



HAL
open science

Détection et caractérisation de variants structuraux dans le génome d'une lignée de poules pondeuse

Morgane Gaudin

► **To cite this version:**

Morgane Gaudin. Détection et caractérisation de variants structuraux dans le génome d'une lignée de poules pondeuse. Sciences du Vivant [q-bio]. 2018. dumas-01961544

HAL Id: dumas-01961544

<https://dumas.ccsd.cnrs.fr/dumas-01961544>

Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AGROCAMPUS
OUEST

CFR Angers

CFR Rennes



Année universitaire : 2017 – 2018

Spécialité : BMC

Biologie Moléculaire et Cellulaire

Spécialisation (et option éventuelle) :

Mémoire de Fin d'Études

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

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

d'un autre établissement (étudiant arrivé en M2)

Détection et caractérisation de variants structuraux dans le génome d'une lignée de poules pondeuse

Par : Morgane GAUDIN

Soutenu à Rennes le 11/06/2018*

Devant le jury composé de :

- Frédéric Lecerf

- Sébastien Huet

- Denis Tagu

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

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⁾ :

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 :

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

AGROCAMPUS
OUEST

CFR Angers

CFR Rennes



Année universitaire : 2017 – 2018

Spécialité : BMC

Biologie Moléculaire et Cellulaire

Spécialisation (et option éventuelle) :

.....

Mémoire de Fin d'Études

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

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

d'un autre établissement (étudiant arrivé en M2)

Détection et caractérisation de variants structuraux dans le génome d'une lignée de poules pondeuse

Par : Morgane GAUDIN

Soutenu à Rennes le* 11/06/2018

Devant le jury composé de :

- Frédéric Lecerf
- Sébastien Huet
- Denis Tagu

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

Copy number variations (CNV), insertions or deletions of 50 consecutive bases or more, are an important source of genetic diversity, since they affect more bases than single nucleotide polymorphisms (SNP). Therefore, they can have an important impact on phenotypic diversity and even alter traits of economic importance in breeding. They have previously been studied in humans, and have been linked to several diseases. Additionally they have been studied in animals, especially cattle, for which they were connected to several phenotypes of economical interest. Here we performed a genome-wide CNV detection on the genome sequence of 50 males of a line of laying hens (Rhode Island).

The detection was performed using three different tools based on two methods, paired-end mapping and split-read analysis, using three tools combining those two methods, Delly, Pindel and Lumpy. After the detection, the quality of the deletions was checked. We kept only those which were supported by more than three paired-end reads. Moreover another quality check was performed before the association analysis, where variants with a MAF below 5% and a call rate below 95% were removed.

In total, about 50000 deletions were identified, covering about 3% of the genome. Most of those were found in only one sample. They were evenly distributed on the genome. The size of the deletion identified ranged from 1 base to a few megabases. Those over 20kb were removed due to their size. Depending on which tools was used, the size and the distribution of these deletions in the population varied greatly, especially between Pindel and Delly. Interestingly, while the variants detected by Delly were mostly found in only one sample, those found by Pindel were mostly found in three or more samples.

Association analysis were performed for 7 traits: egg weight, shell color, egg shell resistance, deformation, diameter, albumen height and shape index. Those traits were calculated using daughters' performances and are among those used to select these chicken. In total, 23 deletions significantly associated with at least one trait have been identified. Among these, 16 overlap genes.

Older GWAS performed on SNP for the same traits were also used to analyze our variants. Variants in linkage disequilibrium with the significant SNP identified were selected, and their consequences were studied. None of the deletions in linkage disequilibrium with the significant SNP were significant in the association analysis, however, 73 of these overlapped genes, and 40 had a high impact on those genes, such as the loss of a stop codon.

Finally, a list of gene potentially involved in egg shell formation was also available, and the variants overlapping these genes were also further studied. In total about 3000 interesting

deletions were identified, overlapping about 350 of the genes of interest. Additionally, 66 were modified by deletion from the three tools.

The results of the last two analysis were compared, to find whether the deletions in regions significantly associated with a trait were also overlapping genes of interest. It appears to be the case, as we identified 112 deletions overlapping 30 genes. Among these 30 genes are modified from deletions from the three tools. These genes and their functions were further studied.

These results are a promising first approach to CNV in this line. However much remains to be done, especially since the population studied was too small to obtain really significant associations between the genotypes and the phenotypes. Moreover any interest variants we have identified here still needs to be confirmed by molecular biology, as these tools might produce many false positives, as it is usually the case for analysis on NGS data and we have little mean to check these.

Tables des matières

Introduction.....	1
Matériel et méthode.....	3
Population étudiée.....	3
Ressources.....	3
Détection de variants.....	4
Méthodes.....	4
Outils.....	5
Analyse des variants.....	5
Analyse d'association.....	5
Comparaison à des régions d'intérêt identifiées par GWAS.....	6
Comparaison à une liste de gènes impliqués dans la formation de la coquille.....	6
Conséquences fonctionnelles des variants.....	6
Comparaison des résultats.....	6
Résultats.....	8
Détection de variants.....	8
Analyse des variants.....	10
Étude d'association.....	10
Comparaison à des régions d'intérêt identifiées par GWAS.....	12
Comparaison à une liste de gènes impliquées dans la formation de la coquille.....	12
Comparaison des résultats.....	14
Discussion.....	14
Références bibliographiques.....	17

Liste d'abréviations

ADN : acide désoxyribonucléique

CHIA : chitinase acide

CNV : copy number variation

De : déformation

Di : diamètre

FSTL1 : follistatine like 1

GEV ; genomic estimated breeding value

GLM : generalized linear modèle

GWAS : genome wide association study

HA : hauteur d'albumen

Indel : insertion délétion

INRA : institut national de recherche agronomique

LAB : axes de décomposition de la lumière

MAF : minor allele frequency

NGS : next generation sequencing

PEM : paired-end mapping

PO : poids d'oeuf

PTN : pléiotrophine

PTPRF : protein Tyrosine Phosphatase, Receptor Type F

SI : shape index

SNP : Single nucleotide polymorphism

SR : split read

SV : structural variant

VEP : variant effect predictor

VIH : virus de l'immunodéficience humaine

Introduction

Le génotypage des Single Nucleotide Polymorphisms (SNP), polymorphismes de substitution d'une base, est une technique qui s'est rapidement répandue au cours des dernières années, que ce soit à des fins de recherche pour des études d'association entre génotype et caractère (GWAS) ou commerciales pour la sélection de reproducteurs dans les filières agronomiques (sélection génomique).

Cependant ce ne sont pas les seuls variants modifiant le génome. Les variants structuraux (SV) sont des variants affectant des fragments de génome d'une taille supérieure à 50 paires de bases consécutives (Alkan et al., 2011). Ils peuvent être divisés en deux catégories selon qu'ils modifient ou non la taille du génome. Les inversions et les translocations ne modifient pas le nombre de bases du génome, tandis que les insertions, les délétions et les duplications augmentent ou diminuent la taille du génome. Ces événements qui modifient le nombre de copies d'une région sont appelés Copy Number Variants ou CNV.

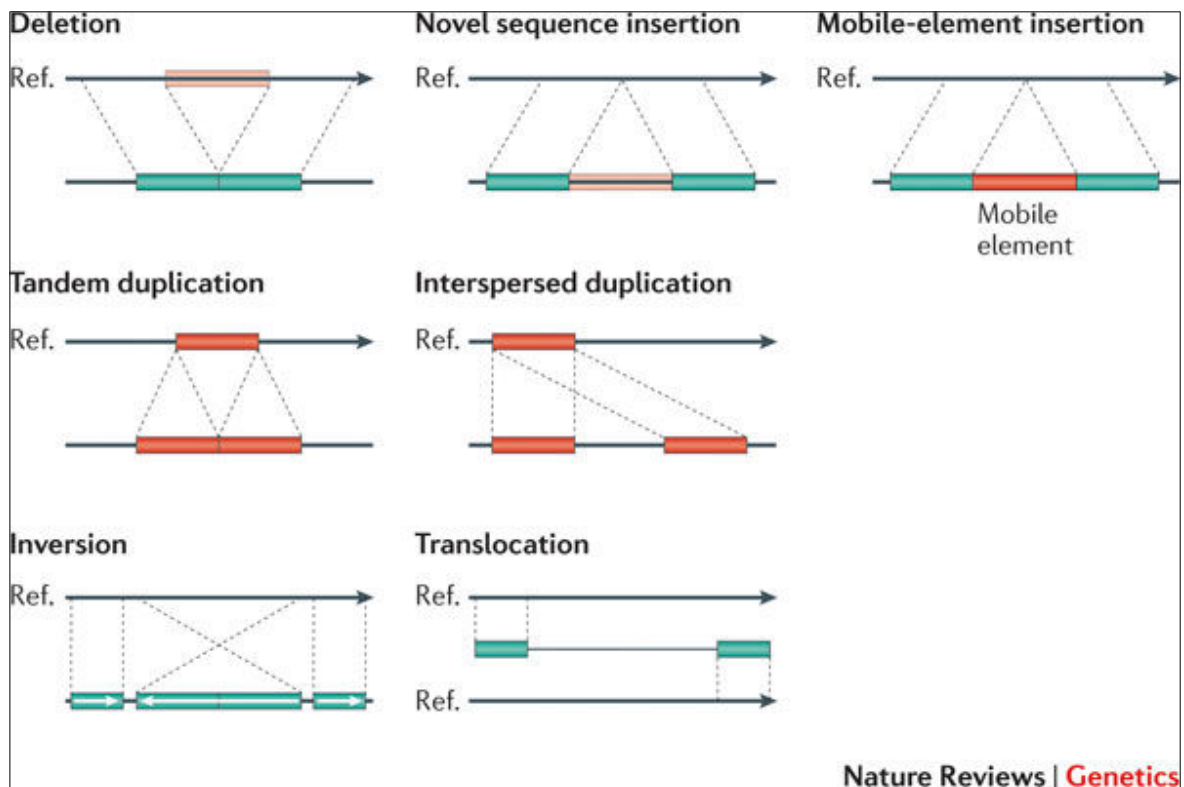


Figure 1: Classes de variations structurales (Alkan et al., 2011)

Les CNV sont moins fréquents que les SNP ou les petites insertions/délétions (InDels). Cependant du fait de leur taille, ils couvrent au total une plus grande partie du génome, et pourraient donc avoir un effet plus important sur les gènes (Bickhart et al., 2012).

Par exemple chez l'humain on estime que 12 % du génome est composé de régions à nombres de copies variables. Des CNV ont été associés à plusieurs maladies génétiques, comme l'autisme (Sebat et al., 2007) ou la schizophrénie (Stone et al., 2008), mais aussi à des résistances aux pathogènes, comme une résistance au VIH (Liu et al., 2010). D'autres phénotypes tels que l'asthme, le psoriasis, la polyarthrite rhumatoïde ont aussi été mis en relation avec des CNV (Ionita-Laza et al., 2009)

Plusieurs études ont également été menées chez les animaux, en particulier chez les animaux d'élevage. La majorité des travaux menés se focalise sur la détection de CNV, mais d'autres études cherchent aussi à identifier les conséquences fonctionnelles des variants identifiés, en particulier en lien avec des caractères d'intérêt agronomiques. Par exemple chez les bovins, des CNV ont été associés à la résistance ou la susceptibilité aux parasites (Hou et al., 2012), la production de lait et la fertilité (Kadri et al., 2014) ou encore à des phénotypes sans cornes (Chen et al., 2017). Ces premiers résultats confirment l'intérêt de l'approfondissement des connaissances sur les CNV chez les animaux d'élevage.

Chez la poule, des phénotypes particuliers tel que le caractère peau noire de la race Poule soie (Nègre soie) (Fan et al., 2013), le phénotype « Pea-comb » affectant le développement de la crête (Moro et al., 2015) ou encore la vitesse d'emplumement et la couleur de plume ont été associés à des CNV (Wang and Byers, 2014).

Les CNV semblent être une source de diversité importante qui est encore peu étudiée. L'objectif de ce travail est dans un premier temps d'identifier les CNV présents dans le génome d'une lignée de poules pondeuses. Dans un second temps les conséquences des délétions identifiées sur des caractères associés à la production et la qualité des œufs seront étudiées par différentes approches.

Matériel et méthode

Population étudiée

Les séquences de 50 coqs d'une lignée de poules pondeuses Rhode Island de l'entreprise Novogen ont été étudiées. Ces coqs font partie d'une population composée de 437 coqs, et ont été génotypés et phénotypés dans le cadre du projet Utopige. De ces 437 individus, 50 ont été choisis pour être séquencés et tous ont été génotypés. Le choix a été fait afin de retenir pour le séquençage les individus représentant la plus grande diversité haplotypique dans la population. Ce choix s'est basé sur les données de génotypage HD disponibles pour tous les individus.

Les séquences ont été obtenues par séquençage en paired-end sur une machine Illumina HiSeq 3000. En séquençage en paired-end, chaque fragment d'ADN est séquencé par les deux extrémités. Les individus ont également été phénotypés pour différents caractères de qualité d'œuf obtenus à partir des performances de leurs filles, dont le poids d'œuf (PO), la couleur de la coquille (LAB), l'indice de déformation (De), le diamètre (Di), la hauteur d'albumen (HA) et la forme de l'œuf (SI). Les valeurs de phénotypes disponibles sont des *genomic estimated breeding values* (GEBV), des valeurs de phénotypes estimées à partir des génotypes. Ces valeurs sont obtenues en associant les génotypes des pères aux phénotypes des filles sur la population entière, ce qui permet par la suite d'estimer la valeur d'un phénotype uniquement grâce à des données de génotypage.

Ressources

La majorité des analyses a été effectuée sur le cluster de la plateforme d'analyse biologique Genotoul à Toulouse (<http://bioinfo.genotoul.fr/>). Un cluster est un regroupement de serveurs, ou de machines, permettant d'effectuer des calculs intensifs à distance.

Détection de variants

Méthodes

Les fichiers de séquences (.bam) ont été nettoyés et indexés à l'aide de samtools (Li, 2011). Les nombreux ADN "scaffolds" présents dans le génome de poule ont été retirés des séquences analysées et du génome de référence, Gallus gallus version 5 (Galgal5).

Il existe plusieurs méthodes de détection des CNV, selon les données disponibles. Il est possible de détecter les CNV en se basant sur le niveau de fluorescence émis lors de génotypages sur puce SNP, via une hybridation comparative génomique (arrayCGH) ou encore sur des données de séquençage (NGS). La détection sur données NGS permet de découvrir un plus grand nombre de CNV, et d'obtenir la localisation précise de leur breakpoints (bases de début et de fin du variant). C'est la méthode qui a été utilisée ici.

La détection de CNV sur données de séquence peut être réalisée de différentes manières selon le type de séquençage utilisé (paired-end ou non) et la profondeur de séquençage. Les deux méthodes utilisées pour cette étude sont le pair end mapping (PEM ou RP) et le split-read (SR). L'approche PEM considère que la distribution de la taille des inserts, c'est-à-dire l'écart entre chaque extrémité séquencée, suit une loi normale. Les SV sont donc détectés en identifiant les paired reads séparés par une distance significativement différente de cette distribution. En SR, les SV sont repérés en identifiant les reads qui ne s'hybrident pas sur toute leur longueur sur le génome de référence. La partie qui ne s'hybride pas est ensuite coupée et réalignée sur le génome de référence, ce qui permet de détecter précisément les breakpoints (bases de début et de fin) du variant identifié (Zhao et al., 2013).

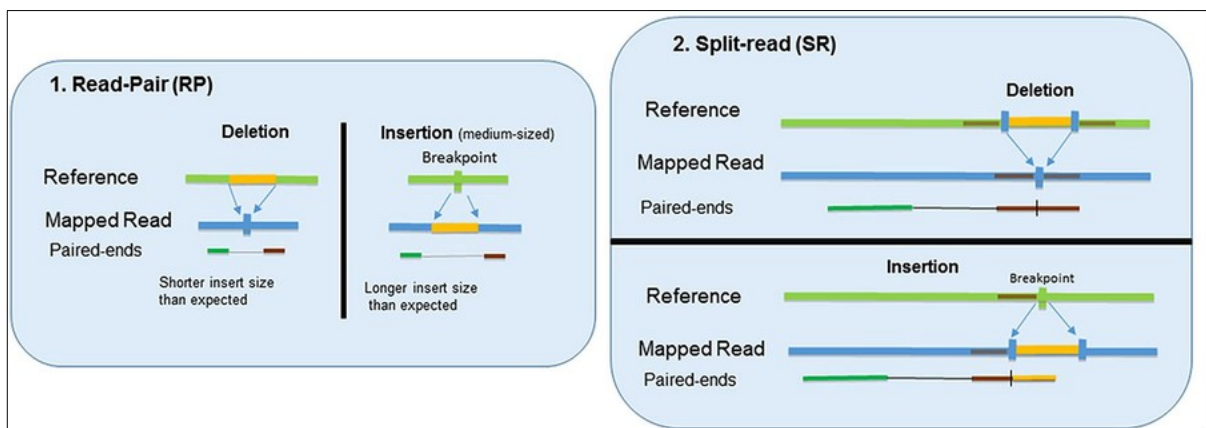


Figure 2: Figure 1: Les deux méthodes de détection de CNV sur données NGS utilisées (Pirooznia et al., 2015)

Outils

La détection de variant a été réalisée à l'aide de trois logiciels : Delly (Rausch et al., 2012), Lumpy (Layer et al., 2014) et Pindel (Ye et al., 2009). Ces trois outils combinent les méthodes de paired-end mapping et split read. Les paramètres par défaut ont été utilisés pour ces trois logiciels. Les fichiers de sortie de ces trois outils sont au format vcf. Ces fichiers contiennent une ligne par variant détectés. Les analyses étant faites individus par individus, 50 fichiers par outils sont produits. Pour chaque variant détecté, les informations de localisation, de taille, de type de variant, de qualité du variant, ainsi que le nombre de reads supportant les variants sont indiquées. Les variants peuvent être notés comme imprécis si aucun split-read n'a permis d'identifier leurs bases de début et de fin. Les variants rapportés par Lumpy et Pindel ne comprennent que les variants supportés par plus de 4 reads, tandis que Delly rapporte tous les variants, en appliquant un flag « LowQual » pour les variants supportés par 3 reads ou moins.

Ces fichiers ont ensuite été génotypés par Svtiper (Chiang et al., 2015) pour chaque individu puis regroupés avec GATK (McKenna et al., 2010) pour au final obtenir un fichier comprenant le génotype des 50 individus par outil. Étant donné que tous les individus ne possèdent pas un variant à chaque position, lors du regroupement des fichiers individuels, de nombreux génotypes ont été marqués comme manquants.

Analyse des variants

Les résultats des trois logiciels utilisés ont été analysés séparément. Ces analyses ont été effectuées sous R.

Analyse d'association

Avant de pouvoir réaliser les analyses d'association, les fichiers de variants ont du être manipulés, pour plusieurs raisons. En premier lieu, le package d'analyse d'association utilisé ne peut pas analyser des variants autres que des SNP. Les délétions ont donc été recodées. Un homozygote pour une délétion (DEL/DEL) a été recodé en A/A, un hétérozygote en A/T et un homozygote pour la référence en T/T. De même, une absence de détection d'une délétion engendrait un génotype manquant (./.) ce qui a été recodé en T/T.

Les analyses d'associations ont été réalisées sous R, à l'aide du package GenABEL (Aulchenko et al., 2007), ce qui a nécessité la conversion des fichiers de sortie au format plink (ped, fam, map). Les génotypes ont été nettoyés pour n'avoir que des variants présentant une

fréquence d'allèle mineur (MAF) supérieure à 5 % et une call rate supérieure à 95 %. Les individus présentant une callrate inférieure à 95 %, une hétérozygotie moyenne supérieure à 1 % et une identity by state (IBS) supérieure à 95 % ont aussi été exclus du reste de l'analyse. La déviation de l'équilibre de Hardy-Weinberg n'a pas été testée, car elle est peu significative sur un échantillon de 50 individus sélectionnés. Un modèle linéaire généralisé (GLM) a été appliqué à ces données nettoyées. Un seuil de significativité $-\log(p\text{value})$ supérieur à 5 a été choisi, et aucune correction pour tests multiples n'a été appliquée aux résultats.

Comparaison à des régions d'intérêt identifiées par GWAS

Des GWAS effectuées précédemment dans la même lignée et sur les mêmes caractères ont permis d'identifier des SNP significativement liés aux caractères (Romé et al., 2015). Afin d'identifier si les délétions détectées sont localisées dans des régions associées à des phénotypes, les variants en déséquilibre de liaison avec ces SNP ont été sélectionnés. Une fenêtre de 20kb autour de chaque SNP a été choisie.

Comparaison à une liste de gènes impliqués dans la formation de la coquille

Une étude sur l'expression différentielle des gènes dans l'oviducte de la poule effectuée par l'INRA de Tour a permis de dresser une liste de gènes potentiellement impliqués dans la formation de la coquille. Afin d'identifier de possibles conséquences fonctionnelles des variants détectés, les délétions co-localisées aux gènes d'intérêt, c'est-à-dire recouvrant partiellement ou totalement ces gènes, ont été sélectionnés. Concrètement, les variants ayant leur première base ou leur dernière base dans le gène ont été retenus.

Conséquences fonctionnelles des variants

Les variants sélectionnés dans ces trois approches ont été analysés avec le Variant Effect Predictor (VEP) (McLaren et al., 2016) afin d'identifier leurs conséquences fonctionnelles sur les gènes.

Comparaison des résultats

Les résultats issus des trois approches utilisées ont été comparés les uns aux autres. Les variants et gènes retrouvés dans deux ou trois de ces approches ont été retenus.

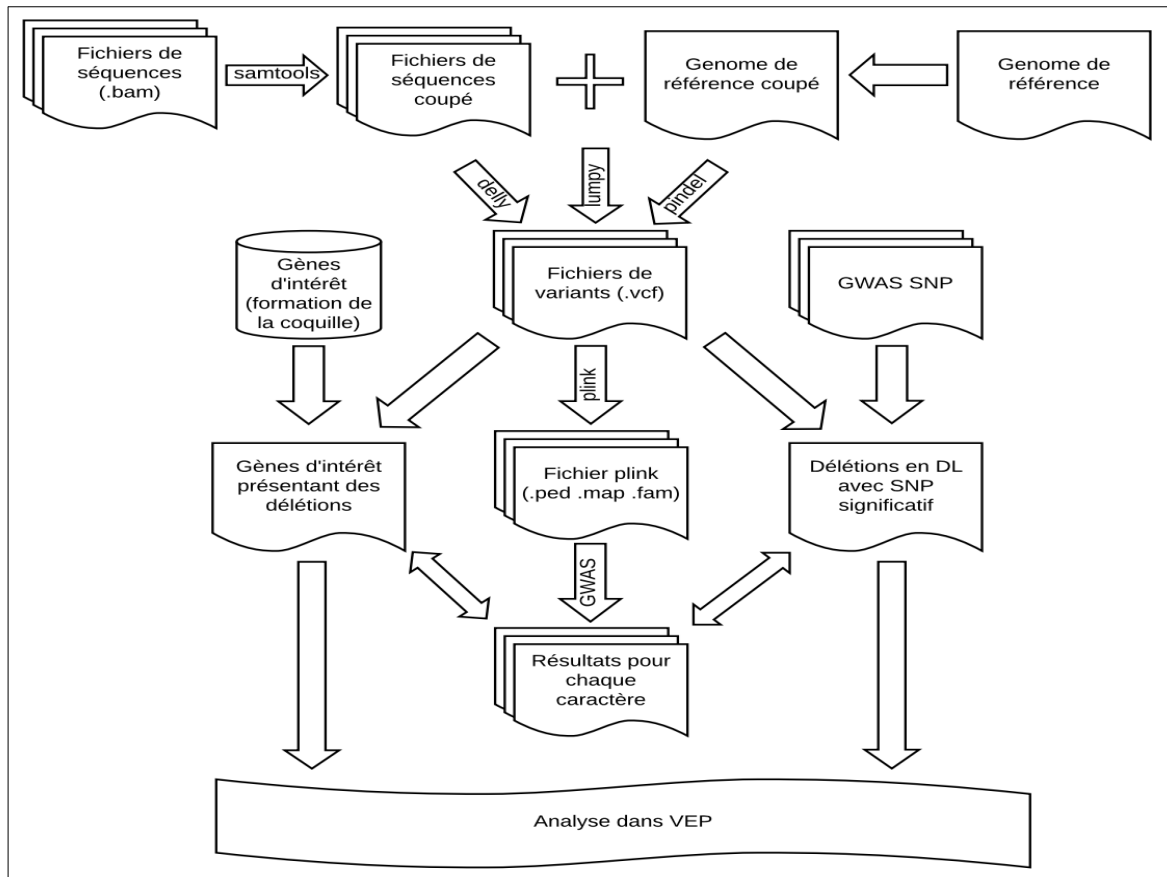


Figure 3: Diagramme des différentes étapes de détection et caractérisation des variants

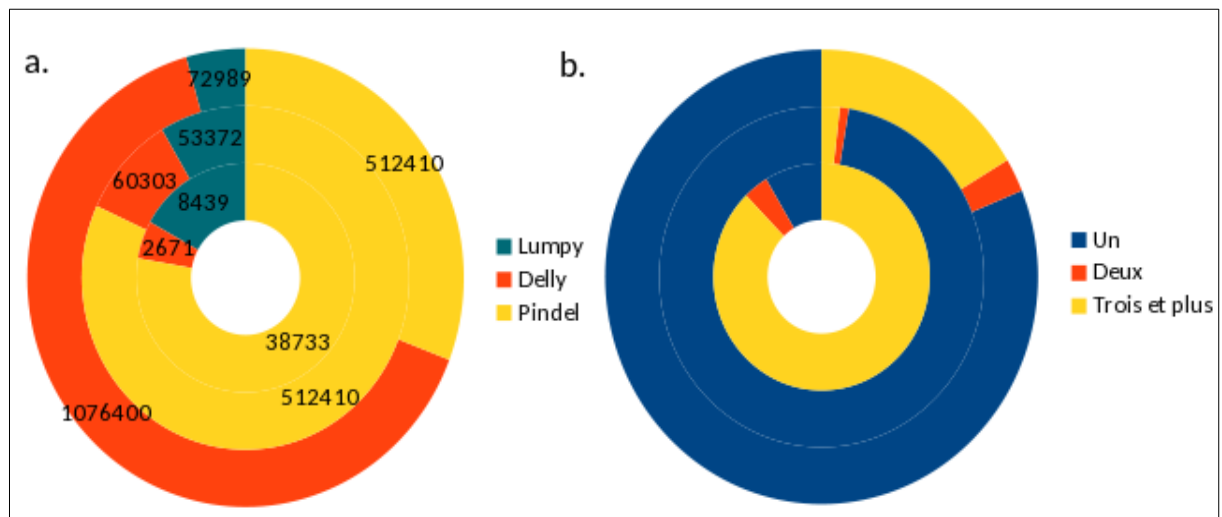


Figure 4: a. Répartition des variants par outil avant filtration (cercle extérieur) et après chacune des étapes de filtration : filtration sur le nombre de reads (>4) validant le variant (cercle du milieu) et filtration sur la MAF, callrate, etc. (cercle intérieur), b. Distribution des variants dans la population. De l'extérieur vers l'intérieur : Lumpy, Delly, Pindel

Résultats

Détection de variants

En tout, 73 989 variants ont été détectés par Lumpy, 1 076 400 par Delly et 512 410 par Pindel. Après un premier contrôle qualité sur la fiabilité des variants (nombre de reads > 4) et une sélection des délétions uniquement, 53 372 variants ont été retenus pour Lumpy, 60 303 pour Delly et 512 410 pour Pindel.

Après le contrôle qualité effectué pour les GWAS (callrate, MAF), 8439 variants ont été conservés pour Lumpy, 2671 pour Delly et 38 733 pour Pindel (figure 4.a). La majorité des variants a été exclue à cause d'une MAF trop faible.

Avant filtration, la plus grande taille de variants obtenus pour Lumpy est de 195 Mb, 183 Mb pour Delly et 16 kb pour Pindel. Cependant, la majorité des variants pour Lumpy mesure entre 200 et 600 bp, entre 400 et 600 bp pour Delly et autour de 1 bp pour Pindel (figure 5).

Un nombre sensiblement plus élevé de variants a été détecté par Pindel par rapport à Delly et Lumpy. Cependant la taille moyenne des variants pour Pindel est de 17 bases, alors qu'elle est d'environ 300 bases pour Lumpy et 900 bases pour Delly. Les variants détectés par Delly et Lumpy sont en majorité des variants retrouvés chez un seul individu, alors qu'au contraire pour Pindel, les variants sont en majorité retrouvés dans au moins 3 individus (figure 4.b). En moyenne, après filtration, les variants sont présents dans 3,6 individus pour Lumpy, 1,2 individus pour Delly et 22,3 individus pour Pindel (figure 6).

Les délétions sont réparties de façon relativement homogène sur le génome (figure 7), et il n'y a pas de chromosomes sur ou sous représentés.

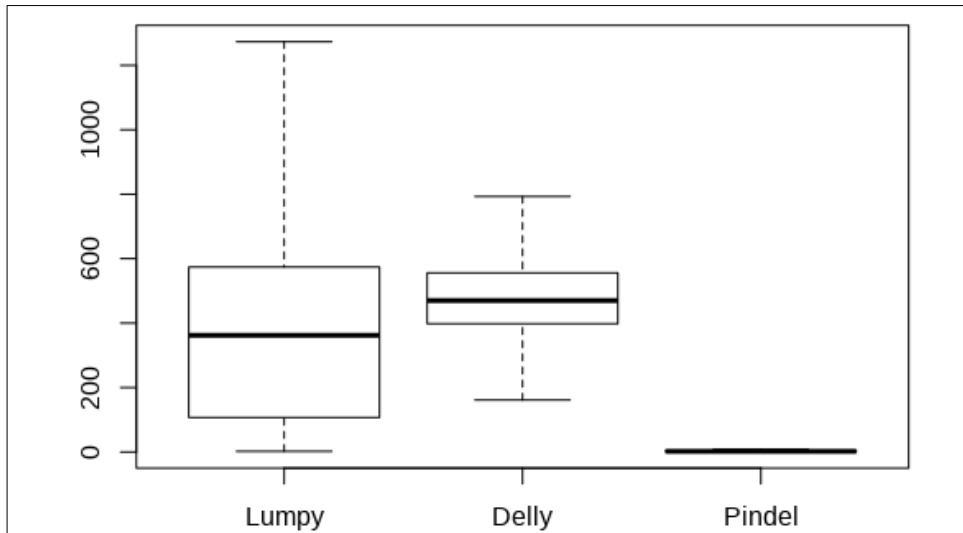


Figure 5: Boxplot de la taille des délétions identifiées

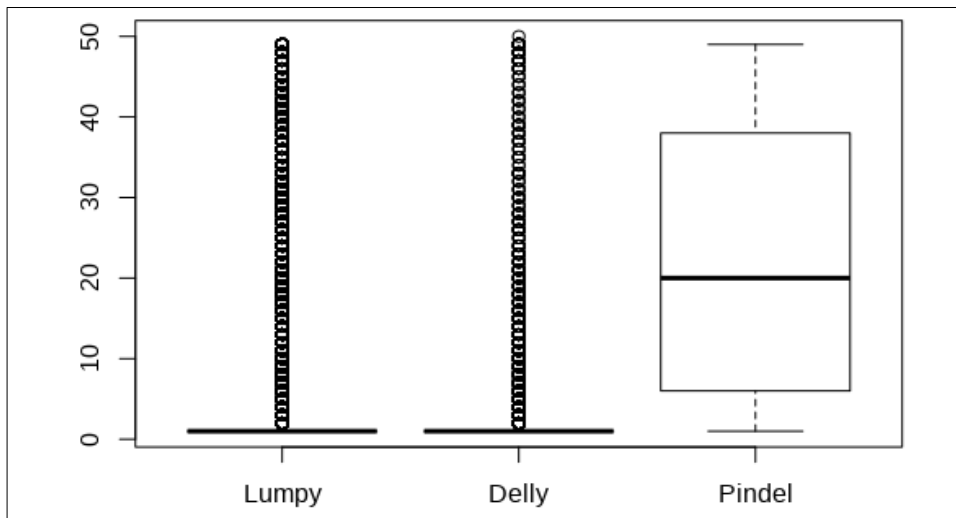


Figure 6: Boxplot du nombre d'individus par variant pour les trois outils

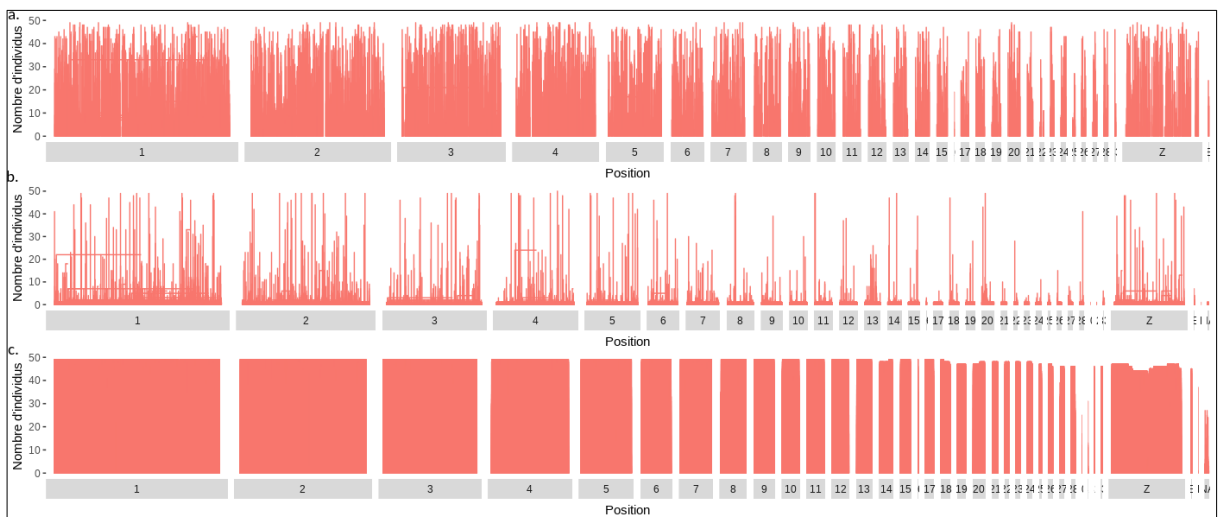


Figure 7: Répartition des variants sur le génome. En abscisse le numéro du chromosome et la position du variant, en ordonnée le nombre d'individus portant ce variant. a. Lumpy, b. Delly, c. Pindel

Analyse des variants

Étude d'association

Au total, 23 variants possédant un effet significatif sur au moins un des caractères (p value > 5) ont été identifiés. Les caractères affectés sont HA, PO, Di, LAB et SI. Six ont été détectés avec Lumpy et 17 avec Pindel. Parmi ceux-ci, 16 sont localisés dans des gènes (tableau 1).

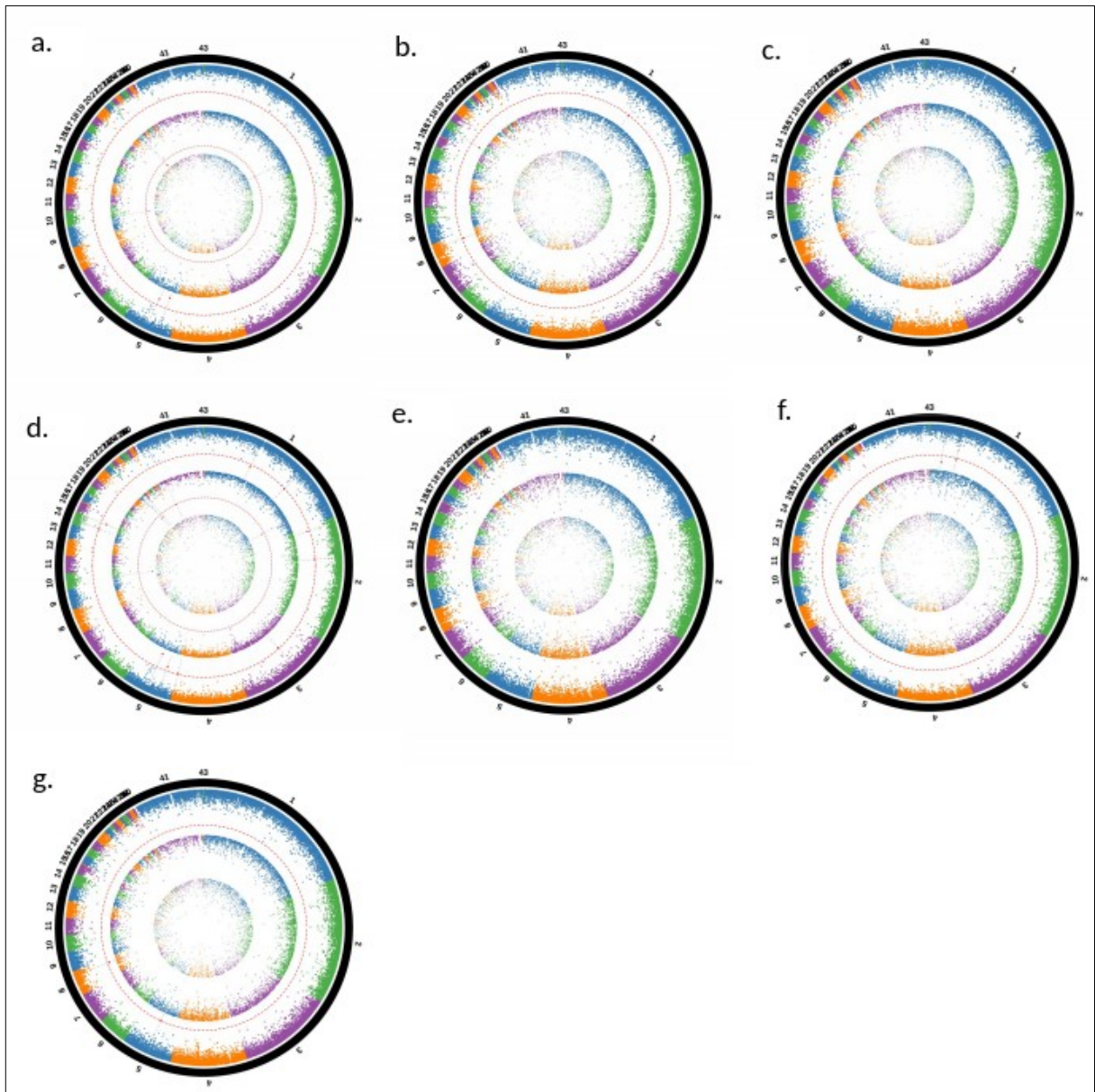


Figure 8: Manhattan plot des analyses d'association par caractères (a. PO, b. LAB, c. FF, d. Di, e. De, f. HA, g. SI). De l'extérieur vers l'intérieur : Pindel, Lumpy, Delly

Tableau 1: Variants significativement associés à un ou plusieurs caractères

Position	Caractère	Outils	Gène	Nb d'ind.
1:18442167-18442211	HA	Pindel	Aucun	21
1:40762472-40762482	HA	Pindel	Aucun	42
1:70899730-70899731	Po et Di	Pindel	WNT7B	49
1:106573172-106573178	LAB	Pindel	Gène inconnu	37
1:130631332-130631338	Di	Pindel	HERC2	49
1:173534780-173534834	Po et Di	Lumpy	Aucun	39
2:5857352-5857353	Di	Pindel	ACVR2B	13
2:48544550-48544554	Di	Pindel	PDE1C	35
3:43259708-43259720	Di	Pindel	Aucun	48
4:89259595-89259602	Di	Pindel	Aucun	39
5:15701341-15701825	Po et Di	Pindel	CD151	19
5:30231346-30231389	SI	Pindel	RYR3	29
5:30722793-30722797	Po et Di	Pindel	FAM98B	49
8:23920802-23920809	LAB	Pindel	FAF1	47
8:25395201-25395204	SI	Pindel	USP24	49
10:2365165-2365238	Po et Di	Lumpy	Gène inconnu	38
14:4224502-4224507	Di	Pindel	RNF216	42
14:14530683-14530691	Di	Pindel	Aucun	28
18:1121599-1121602	LAB	Pindel	DNAH9	49
21:6422114	PO	Lumpy	PPIH	14
21:6523011-6523057	Po et Di	Lumpy	ZNF362	8
23:1199802-1213122	Di	Lumpy	FOXO6	6
Z:10770409-10770432	Di	Lumpy	Aucun	3

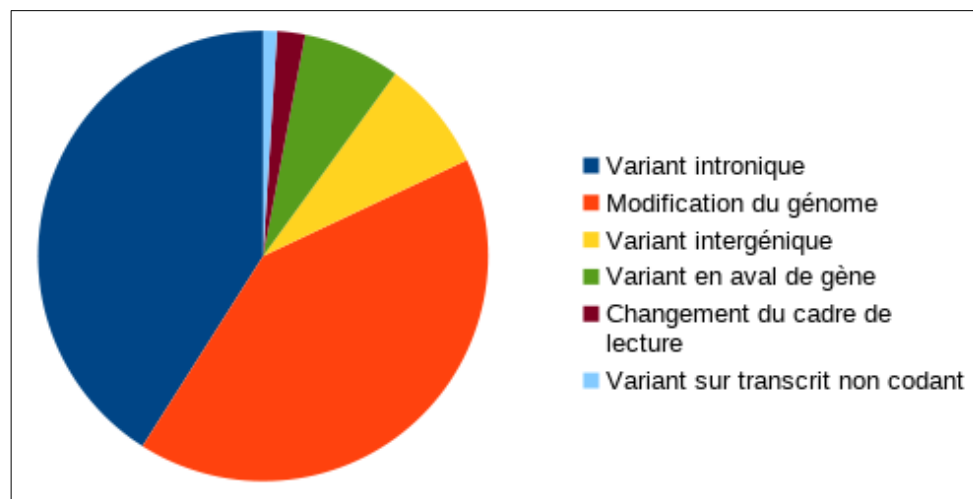


Figure 9: Conséquence sur le génome des variants significativement associés à un phénotype

Comparaison à des régions d'intérêt identifiées par GWAS

Des CNV en déséquilibre de liaison avec des SNP associés aux phénotypes étudiés ont été identifiés, 919 pour Lumpy, 966 pour Delly et 8441 pour Pindel. Cependant aucun de ces CNV ne dépasse le seuil de significativité dans nos analyses d'association.

Les conséquences sur le génome de ces variants ont été étudiées, et il s'avère qu'au total 73 des délétions (16 pour Lumpy, 22 pour Delly et 35 pour Pindel) ont des conséquences sur la séquence codante, dont 40 (14 pour Lumpy, 21 pour Delly et 5 pour Pindel) ont un impact fort, tels qu'une perte de codon stop ou une ablation de transcrit.

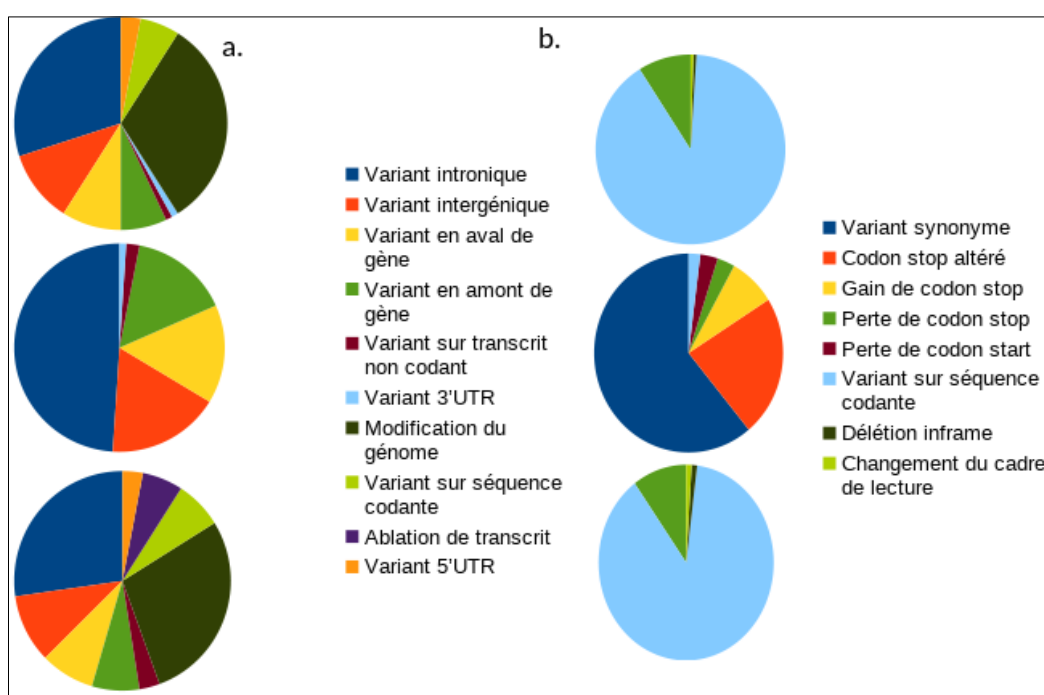


Figure 10: Conséquences sur le génome des variants en DL avec des SNP pour chaque outils. a. Conséquences complètes, b. Conséquences codantes. De haut en bas, Lumpy, Pindel, Delly

Comparaison à une liste de gènes impliquées dans la formation de la coquille

Parmi les délétions identifiées par Lumpy, 256 recouvrent complètement ou partiellement 135 gènes dont 97 sont dans la liste des gènes potentiellement impliqués dans la formation de

la coquille. Les conséquences de ces délétions sont résumées dans la figure 7. 12 délétions induisent une perte de codon stop ou une ablation de transcrit dans 10 des gènes d'intérêt.

Pour Delly, 350 délétions superposées à 147 gènes dont 97 des gènes d'intérêt ont été identifiées. 4 gènes sont fortement altérés par 5 délétions, qui ont pour conséquence une perte de codon stop ou une ablation de transcrit. Le reste des conséquences de ces délétions est présenté dans la figure 7

Enfin pour Pindel, 2366 délétions superposées à 286 gènes dont 168 des gènes d'intérêt ont été identifiées. De ces délétions, 7 ont d'importantes conséquences fonctionnelles sur 6 gènes d'intérêt.

De tous ces gènes d'intérêt modifiés par des délétions, 66 sont modifiés par des délétions identifiées par les trois outils. Parmi ces 66 gènes, un a retenu notre attention, CHIA. Une des délétions identifiées a un impact fort sur la séquence de ce gène (suppression du début du gène).

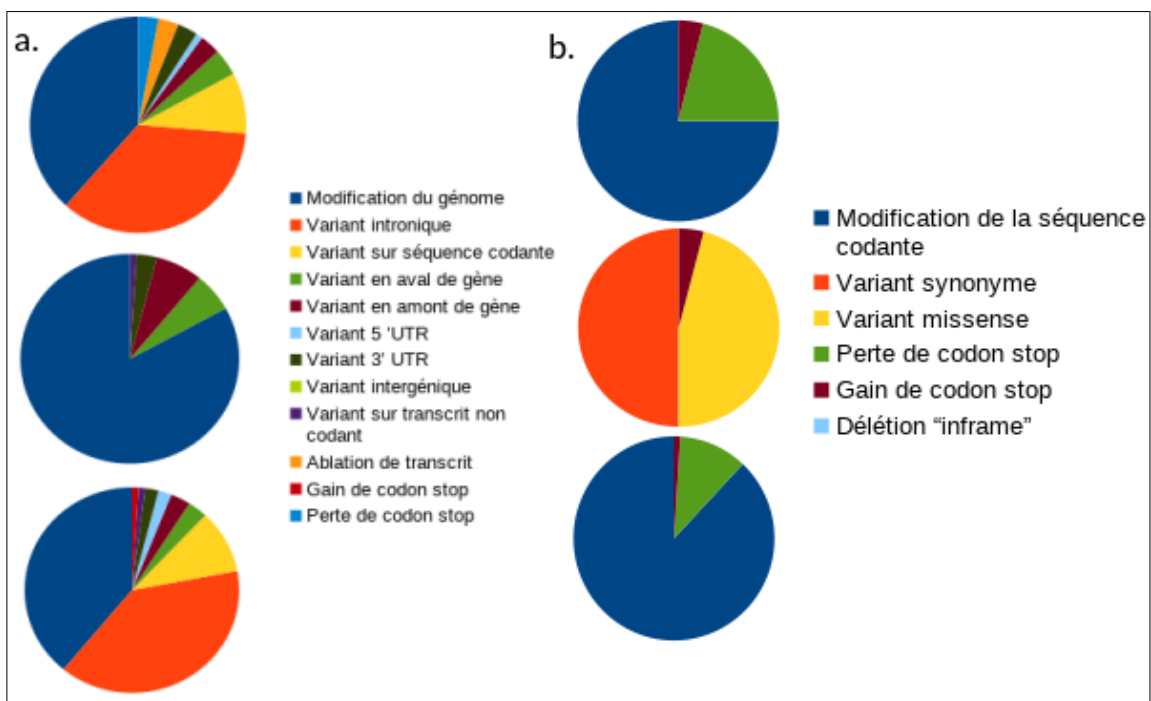


Figure 11: Conséquences sur le génome des variants à proximité des gènes d'intérêt pour chaque outils. a. Conséquences complètes, b. Conséquences codantes. De haut en bas, Lumpy, Pindel, Delly

Comparaison des résultats

Les variants significatifs identifiés par analyse d'association ne sont retrouvés ni dans les variants en déséquilibre de liaison avec les SNP, ni dans les gènes d'intérêt. En revanche, 112 variants (6 pour Lumpy, 15 pour Delly et 91 pour Pindel) en déséquilibre de liaison avec des SNP recouvrent partiellement ou totalement 30 des gènes d'intérêt (4 pour Lumpy, 11 pour Delly et 15 pour Pindel). Parmi ces gènes, trois sont modifiés par des délétions identifiées par les trois outils, FSTL1, PTN et PTPRF.

Tableau 2 : gènes d'intérêts et caractères en déséquilibre de liaison avec les délétions sur ces gènes

Gène	Caractère associé
FSTL1	Di et HA
PTN	FF
PTPRF	De

Discussion

Seules les délétions ont été étudiées ici, pour différentes raisons, la principale étant un manque de temps. Ensuite, les résultats des trois outils n'étaient pas homogènes, Pindel par exemple faisant une différence entre duplication et duplication en tandem.

Les chromosomes « scaffolds » sont des contigs qui n'ont pas pu être assignés à un endroit précis du génome. Au nombre d'environ 20 000 dans la version 5 de Gallus gallus, ces fragments ralentissaient énormément le processus de détection des variants structuraux. De plus, n'ayant pas de localisation précise, et ne portant pas de gènes ou de QTL connus la plupart du temps, les CNV potentiellement identifiables sur ces fragments n'auraient pas d'intérêt. Pour ces raisons de manque d'efficacité et d'information, les ADN « scaffolds » ont été retirés.

Il est difficile d'estimer le nombre de faux positifs obtenus lors de la détection par les différents outils, à l'exception des variants de taille excessivement importante. Il est en effet peu probable d'observer la délétion quasi-complète d'un chromosome, ce qui est un résultat

observé ici. Par exemple Lumpy détecte un variant de 195 Mb sur le chromosome 1, alors que la taille du chromosome 1 est de 196 Mb. Un tel résultat peut être dû à un mauvais alignement d'un split-read dans une région hautement répétée, par exemple. Globalement, la détection de CNV est moins fiable et plus compliquée dans les régions hautement répétées. C'est la raison pour laquelle il sera nécessaire de valider les variants intéressants par biologie moléculaire, ce qui est en cours pour une délétion identifiée sur CHIA.

Étant donné que la détection de variants s'effectue individu par individu et que le regroupement des fichiers n'est effectué qu'après, un grand nombre de génotype était considéré comme manquant lors du regroupement. Ceci entraînait la suppression de la quasi-totalité des variants lors du contrôle qualité pour les analyses d'association, à cause d'une call rate trop faible. Or ce comportement n'était pas voulu, étant donné que les génotypes manquant ne sont pas mal génotypés, ils sont simplement identiques à la référence. C'est la raison pour laquelle il a été nécessaire de modifier les génotypes manquants en génotypes de référence.

L'absence de correction pour test multiple entraîne le risque de retenir comme significatif des variants non significatifs. Cependant, étant donné la taille restreinte de la population et le faible nombre de variants, aucun résultat n'aurait été retenu comme significatif avec une correction. De plus, cette étude est une analyse préliminaire des CNV dans le génome de poules pondeuses. Ce manque de puissance des analyses d'association est la raison pour laquelle nous avons tenté de caractériser les variants d'autres façons.

Les délétions identifiées par Pindel sont sensiblement plus petites et plus nombreuses que pour les deux autres outils. Une grande partie des variants détectés par Pindel pourraient être qualifiés d'indel, des petites insertions/délétions, généralement de moins de 50 bases. Ces variants sont plus répandus dans la population que les variants de grande taille. La part importante de délétions de mauvaise qualité identifiée par Delly est due à l'absence de contrôle qualité du logiciel à l'issue de la détection. De plus un grand nombre de variants identifiés par ce logiciel ne sont présents que chez un seul individu. Ceci explique qu'une partie importante des variants ait été exclue à cause de la valeur de la MAF : ils sont peu représentés dans la population.

Un seuil de 20kb a été choisi pour estimer qu'un variant était lié à un SNP. En effet 20kb donne en moyenne sur tous les chromosomes du génome de la poule un déséquilibre de liaison supérieur ou égal à 0,5 (Hérault et al., 2018), que l'on peut considérer comme assez

fort pour affirmer que le SNP et le variant sont liés. L'absence de CNV significatifs co-localisés à des SNP significatifs peut s'expliquer par le petit effectif disponible pour les analyses d'association. Comme dit précédemment, les analyses d'association effectuées manquent de puissance.

Parmi les gènes identifiés qu'il faudra regarder plus en détail, quatre ont particulièrement retenu notre attention : CHIA, FSTL1, PTN et PTRF. FSTL1 code pour la follistatine like 1, une protéine impliquée dans le développement embryonnaire, et qui participe à la formation des cellules musculaires cardiaques (GenRef). La pléiotrophine, protéine codée par PTN joue un rôle dans les carcinomes ovariens et est un marqueur génétique dans les somites aviaires et le développement des tendons. Le récepteur type tyrosine-protéine phosphatase F, codé par PTRF participe aux jonctions adhérentes et donc joue un rôle dans l'adhésion cellulaire. Enfin, CHIA code pour la chitinase acide. Afin de confirmer l'existence de délétions à proximité de ces gènes, puis potentiellement de confirmer leurs conséquences plusieurs étapes de biologies moléculaires sont encore nécessaires. La validation de la délétion identifiée à proximité de CHIA est en cours.

En conclusion, malgré quelques faiblesses des analyses dues à la taille de la population, des variants intéressants ont été identifiés et leur existence devra être confirmée par l'expérimentation. S'il s'avère que ces variants existent dans cette lignée, il faudra par la suite confirmer leur existence dans d'autres lignées, voire comparer des lignées de poules pondeuses avec des lignées de poulets de chair.

Références bibliographiques

- Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376.
- 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.
- Bickhart, D.M., Hou, Y., Schroeder, S.G., Alkan, C., Cardone, M.F., Matukumalli, L.K., Song, J., Schnabel, R.D., Ventura, M., Taylor, J.F., et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* 22, 778–790.
- Chen, L., Chamberlain, A.J., Reich, C.M., Daetwyler, H.D., and Hayes, B.J. (2017). Detection and validation of structural variations in bovine whole-genome sequence data. *Genet. Sel. Evol.* 49.
- Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* 12, 966–968.
- Fan, W.-L., Ng, C.S., Chen, C.-F., Lu, M.-Y.J., Chen, Y.-H., Liu, C.-J., Wu, S.-M., Chen, C.-K., Chen, J.-J., Mao, C.-T., et al. (2013). Genome-Wide Patterns of Genetic Variation in Two Domestic Chickens. *Genome Biol. Evol.* 5, 1376–1392.
- Héroult, F., Herry, F., Varenne, A., Burlot, T., Picard-Druet, D., Recoquillay, J., Macé, C., Fagnoul, F., Allais, S., and Le Roy, P. (2018). A linkage disequilibrium study in layers and roiler commercial chicken populations.
- Hou, Y., Liu, G.E., Bickhart, D.M., Matukumalli, L.K., Li, C., Song, J., Gasbarre, L.C., Van Tassell, C.P., and Sonstegard, T.S. (2012). Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics* 12, 81–92.
- Ionita-Laza, I., Rogers, A.J., Lange, C., Raby, B.A., and Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* 93, 22–26.

Kadri, N.K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U.S., Panitz, F., Aamand, G.P., Schulman, N., et al. (2014). A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. *PLoS Genet.* *10*, e1004049.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* *15*, R84.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* *27*, 2987–2993.

Liu, S., Yao, L., Ding, D., and Zhu, H. (2010). CCL3L1 Copy Number Variation and Susceptibility to HIV-1 Infection: A Meta-Analysis. *PLoS ONE* *5*, e15778.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* *20*, 1297–1303.

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 Biol.* *17*.

Moro, C., Cornette, R., Vieaud, A., Bruneau, N., Gourichon, D., Bed'hom, B., and Tixier-Boichard, M. (2015). Quantitative Effect of a CNV on a Morphological Trait in Chickens. *PLOS ONE* *10*, e0118706.

Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. *Front. Genet.* *06*.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* *28*, i333–i339.

Romé, H., Varenne, A., Hérault, F., Chapuis, H., Alleno, C., Dehais, P., Vignal, A., Burlot, T., and Le Roy, P. (2015). GWAS analyses reveal QTL in egg layers that differ in response to diet differences. *Genet. Sel. Evol.* 47.


Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong Association of De Novo Copy Number Mutations with Autism. *Science* 316, 445–449.

Stone, J.L., O'Donovan, M.C., Gurling, H., Kirov, G.K., Blackwood, D.H.R., Corvin, A., Craddock, N.J., Gill, M., Hultman, C.M., Lichtenstein, P., et al. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237–241.

Wang, X., and Byers, S. (2014). Copy Number Variation in Chickens: A Review and Future Prospects. *Microarrays* 3, 24–38.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14, S1.

	Diplôme : Ingénieur agronome Spécialité : Biologie moléculaire et cellulaire Spécialisation / option : Enseignant référent : Frédéric Lecerf
Auteur(s) : Morgane Gaudin Date de naissance* : 11/11/1995	Organisme d'accueil : INRA UMR PEGASE Adresse : 16, Le Clos, Saint-Gilles 35590
Nb pages : 20 Annexe(s) : 0	
Année de soutenance : 2018	Maître de stage : Frédéric Hérault
Titre français : Détection et caractérisation de variants structuraux dans le génome d'une lignée de poules pondeuses Titre anglais : Detection and characterization of structural variants in the genome of a line of laying hens	
Résumé (1600 caractères maximum) : Les variations du nombre de copies (CNV), des insertions ou délétions de 50 bases consécutives ou plus, sont une importante source de diversité génétique. Du fait de leur taille elles peuvent avoir un fort impact phénotypique. Dans cette étude une détection sur tout le génome des CNV a été effectuée sur les séquences de 50 coqs d'une lignée de poules pondeuses Rhode Island. La détection a été effectuée au moyen de trois outils différent combinant deux méthodes de détection : paired-end mapping et split-read. Au total, environ 50 000 délétions ont été identifiées, représentant environ 3% du génome, allant de une base à quelques mégabases. Des analyses d'association ont été effectuées pour 7 caractères. Au total, 23 délétions localisées dans 16 gènes ont été identifiées. Les délétions identifiées ont aussi été comparées à des résultats de GWAS, et 73 des délétions en déséquilibre de liaison avec des SNP tombent dans des gènes. Enfin, les délétions ont été comparées à une liste de gènes en lien avec la formation de la coquille d'œuf et 66 gènes modifiés par des délétions ont été identifiés. En conclusion, des variants intéressants ont été identifiés, et premier aperçu des délétions dans cette lignée de poules pondeuses a été dressé.	
Abstract (1600 caractères maximum) : Copy number variants (CNV) are insertions or deletions of 50 consecutive bases or more and are an important source of genetic diversity. Because of their sheer size, their impact on phenotypes can be important. Here we performed CNV detection on the genome of 50 chicken from a line of laying hens (Rhode Island). The detection was performed with three tools combining two methods of detection: paired-end mapping and split-read analysis. In total about 50000 deletions were identified, covering about 3% of the genome and ranging in length from one base to a few megabases. Association analysis were performed for 7 traits, egg weight, shell color, egg shell resistance, deformation, diameter, albumen height and shape index. 23 deletions located in 16 genes were identified. The deletions were also compared to GWAS results and 73 of the deletions in linkage disequilibrium with significant SNP were located in genes. Finally, the deletions were compared to a list of genes involved in egg shell formation 66 genes were impacted by deletions from all three tools. In conclusion, interesting variants have been identified and this study gives a first overview of CNV in this line of chicken.	
Mots-clés : génétique, génomique, variants structuraux, cnv, poules pondeuses, délétions, ADN Key Words: genetic, genomics, structural variants, cnv, laying hen, deletions, DNA	

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