parmi les 8 581, soit 2%) seraient de nouveaux *stop_gained*. La distribution par déciles dans la séquence codante a été observée pour ces différents cas (Fig. 5-A et B - barplots). Celle correspondant aux *stop_gained* ré-évalués se rapproche de la distribution obtenue pour l'ensemble des *stop_gained*, à l'inverse de celle observée pour les *stop_gained* avant ré-évaluation (Fig. 5-A et B). Il apparaît également que 8 012 *missense_variant* parmi les 8 581 sont bien prédits cependant 67% (5 135) ne correspondent pas au bon acide aminé. Cette analyse a également été effectuée avec les 246 codons ayant 3 SNP phasés. Les résultats indiquent que 84% (21/25) des *stop_gained* prédits initialement sont des *missense_variant*, et 13 codons indiqués comme *missense_variant* sont en réalité des *stop_gained.*

DISCUSSIONS

Détection de SNP et génotypes par RNA-seq - comparaison avec le DNA-seq.

Détection de SNP : Comme attendu, le nombre de SNP détectés à l'échelle du génome par RNA-seq (~1,4M) est inférieur au DNA-seq (~8M), car seuls les variants présents dans les régions transcrites sont détectées. Le nombre de SNP détectés par RNA-seq à l'échelle des gènes exprimés seulement (exons et introns) passe alors à ~1,05M suggérant que les 350 000 SNP perdus sont probablement situés dans des gènes encore non modélisés. Ceci est corroborée par l'identification continuelle de nouveaux gènes, en particulier des longs non codants, dans les génomes des espèces d'élevage (Jehl et al., 2019b). Afin de fournir un meilleur terrain de comparaison entre les résultats obtenus avec des données DNA-seq 20X et RNA-seq, nous avons limité l'analyse du génome aux seuls exons exprimés, éliminant donc tous les introns

plus ou moins exprimés en RNA-seq ainsi que d'éventuels exons non exprimés. Le nombre de SNP identifiés par RNA-seq passe alors de 1,05M à ~220 000 suggérant que 80% des SNP captés sont situés dans les introns, connus pour être beaucoup plus polymorphes que les régions codantes. En effet, la différence de pression de sélection entre les régions codant les protéines et les régions non codantes dont les introns peut expliquer ces observations. Une altération de la protéine est *a priori* plus délétère qu'une modification dans les régions non codantes du génome, même si le rôle régulateur de certaines d'entre elles émerge (Barrett et al., 2012).

Nous avons ensuite caractérisé plus en profondeur, dans ces régions exoniques exprimées, les SNP communs aux 2 technologies (~200 000) ainsi que les SNP minoritaires spécifiques au DNA-seq et RNA-seq (~35 000 et ~20 000 respectivement). Nous montrons que 85% des SNP détectés en DNA-seq 20X sont retrouvés en RNA-seq, montrant une bonne fiabilité du RNA-seq; des résultats similaires (82%) ont été rapportés par Adetunji et al, 2019. Notons néanmoins que ce pourcentage dépend de l'effort de séquençage fait en RNA-seq, comme le montrent nos résultats sur la population FLLL où seulement 72% des données DNA-seq sont retrouvés en RNA-seq; ceci est probablement dû à la longueur moindre des *reads* de 100 pb contre 150 pb pour la population RpRm. En revanche, nous montrons pour les deux populations ici étudiées RpRm et FLLL que plus de 90% des SNP détectés par le RNA-seq le sont en DNA-seq montrant ainsi une bonne fiabilité. Par ailleurs, on attend une sensibilité accrue avec le nombre de tissus analysés puisqu'une augmentation systématique du nombre de SNP est observée lorsque plusieurs tissus sont analysés. Cette augmentation est en partie due à l'ajout de gènes qui sont exprimés de façon tissu-spécifique. Ceci est cohérent avec des travaux récemment menés au laboratoire sur 25 tissus de l'espèce poule montrant 10% et 25% de gènes tissu-spécifiques respectivement pour les gènes codants des protéines et gènes longs non codants (Jehl et al., 2019a). Concernant les SNP détectés uniquement par une des méthodes, quatre caractéristiques ont retenu notre attention : *i*) Une très forte proportion (40% à 50%) de SNP répondant au filtre GATK de type *SnpCluster* que ce soit pour les SNP spécifiques au RNA-seq ou au DNA-seq. Au vu de ces résultats, étant donné que ce filtre n'est pas recommandé pour le DNA-seq et que les filtres GATK pour le RNA-seq sont déclarés par leurs auteurs comme "devant être validés par les utilisateurs" (GATK, 2020), nous avons décidé de ne pas supprimer ces SNP de notre ensemble de données. *ii*) Un fort excès de SNP dans les jonctions exons-exons dû probablement à la difficulté d'aligner les *reads* qui chevauchent deux séquences plus ou moins éloignées dans le génome pouvant provoquer des SNP erronés (Baruzzo et al., 2017), ce type de *reads* des transcrits qui sont très souvent multi-exoniques et épissés n'existent pas en DNA-seq. *iii*) Un fort déficit de SNP dans les régions 3'UTR probablement dû à la dégradation des transcrits matures par l'action de nucléases agissant en extrémité 3'

(Houseley et al., 2009) rendant impossible leur séquençage par RNA-seq. *iv*) une plus faible couverture en *reads* des SNP détectés uniquement dans une des méthodes comparé aux SNP détectés par les 2 méthodes, sachant que pour le RNA-seq, le nombre de *reads* par SNP est parfois de l'ordre de quelques unités pour l'ensemble de la population (5% des SNP sont couverts par entre 1 et 5 *reads*, 1er quartile à 22 *reads*) ; cela laisse supposer une fraction de SNP faux positifs due à cette faible profondeur. Enfin, un certain nombre de SNP spécifiques du RNA-seq sont probablement dû à au processus biologique de l'édition de l'ARN, processus qui modifie la séquence de l'ARN au niveau d'un nucléotide alors que cette position dans l'ADN est homozygote. Cependant des travaux conduits chez différentes espèces dont la souris (Lagarrigue et al., 2013a) ou encore la poule (Roux et al., 2015, Frésard et al., 2015) ont montré que ce processus reste rare à l'échelle du génome. Quant aux SNP DNA-spécifiques, i.e. non retrouvés en RNA-seq, ils peuvent être en partie dus à une expression allèle spécifique (*ASE* en anglais pour *Allele Specific Expression*) extrême, également relativement rare (Aguet et al., 2019) : cas où une des deux copies alléliques est transcrite alors que l'autre ne l'est pas, conduisant à une absence de polymorphisme au niveau ARN alors que l'ADN est polymorphe à la position.

En conclusion, le RNA-seq permet de détecter plus de 85 % des SNP captés en DNA-seq dans les régions exprimées et probablement plus avec plusieurs tissus analysés et cela avec une concordance avec l'ADN de plus de 90%. Notons que de plus en plus d'études centrées sur l'analyse de transcriptomes tissulaires sont réalisées sur plusieurs tissus, du fait de la diminution des coûts du RNA-seq, entraînant l'accumulation de données au sein d'une même population. Dans les espèces d'élevage où aucune technologie de séquençage exomique (WES - *Whole Exome Sequencing*), à plus bas coût que le DNA-seq "plein génome", n'a été développée, la détection de variants par RNA-seq dont les données s'accumulent, fournirait donc une alternative intéressante pour explorer les SNP à fort impact fonctionnel, en particulier les SNP situés dans les régions codantes pouvant provoquer des pertes de fonctions en induisant un dysfonctionnement de la protéine associée, ou encore des SNP dans les régions 3'UTR pouvant être impliqués dans la régulation du niveaux des transcrits.

Appliquée à 10 populations de poulets, nous avons ainsi obtenu un total de 9 949 072 SNP dont 24% encore inconnus en référence aux 23,8 M de SNP de la base donnée de référence "dbSNP" de *Ensembl* (v92 - octobre 2018, liste toujours à 23M en 2020 version v100). Avant de les caractériser en termes d'impacts fonctionnels, nous avons d'abord utilisé ces 9,9M de SNP pour faire une première étude exploratoire de la génétique de ces 10 populations

Caractérisation génétique de 10 populations de poules à partir des SNP de RNA-seq

L'analyse des données de RNA-seq issus d'un même tissu, le foie, permet de comparer entre populations le nombre de SNP détectés comme polymorphes (cf table 1), i.e. où au moins un nucléotide observé dans la population est différent de celui indiqué dans le génome de référence, peu importe sa fréquence. Ce type d'analyse est souvent menée pour fournir une première caractérisation de populations avec des données de DNA-seq, ces données étant rarement générées sur un nombre suffisant d'individus pour pouvoir calculer des fréquences alléliques (e.g. projet 1000 genome Poule, Tixier-Boichard et al., 2019). Ainsi le nombre de SNP détectés par population à partir de RNA-seq de foie est de l'ordre de 1,8M, avec des variations entre populations entre 1,2M à 1,8M de SNP excepté pour la population commerciale Cobb qui atteint 4,1M. Cette grande différence entre Cobb et les autres populations peut s'expliquer par la nature pyramidale des schémas de sélection en volailles. Sans entrer dans les détails, ces derniers sont caractérisés par 3 étages d'élevages allant de la sélection à la production. Les animaux dits "terminaux" utilisés pour la production sont tous des individus "croisées", mélange de 3 à 4 lignées sélectionnées en amont. La population Cobb est la seule des 10 populations ici étudiées, correspondant à de tels individus "terminaux" pouvant donc expliquer son haut degré de polymorphismes. Notons que cette population Cobb est toujours en première position en termes de SNP détectés sans ou avec le filtre "génotypes disponibles" et présente la proportion d'allèles alternatifs fixés la plus faible (7%). Cela s'explique par son origine génétique et la taille élevée de la population dont elle est issue. A l'autre extrême, la Red Jungle fowl, considéré comme la race ancêtre de toutes les populations de poules domestiquées et que l'on aurait pu penser comme étant très polymorphe car jamais sélectionné ou domestiqué, semble pourtant être la moins polymorphe au sens de la définition donnée plus haut (pour rappel table 1, résultats mono-tissu "foie"); ceci peut s'expliquer par le fait que le génome de référence de l'espèce poule a été déterminé à partir du génome d'un individu de cette population, limitant donc la détection de variation pour cette population. Parmi les 9 949 072 SNP détectés sur l'ensemble des 10 populations, nous avons observé une intersection de 85 127 SNP avec au moins un allèle alternatif dans chaque population qui pourra être utilisé pour caractériser plus finement la génétique de ces populations notamment en recherchant les traces de sélection comme déjà réalisés par certains auteurs dans d'autres espèces (Li et al., 2019, Qanbari et al., 2019). Notons que pour ces SNP, les fréquences alléliques peuvent êtres très variables d'une population à l'autre avec des allèles qui peuvent être rares, fréquents voire fixés. De plus, les 2 017 SNP ayant des fréquences alléliques supérieures à 10% dans toutes les populations analysées pourraient servir à l'enrichissement des puces de génotypage de moyenne densité (Groenen et al., 2011) utilisées pour des analyses d'associations "génotype-phénotype".

Nous avons ensuite analysé les liens génétiques entre ces 10 populations en utilisant une des listes de SNP polymorphes dans chacune des populations (i.e. ayant au moins un allèle ALT parmi les individus de chaque population) ; nous montrons que la population Red Jungle Fowl (RJFh) se détache nettement de toutes les autres populations, quelle que soit la méthode - CAH ou ACP - utilisée. En effet, cette population est considérée comme la population ancestrale, à l'origine de toutes les races domestiquées depuis 6 000 ans (Miao et al., 2013). Elle a d'ailleurs été choisie pour la production de la séquence de référence du génome de la poule. Nous retrouvons ensuite bien le groupe de populations pondeuses séparé du groupe de populations de poulets de chair. Cette dichotomie s'explique par des générations de sélection sur des objectifs très distincts entre les deux filières viande et oeuf qui ont de fait une base génétique différente : grossièrement croissance d'un côté et nombre d'oeufs de l'autre. Nous pouvons également distinguer au sein de ces deux grands groupes génétiques les lignées RpRm et lignées Novogen, toutes deux étant des lignées de pondeuse de même base génétique "la Rhode Red Island" à plumage et oeuf roux mais qui ont une histoire de sélection différente puisque les A3A3 et N4A3 sont des lignées commerciale "parentales" en sélection alors que la RpRm est devenue une population expérimentale fermée dans les années 80 qui a fait l'objet d'une sélection divergente sur l'efficacité alimentaire et qui est à l'origine des 2 sous populations Rp et Rm que nous distinguons également par CAH mais jamais par l'un des 4 axes de l'ACP, les différences devant être plus ténues. La même configuration se retrouve pour la lignée commerciale Cobb et la population FLLL devenue expérimentale dans les années 80 et qui a fait l'objet d'une sélection divergente sur l'adiposité corporelle étant à l'origine des 2 sous populations FL et LL que nous distinguons par CAH. Enfin nous observons les populations Fayoumi et FrAg, qui sont respectivement une population égyptienne ancienne et robuste (i.e. assez résistante aux maladies) et une population Leghorn de pondeuses à plumage et oeufs blancs. Nous remarquons que la lignée de ponte Leghorn n'est pas plus proche des autres lignées de pondeuses que des autres populations indiquant que les 2 rameaux de pondeuse à oeufs roux et oeufs blancs sont assez anciens. L'ensemble de ces analyses montrent donc des résultats en cohérence avec l'histoire de ces populations.

Prédictions des conséquences fonctionnelles associées aux SNP détectés

Comme attendu, les annotations fonctionnelles sont en majorité des variants silencieux (*synonymous_variant*) et faux sens (*missense_variant)* qui respectivement, ne sont pas ou peu soumis à la pression de sélection (Hunt et al, 2009). Les variants *a priori* hautement impactant comme les non-sens (*stop_gained*) sont en revanche très peu nombreux, représentant ici moins de 0,01% (1746/16,5M) des conséquences prédites pour l'ensemble des 9,9M de SNP détectés pour les 10 populations. De façon intéressante, nous montrons que les variants

synonymous_variant et *missense_variant* se répartissent uniformément dans la séquence codante alors que les SNP *stop_gained* sont plus fréquents en début et en fin de séquence codante sans pour autant se l'expliquer. En effet, nous aurions attendu un excès de ces codons STOP en fin de séquences codantes, étant a priori, moins impactant à cette position sur la fonction de la protéine puisque cette dernière est moins tronquée. De plus, les transcrits avec codons STOP prématurés sont en général dégradés par le système NMD *(Nonsense-Mediated mRNA Decay)* (Kurosaki et al., 2019). Notre priorité dans un futur proche sera d'étudier la répartition des *stop_gained* en fonction des génotypes des individus afin de savoir s'il n'y a pas un excès d'hétérozygotes, ce qui pourrait fournir une explication à cette distribution.

Nous avons dans un second temps analysé les doubles ou triples SNP phasés et situés dans un même codon. A notre connaissance, deux études se sont intéressées à la prise en compte des variants présents au sein d'un même codon (Vergara et al., 2012, Cheng et al., 2017), cependant elles ne considéraient pas la phase. Dans le cadre de notre étude, cela concerne peu de codons, 9178 pour les doubles SNP phasés et 246 pour les triples soit moins de 0,2% des 9,9 M de SNP totaux détectés. Nous montrons que 96% des 346 SNP en doublet dans des codons *stop_gained* sont des faux positifs et ont une distribution uniforme de leur position le long de la séquence codante. A l'inverse, 177 codons *missense_variant* deviennent des codons *stop* et présentent comme les autres codons *stop* un excès de position en début et fin de séquence codante. Notons que 67% de ces double SNP *missense_variant* phasés ont une prédiction erronée concernant l'acide aminé modifié. En résumé, l'ensemble de ces résultats permet d'évaluer pour la première fois, sur un nombre conséquent de SNP issus de multiples populations, l'incidence des doubles ou triples SNP phasés et situés dans un même codon. Cette étude révèle également une distribution très particulière et inattendue des positions des SNP de type *stop_gained* le long de la séquence codante, ce qui demande à être approfondie.

Par ailleurs, cette étape d'annotation appliquée aux 9,9M de variants des 10 populations a ainsi révélé au total 1590 SNP *stop_gained*. Nous proposons dans un avenir proche d'analyser les gènes impactés par l'ensemble de ces SNP, en particulier ceux positionnés dans la première moitié de la séquence codante associée, ayant une fréquence allélique non négligeable (>= 10%) dans au moins une population en parallèle de fréquences différentielles entre populations. Un tel SNP pourrait alors jouer un rôle en lien avec une variation phénotypique entre populations.

En conclusion, ce travail montre que les données issues du séquençage ARN sont utilisables pour accéder aux variants, dans les régions exprimées du génome. Ces variants sont en nombre suffisant dans le cadre d'études de diversité génétique ou d'association génotype-phénotype. Cette étude permet de mettre également en lumière l'importance de la phase entre SNP proches, dans le cadre de la prédiction de conséquences dans les régions codant les protéines.

RÉFÉRENCES

Adetunji, M.O., Lamont, S.J., Abasht, B., and Schmidt, C.J. (2019). Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLOS ONE 14, e0216838.

Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. BioRxiv 787903.

Auwera, G.A.V. der, Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43, 11.10.1-11.10.33.

Barrett, L.W., Fletcher, S., and Wilton, S.D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci. 69, 3613–3634.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods 14, 135–139.

Brown, W.R.A., Hubbard, S.J., Tickle, C., and Wilson, S.A. (2003). The chicken as a model for large-scale analysis of vertebrate gene function. Nature Reviews Genetics 4, 87–98.

Cheng, S.-J., Shi, F.-Y., Liu, H., Ding, Y., Jiang, S., Liang, N., and Gao, G. (2017). Accurately annotate compound effects of genetic variants using a context-sensitive framework. Nucleic Acids Res 45, e82–e82.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6, 80–92.

Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., Sun, H., Li, T., Zhang, J., Qiu, X., et al. (2018). VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics 19, 135.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

FAO - Food and Agriculture Organization of the United Nations (2020). Chickens | Gateway to poultry production and products. Consulté à l'adresse : http://www.fao.org/poultry-production-products/production/poultry-species/chickens/en/

Frésard, L., Leroux, S., Roux, P.-F., Klopp, C., Fabre, S., Esquerré, D., Dehais, P., Djari, A., Gourichon, D., Lagarrigue, S., et al. (2015). Genome-Wide Characterization of RNA Editing in Chicken Embryos Reveals Common Features among Vertebrates. PLoS One 10.

GATK (2017). Calling variants in RNAseq. Consulté à l'adresse https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq

GATK (2020). Hard-filtering germline short variants. Consulté à l'adresse https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

Goede, O.M. de, Ferraro, N.M., Nachun, D.C., Rao, A.S., Aguet, F., Barbeira, A.N., Castel, S.E., Kim-Hellmuth, S., Park, Y., Scott, A.J., et al. (2019). Long non-coding RNA gene regulation and trait associations across human tissues. BioRxiv 793091.

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., Gilbert, H., and Louveau, I. (2017). A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics 18, 244.

Groenen, M.A., Megens, H.-J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M., and Cheng, H.H. (2011). The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12, 274.

Guo, Y., Zhao, S., Sheng, Q., Samuels, D.C., and Shyr, Y. (2017). The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. BMC Genomics 18, 690.

Houseley, J., and Tollervey, D. (2009). The Many Pathways of RNA Degradation. Cell *136*, 763–776.

Hunt, R., Sauna, Z.E., Ambudkar, S.V., Gottesman, M.M., and Kimchi-Sarfaty, C. (2009). Silent (Synonymous) SNPs: Should We Care About Them? In Single Nucleotide Polymorphisms: Methods and Protocols, A.A. Komar, ed. (Totowa, NJ: Humana Press), pp. 23–39.

Jehl, F., Désert, C., Klopp, C., Brenet, M., Rau, A., Leroux, S., Boutin, M., Lagoutte, L., Muret, K., Blum, Y., et al. (2019a). Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics 20.

Jehl, F., Muret, K., Bernard, M., Esquerre, D., Acloque, H., Giuffra, E., Djebali, S., Foissac, S., Derrien, T., Zerjal, T., Klopp, C., Lagarrigue, S. (2019b). An atlas of chicken long non-coding RNAs gathering multiple sources : gene models and expression across more than twenty tissues. Presented at PAG XXVII - Plant & Animal Genome Conference, San Diego, USA (2019-01-12 - 2019-01-16).

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Krueger, F. (2020). GitHub : TrimGalore. Consulté à l'adresse : https://github.com/FelixKrueger/TrimGalore

Kurosaki, T., Myers, J.R., and Maquat, L.E. (2019). Defining nonsense-mediated mRNA decay intermediates in human cells. Methods 155, 68–76.

Lagarrigue, S., Hormozdiari, F., Martin, L.J., Lecerf, F., Hasin, Y., Rau, C., Hagopian, R., Xiao, Y., Yan, J., Drake, T.A., et al. (2013a). Limited RNA editing in exons of mouse liver and adipose. Genetics 193, 1107–1115.

Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.-F., Pan, C., van Nas, A., Demeure, O., Cantor, R., Ghazalpour, A., Eskin, E., et al. (2013b). Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics 195, 1157–1166.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323.

Li, D., Li, Y., Li, M., Che, T., Tian, S., Chen, B., Zhou, X., Zhang, G., Gaur, U., Luo, M., et al. (2019). Population genomics identifies patterns of genetic diversity and selection in chicken. BMC Genomics 20, 263.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology 17, 122.

Miao, Y.-W., Peng, M.-S., Wu, G.-S., Ouyang, Y.-N., Yang, Z.-Y., Yu, N., Liang, J.-P., Pianchou, G., Beja-Pereira, A., Mitra, B., et al. (2013). Chicken domestication: an updated perspective based on mitochondrial genomes. Heredity 110, 277–282.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 621–628.

Mou, C., Pitel, F., Gourichon, D., Vignoles, F., Tzika, A., Tato, P., Yu, L., Burt, D.W., Bed'hom, B., Tixier-Boichard, M., et al. (2011). Cryptic Patterning of Avian Skin Confers a Developmental Facility for Loss of Neck Feathering. PLoS Biol 9.

Muret, K., Klopp, C., Wucher, V., Esquerré, D., Legeai, F., Lecerf, F., Désert, C., Boutin, M., Jehl, F., Acloque, H., et al. (2017). Long noncoding RNA repertoire in chicken liver and adipose tissue. Genet Sel Evol 49.

Oikkonen, L., and Lise, S. (2017). Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome Open Res 2.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413–1415.

Pickrell, J.K. (2014). Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. The American Journal of Human Genetics *94*, 559–573.

Piskol, R., Ramaswami, G., and Li, J.B. (2013). Reliable Identification of Genomic Variants from RNA-Seq Data. The American Journal of Human Genetics 93, 641–651.

Qanbari, S., Rubin, C.-J., Maqbool, K., Weigend, S., Weigend, A., Geibel, J., Kerje, S., Wurmser, C., Peterson, A.T., Jr, I.L.B., et al. (2019). Genetics of adaptation in modern chicken. PLOS Genetics 15, e1007989.

Quinn, E.M., Cormican, P., Kenny, E.M., Hill, M., Anney, R., Gill, M., Corvin, A.P., and Morris, D.W. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLoS One 8.

Roux, P.-F., Frésard, L., Boutin, M., Leroux, S., Klopp, C., Djari, A., Esquerré, D., Martin, P.G.P., Zerjal, T., Gourichon, D., et al. (2015). The Extent of mRNA Editing Is Limited in Chicken Liver and Adipose, but Impacted by Tissular Context, Genotype, Age, and Feeding as Exemplified with a Conserved Edited Site in COG3. G3 (Bethesda) 6, 321–335.

Savary, C., Kim, A., Lespagnol, A., Gandemer, V., Pellier, I., Andrieu, C., Pagès, G., Galibert, M.-D., Blum, Y., and de Tayrac, M. (2020). Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. Scientific Reports 10, 1224.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–132.

Tang, X., Baheti, S., Shameer, K., Thompson, K.J., Wills, Q., Niu, N., Holcomb, I.N., Boutet, S.C., Ramakrishnan, R., Kachergus, J.M., et al. (2014). The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Res. 42, e172.

Tixier-Boichard, M., Lecerf, F., Bardou, P., Klopp, C. (2019). A French Pilot Project to Test the Concept of a 1000 Gallus Genomes Initiative. Presented at PAG XXVII - Plant & Animal Genome Conference, San Diego, USA (2019-01-12 - 2019-01-16).

Vergara, I.A., Frech, C., and Chen, N. (2012). CooVar: Co-occurring variant analyzer. BMC Res Notes 5, 615.

Wolfien, M., Rimmbach, C., Schmitz, U., Jung, J.J., Krebs, S., Steinhoff, G., David, R., and Wolkenhauer, O. (2016). TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics 17, 21.

Wang, C., Davila, J.I., Baheti, S., Bhagwate, A.V., Wang, X., Kocher, J.-P.A., Slager, S.L., Feldman, A.L., Novak, A.J., Cerhan, J.R., et al. (2014). RVboost: RNA-seq variants prioritization using a boosting method. Bioinformatics 30, 3414–3416.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164–e164.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328.

AGRO CAMPUS OUEST	Diplôme : **Ingénieur Agronome** Spécialité : **Biologie Moléculaire et Cellulaire** Spécialisation / option : Enseignant référent : **Sandrine Lagarrigue**

Auteur(s) : **DEGALEZ Fabien** Date de naissance* : **09/05/1997**	Organisme d'accueil : **INRAE – PEGASE – UMR 1348** Adresse : **16 ,Le Clos, Domaine de La Prise, 35590 Saint-Gilles**
Nb pages : **18** Annexe(s) : **0**	Maître de stage : **Sandrine Lagarrigue**
Année de soutenance : **2020**	

Titre français :

Potentiels du séquençage des ARN pour explorer les micro-variations du génome

Titre anglais :

Potential of RNA sequencing to explore micro-variations of the genome

Résumé (1600 caractères maximum) :

Dans ce stage, nous avons finalisé un ensemble de programmes informatiques permettant la détection de variants mononucléotidiques (SNP) fiables à partir de données de séquençage d'ARN (RNA-seq) et les avons appliqués sur des données RNA-seq de 10 populations de poules commerciales et expérimentales de chair et de ponte. Une analyse de la concordance des variants obtenus en RNA-seq et en DNA-seq 20X (séquençage de l'ADN génomique des même tissus) a été réalisée, pour la première fois, sur les mêmes individus et de surcroît en nombre important (15 poules). Nous montrons ainsi que les données de RNA-seq sont une ressource de polymorphismes intéressante à exploiter car, à régions exprimées égales, le RNA-seq permet de détecter plus de 85% des SNP captés en DNA-seq - plus encore avec plusieurs tissus analysés - et cela avec une concordance avec l'ADN de plus de 90%. D'autre part le nombre de SNP détectés est conséquent : 9,9M de SNP détectés pour l'ensemble des 10 populations avec en moyenne 1,8M de SNP détectés par population dont ~0,5M avec des génotypes renseignés. Finalement, 250 000 SNP avec génotypes renseignés sont communs aux 10 populations étudiées. Cette liste de SNP a permis une première caractérisation des liens génétiques entre ces 10 populations, qui est cohérente avec l'histoire phylogénétique de ces dernières. Nous avons également analysé l'impact fonctionnel de ces 9,9M de SNP sur les transcrits et protéines associés et avons identifié 1590 SNP *stop_gained* qui nous restent à analyser. Nous avons également développé un programme permettant de prédire l'impact fonctionnel de double ou triple SNP phasés au sein d'un même codon, donnée non prise en compte encore dans les programmes actuels. Bien que rares (concernent 0,2% des variants), nous montrons qu'environ 70% des prédictions fonctionnelles sur ces codons à double ou triple SNP phasés sont erronées.

Abstract (1600 caractères maximum) :

In this internship, we finalized a pipeline for the detection of reliable mononucleotide variants (SNP) from RNA sequencing data (RNA-seq) and applied it to RNA-seq data from 10 commercial and experimental broiler and egg-laying hen populations. A concordance analysis of the variants obtained in RNA-seq and DNA-seq 20X (genomic DNA sequencing of the same tissues) was carried out, for the first time, on the same individuals and in significant numbers (15 hens). We thus show that RNA-seq data are an interesting polymorphism resource to exploit because, at equal expressed regions, RNA-seq allows the detection of more than 85% of the SNP captured in DNA-seq - even more with several tissues analysed - and this with a DNA match of more than 90%. On the other hand, the number of SNP detected is consistent: 9.9M SNP detected for all 10 populations with an average of 1.8M SNP detected per population, of which ~0.5M with informed genotypes. Finally, 250,000 SNP with informed genotypes are common to the 10 populations studied. This list of SNP allowed a first characterization of the genetic links between these 10 populations, which is coherent with their phylogenetic background. We also analyzed the functional impact of these 9.9M SNP on transcripts and associated proteins and identified 1590 stop_gained SNP that remain to be analyzed. We have also developed a program to predict the functional impact of double or triple phased SNP within the same codon, data not yet taken into account in current programs. Although rare (concerning 0.2% of variants), we show that about 70% of the functional predictions on these double or triple phased SNP codons are erroneous.

Mots-clés : poules, séquençage ARN, séquençage ADN, polymorphisme, SNP, génotype, génétiques, phylogénie, prédictions de conséquences, phase, codon

Key Words : hens, RNA sequencing, DNA sequencing, polymorphism, SNP, genotype, genetic, phylogeny, prediction of consequences, phase, codon

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