Master de Bioinformatique
parcours
du Génome aux Écosystèmes

Année promotionnelle 2019

Rapport de stage

Caractérisation des chromosomes sexuels du génome de l'*Ornithorhynchus anatinus*

Auteure :
Julie Blasquiz

Maître de stage :
Julie Hussin, PhD
IVADO Professeur Assistant
Université de Montréal

1er septembre 2019

Table des matières

Nomenclature

ACP Analyse en Composantes Principales, page 13

ADN Acide DésoxyriboNucléique, page 9

BAM Binary Alignment Map, page 10

CNSW Central New South Wales, page 26

CP Composantes Principales, page 20

FASTQ Fast (Alignment) Quality, page 10

GCA GenBank assembly accession, page 9

ICM Institut de Cardiologie de Montréal, page 5

IGV Integrative Genomics Viewer, page 14

IVADO Institut de Valorisation des Données, page 1

NGS Next Generation Sequencing, page 9

NNSW North New South Wales, page 26

NQLD North QueensLanD, page 26

PAR Pseudo-Autosomal Region, page 6

PCR Polymerase Chain Reaction, page 10

SGS Second-Generation Sequencing, page 9

SNP Single Nucleotide Polymorphism, page 11

UMAP Uniform Manifold Approximation and Projection, page 13

VCF Variant Call Format, page 11

WGS Whole Genome Shotgun, page 9

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au bon déroulement de mon stage.

Je tiens à remercier vivement ma maître de stage, Mme Julie Hussin, responsable du groupe de bioinformatique OMICS au sein de l'Institut de Cardiologie de Montréal (ICM), pour son accueil chaleureux, ses recommandations, ses suggestions et conseils apportés tout le long du stage.

Je remercie également toute l'équipe du laboratoire de Mme Julie Hussin pour leur accueil, leur esprit d'équipe et en particulier Mr Jean-Christophe Grenier, bioinformaticien de l'équipe, qui m'a beaucoup aidée notamment dans l'utilisation du cluster de calcul.

Enfin, je tiens à remercier Dominique Fournelle pour sa coopération étroite tout le long de ce projet notamment pour sa contribution au poster présenté lors de la journée de la recherche de l'institut de cardiologie (cf. annexe A) et de son aide quant à la validation ou non de la catégorisation des contigs.

1 Introduction

L'ornithorynque est un mammifère pondeur qui, à côté de l'échidné, occupe une place unique dans l'arbre phylogénétique des mammifères. Malgré un intérêt général pour sa biologie inhabituelle, on en sait peu sur sa structure de population ou son évolution récente[1].

Durant mes six mois de stage à l'Institut de Cardiologie de Montréal, du 4 mars au 30 août 2019, j'ai pu exploré sa génomique afin de comprendre son système inhabituel et complexe de chromosomes sexuels. L'ornithorynque possède, en effet, cinq paires de chromosomes sexuels.

Mes objectifs ont été la validation de la caractérisation de contigs préalablement identifiés X, Y et autosomaux provenant de cinquante-sept génomes d'ornithorynques de différentes populations australiennes, l'identification de nouveaux contigs X, Y et autosomaux et enfin l'identification des Régions Pseudo-Autosomales (Pseudo-Autosomal Region, PAR). Pour atteindre ces objectifs, j'ai notamment travaillé avec les caractères de données manquantes, de couverture et d'hétérozygotie pour l'identification des contigs X,Y, autosomaux et des PAR.

Ce rapport comprend quatre parties : une première partie dans laquelle le contexte biologique de l'ornithorynque, les méthodes de séquençage de nouvelle génération et les trois critères de données manquantes, de profondeur et d'hétérozygotie sont introduits. Une deuxième partie qui tient compte des matériels et méthodes des résultats obtenus. Une troisième partie qui explique les résultats obtenus. Et enfin, une dernière partie portant sur l'analyse et le bilan de tout le travail réalisé durant ce stage.

2 État de l'art

2.1 L'ornithorynque *Ornithorhynchus anatinus*

Les ornithorynques vivent dans les rivières de l'Est et du Sud de l'Australie ainsi qu'en Tasmanie. Ils sont encore relativement courants dans la nature, mais ont récemment été reclassés comme «vulnérables» en raison de leur dépendance à un environnement aquatique soumis au stress du changement climatique et à la dégradation due aux activités humaines [2].

L'ornithorynque a toujours suscité de l'intérêt et de la controverse dans le monde zoologique [3]. Malgré un physique atypique comprenant un bec de canard, de la fourrure, des pattes palmées et une queue de castor, il a été considéré qu'il s'agissait d'un véritable mammifère. L'ornithorynque (*Ornithorhynchus anatinus*) a été placé avec les échidnés dans un taxon appelé monotrème (qui signifie «trou unique» en raison de leur ouverture externe commune pour les systèmes uro-génital et digestif). Les monotrèmes appartiennent à la sous-classe des mammifères protothériens, qui a divergé de la ligne des thérapsides menant aux thériens et s'est ensuite scindée en marsupiaux et euthériens [2].

L'ornithorynque est le résultat d'un amalgame de caractéristiques ancestrales dérivées des reptiles et des mammifères [2] : l'ornithorynque, tout comme l'ensemble des mammifères, sécrète du lait (celui de l'ornithorynque a la particularité de posséder des propriétés antibactériennes très développées [4]). L'ornithorynque présente également l'une des caractéristiques du taxon des reptiles à savoir la ponte d'oeuf. De plus, le venin des ornithorynques, sécrété uniquement par les mâles et se trouvant dans un aiguillon au niveau des pattes postérieures de l'animal, présente de fortes similarités avec celui des reptiles [2].

L'ornithorynque *Ornithorhynchus anatinus* présente également diverses particularités au niveau de son génome. Le caryotype de cette espèce comprend cinquante-deux chromosomes dont dix chromosomes sexuels : dix chromosomes X chez les femelles et une alternance de cinq X et de cinq Y chez les mâles [5]. Les chromosomes sexuels de l'ornithorynque partagent une homologie plus élevée avec les chromosomes sexuels Z et W des oiseaux qu'avec les chromosomes sexuels des mammifères, malgré l'affichage d'un système XY déterminant le sexe [5][6]. Lors de la méiose masculine, les chromosomes sexuels forment une chaîne alternée de chromosomes X et Y reliés par neuf Régions Pseudo-Autosomales (PAR) [7]. Les PAR sont des régions de chromosomes sexuels recombinants qui se comportent comme des autosomes.

Plus de la moitié du génome de l'ornithorynque est constituée d'éléments génétiques mobiles, appelés transposons, dont certains sont encore des moteurs actifs de l'évolution

génomique chez cette espèce. A cela s'ajoute le fait que la fréquence des répétitions intercalées (de plus de deux répétitions par kilo-bases) chez l'ornithorynque *Ornithorhynchus anatinus* est supérieure à celle de tout génome de métazoaire précédemment caractérisé [2].

Ce génome complexe soulève de nombreuses questions quant au déroulement de la méiose dans cet organisme. L'étude du fonctionnement de celle-ci permettra notamment la compréhension à plus grande échelle de la méiose et de son évolution au sein des mammifères.

2.2 Le séquençage de nouvelle génération (NGS)

Le premier génome de référence de l'ornithorynque *Ornithorhynchus anatinus* (*GenBank assembly accession (GCA) : GCA_000002275.2)*, réalisé sur une femelle, ne laisse place à aucun chromosome Y référencé. Un second génome de référence, *(GenBank assembly accession : GCA_002966995.1)*, a alors été réalisé sur un mâle permettant cette fois le recensement de chromosomes X et Y [1] . Le nouveau génome de référence ainsi que les génomes de cinquante-sept individus ornithorynques ont été séquencés par l'approche de séquençage de nouvelle génération (NGS) [8]. Le NGS permet le séquençage de grandes quantités d'ADN à de faibles coûts. Le séquençage complet des génomes a été réalisé via la méthode dite «globale» ou encore *Whole Genome Shotgun* (WGS).

2.2.1 Pacific Biosciences

Le second génome de référence a été séquencé en utilisant la biotechnologie Pacific Biosciences (ou Pacbio). Cette technologie, aussi appelée *Single molecule sequencing*, fait partie des technologies de séquençage de seconde génération (SGS). Ce sont des séquenceurs capables de générer de très longs *reads* de dizaines de kilo-bases sans avoir besoin de cloner les fragments pour amplifier le signal [9].

L'assemblage du génome de référence a été réalisé avec la méthode d'assemblage *de novo* faisant recourt à des algorithmes utilisant les graphes de Bruijn [10] : les fragments d'ADN chevauchants permettent la construction de contigs et l'assemblage de ces contigs entre eux permet ensuite d'obtenir un *scaffold*.

2.2.2 Dovetail Hi-C *scaffolding*

Lors de la construction d'un assemblage de génome *de novo*, la contiguïté et la précision de l'assemblage sont deux éléments importants. La biotechnologie de Dovetail Genomics a pour objectif d'améliorer un assemblage avec deux méthodes dites de ligations de proximités propriétaires, Chicago et Dovetail Hi-C, et le logiciel de *scaffolding*, HiRise.

Des données de ligation de proximité propriétaires *in vitro* sont utilisées pour créer une contiguïté d'assemblage en réalisant des jointures à longue distance. Le logiciel de *scaffolding* HiRise utilise ces données pour rechercher et corriger les faux raccordements erronés dans l'ensemble d'entrée.

Ensuite, des bibliothèques Dovetail Hi-C sont construites en utilisant des cellules ou des tissus intacts. HiRise utilise les données Dovetail Hi-C pour établir des connexions à portée encore plus longue, jusqu'aux chromosomes complets, augmentant ainsi considérablement la contiguïté.

L'assemblage final est à la fois hautement contigu et très précis (*https ://dovetailgenomics.com/ga_tech_overview/*).

2.2.3 Illumina HiSeq

Les génomes des cinquante-sept individus ornithorynques, quant à eux, ont été séquencés à l'aide du séquenceur Illumina HiSeq X. Le séquençage par synthèse d'Illumina avec le séquenceur HiSeq X utilisé ici ne permet de lire que de courts fragments d'ADN de trois cent paires de bases qu'il faut par la suite ré-assembler pour reconstruire le génome complet.

Cette méthode consiste d'abord par cloner, via PCR (Polymerase Chain Reaction), plusieurs fois les fragments d'ADN afin d'amplifier leur signal. Puis le brin complémentaire de chaque fragment cloné est synthétisé. À chaque incorporation d'un nucléotide, un signal lumineux est détecté et associé à un nucléotide. L'ensemble des données est ensuite enregistré dans un fichier au format FASTQ contenant les séquences des *reads* et leurs scores de qualité (score Phred).

Chaque *read* est ensuite aligné sur le nouveau génome de référence. L'algorithme de Burrows Wheeler [11] est utilisé pour la recherche de correspondance entre les *reads* et la référence. Après cet alignement, on obtient des fichiers BAM associant à chaque *read* ses coordonnées génomiques (contig et position).

2.3 Les caractéristiques de données manquantes, d'hétérozygotie et de profondeur

2.3.1 Obtention des fichiers VCF

Les fichiers BAM obtenus après alignement sont soumis à une étape de nettoyage de données avant l'obtention des fichiers VCF (Variant Call Format) répertoriant les variations de séquence de gènes observées chez les différents individus, appelées SNP [12] (cf. Figure 2.1). Les SNP (Single Nucleotide Polymorphism) correspondent à des variations mineures du génome au sein d'une population [12].

Cette étape de préparation des données comprend deux étapes : la première étape consiste en la conservation des *reads* pairés proprement réalisée via l'outil Samtools [13] et la seconde étape en la filtration des *reads* PCR en double effectuée à l'aide de l'outil *Picard MarkDuplicates tool (http ://broadinstitute.github.io/picard).*

À la fin de la préparation des données, un nouveau fichier BAM est obtenu, pouvant être utilisé pour l'identification des SNP. Cette dernière étape de détection de variants, appelée *variant calling*, est réalisée via l'outil Platypus (signifiant ornithorynque en anglais). L'outil Platypus permet d'identifier les SNP, les *indels* et les insertions de deux cents paires de bases [14]. Les fichiers VCF obtenus reprennent l'ensemble de ces données.

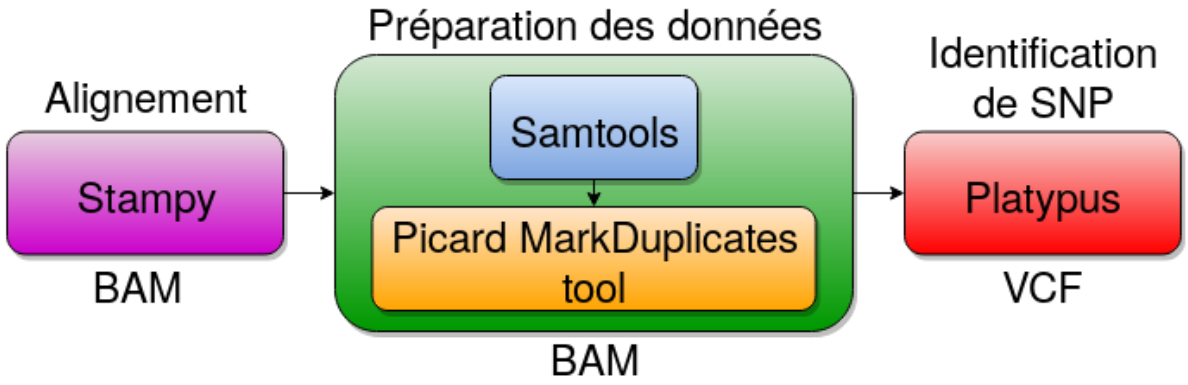

FIGURE 2.1 – Pipeline d'obtention des fichiers VCF

2.3.2 Données manquantes

La valeur de données manquantes, obtenue depuis les fichiers VCF, représente la proportion de SNP dans laquelle un génotype n'est pas présent. Une valeur de données manquantes par individu égale à zéro indique l'absence de données manquantes dans un contig donné et une valeur de un indique au contraire l'absence de SNP sur le contig en question. Plusieurs raisons peuvent justifier la présence de données manquantes dans

un génome : la mauvaise qualité de l'alignement, l'absence de *read* à cette position, une délétion ou encore un soupçon d'erreur de séquençage d'après les scores de qualité des fichiers FASTQ. Dans notre analyse, la valeur de données manquantes des contigs des individus femelles peut également permettre l'identification des contigs Y : le génome de référence étant un mâle, les données manquantes des femelles sont associées aux contigs Y.

2.3.3 Hétérozygotie

L'hétérozygotie est également obtenue via les fichiers VCF. Ce critère donne, par position ou par individu, le taux d'hétérozygotie. Une valeur proche de zéro indique un fort taux d'homozygotie tandis qu'une valeur proche de un indiquera un fort taux d'hétérozygotie. Les contigs Y étant hémizygotes (ne possédant qu'un allèle à un locus donné pour cause d'absence de région homologue), les régions hétérozygotes de ceux-ci correspondent aux PAR.

2.3.4 Profondeur

La profondeur correspond au nombre moyen de *reads* qui se superposent sur la zone d'intérêt (ex. un SNP)[15]. Ce critère, tout comme les deux précédents, peut être obtenu, pour chaque SNP, via les fichiers VCF. Sa valeur, quant à elle, peut varier de zéro à de très grands nombres (un contig d'ADN mitochondrial avec plus de deux cents de profondeur a déjà été observé). Un contig X présente deux fois plus de profondeur sur un individu femelle que sur un individu mâle par le simple fait que les femelles possèdent deux chromosomes X par paires de chromosomes sexuels alors que les les mâles n'en possèdent qu'un.

3 Matériels et Méthodes

L'alignement des cinquante-sept génomes ornithorynques a été fait sur le second génome de référence à l'aide de l'outil Stampy [16] (cf. Figure 2.1). Seuls les *reads* pairés proprement ont été conservés à l'aide de Samtools [13]. Picard MarkDuplicates tool *(http ://broadinstitute.github.io/picard)* a ensuite été utilisé pour filtrer les PCR en double. L'outil Platypus a permis par la suite la découverte des positions variables, les SNP [14]. Le filtrage des régions répétées (RepeatMasker) a été effectué à l'aide de bedtools [17]. Toutes ces étapes ont été réalisées dans une étude précédente [1].

Les calculs des statistiques des données manquantes pour identifier les contigs X,Y et autosomaux ainsi que les statistiques de l'hétérozygotie pour identifier les PAR des contigs sexuels ont également été réalisés avec l'aide de VCFtools [12]. Le calcul des statistiques de profondeur a été réalisé via VCFtools et BCFtools [18]. Les calculs des statistiques se sont faits via VCFtools en gardant uniquement les SNP avec l'annotation PASS (indiquant la bonne qualité du SNP) et les SNP bialléliques. Une sélection des contigs basée sur leurs tailles et leurs nombres de SNP après filtres a été réalisée afin d'éliminer notamment ceux avec trop peu d'informations pour être considérés comme étant informatifs : seuls les contigs de plus de quarante mille kilo-bases et de plus de cinquante SNP après filtres ont été conservés constituant un panel de cinq cent quatre-vingt trois contigs (cf. annexe B).

L'extraction des valeurs de profondeur, de données manquantes et d'hétérozygotie fournit pour chaque critère une matrice dont le nombre de colonnes est égal au nombre d'individus analysés, cinquante-sept ici, et le nombre de lignes au nombre de contigs considérés, ici cent-quarante et un avec les contigs préalablement identifiés. Chaque valeur de la matrice correspond au critère (données manquantes, hétérozygotie ou profondeur) moyenné par individu par contig.

La profondeur est également prélevée par position pour les contigs d'intérêt. La profondeur par position est extraite directement des fichiers BAM à l'aide de l'outil Samtools. Ces valeurs sont ensuite normalisées par la moyenne de profondeur par individu.

L'analyse en composantes principales (ACP) est une technique permettant de réduire la dimensionnalité de grands jeux de données, en augmentant l'interprétabilité, tout en minimisant la perte d'informations. Pour ce faire, l'ACP crée de nouvelles variables non corrélées qui maximisent successivement la variance [19].

L'approximation et projection de variétés uniformes (UMAP) est une technique de réduction de dimensionnalité non linéaire développée pour l'analyse de tout type de données de grande dimension [20].

L'analyse précise des régions d'intérêt des contigs à été effectuée avec IGV, *Integrative Genomics Viewer*, un outil de visualisation hautes performances pour l'exploration interactive de grands ensembles de données génomiques permettant, entre autre, la visualisation des insertions, des délétions, et de la profondeur [21].

L'utilisation de ces différents outils s'est réalisée via des scripts bash et les figures portant sur ces données ont été réalisées avec le langage de programmation R via la librairie de visualisation de données «ggplot2» [22]. Les ACP et UMAP ont spécifiquement été réalisées en R via les librairies respectives «ggfortify» et «umap» en utilisant les paramètres par défaut.

L'ensemble des scripts, matrices, fichiers et figures obtenus tout au long du stage sont contenus dans mon GitHub.

4 Résultats et discussions

4.1 Analyse des chromosomes sexuels

Cent-quarante et un contigs sur un total de quatre-mille-cinq-cent-soixante-douze ont préalablement été identifiés comme étant des X, Y et autosomaux (cf. Table E.1). Les trente-huit contigs X et les soixante-quatre contigs autosomaux ont pu être identifiés comme tels via leurs alignements sur le premier génome de référence femelle. Les trente-neuf contigs restants ont quant à eux été identifiés Y via la présence de gènes Y spécifiques. Ces identifications ont été préalablement effectuées dans une étude antérieure.

Table 4.1 – Nombre de contigs préalablement identifiés par catégories

Catégories	**Nombre de contigs**
Contigs X	38
Contigs Y	39
Contigs autosomaux	64
Total	141

Nous avons déduit la composition des chromosomes sexuels pour chacun des individus ornithorynques sur la base de l'homozygotie sur les chromosomes X et le taux de génotypes non manquants sur le chromosome Y (cf. Figure 4.1).

Une proportion de génotypes homozygotes sur le chromosome X égale à zéro indique que ce chromosome présente une forte hétérozygotie tandis qu'une valeur de un indique une forte homozygotie.

Nous avons identifié les échantillons comme mâle ou femelle, selon les critères suivants : les mâles ont une proportion de génotypes homozygotes sur le chromosome X supérieure à 0,9, et un fort taux de génotypes non manquants sur le chromosome Y. Les femelles ont un faible taux de génotypes non manquants sur le chromosome Y inférieur à 0,4 [23].

Basés sur ces critères, deux individus sont mal répertoriés : un individu est répertorié mâle (en triangle rose dans le groupe des femelles) mais son profil est identique à celui d'une femelle et ,inversement, un individu est répertorié femelle (en rond rose dans le groupe des mâles) mais son génotype est identique à celui d'un mâle. Suite à ces obser-

vations, nous avons corrigé l'attribution des sexes de ces deux individus.

Ce graphique distinguant les mâles des femelles sur les critères de données manquantes et d'hétérozygotie induit le fait que ces deux critères peuvent être considérés pour la distinction entre les contigs X et les contigs Y et ainsi être utilisés pour l'élaboration des profils des contigs.

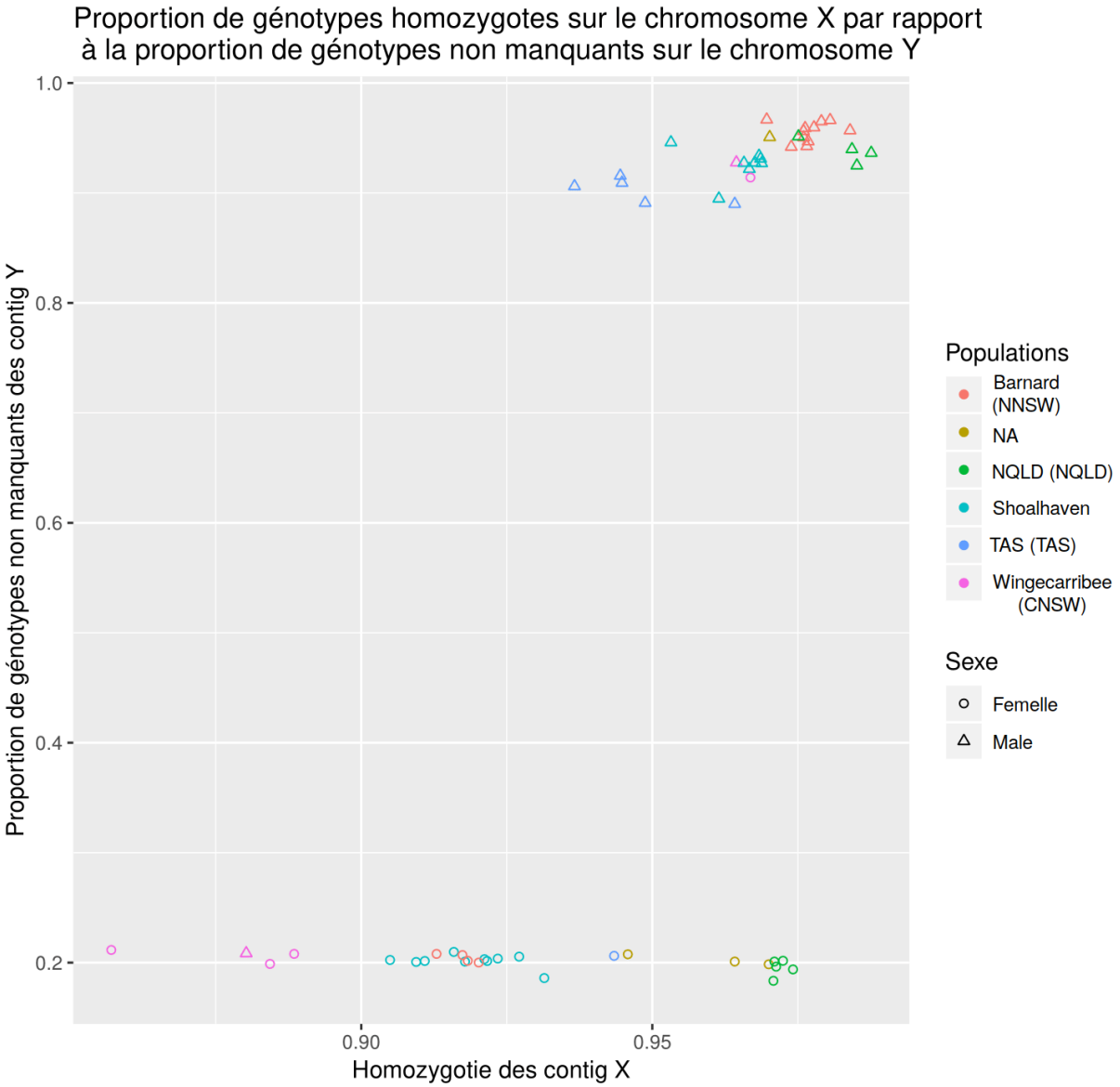

FIGURE 4.1 – Proportion de génotypes homozygotes (égale à un moins la proportion de génotypes d'hétérozygotes) sur le chromosome X par rapport à la proportion de génotypes non manquants sur le chromosome Y

4.2 Profils des contigs X, Y et autosomaux

Nous avons réalisé les profils des contigs X, Y et autosomaux sur la profondeur, sur les valeurs de données manquantes et sur l'hétérozygotie en distinguant les valeurs des mâles en bleu et celles des femelles en rouge (cf. Figure 4.2).

Pour l'ensemble des trois critères, les contigs autosomaux ne présentent qu'une faible différence entre mâles et femelles. Les contigs autosomaux sont en effet diploïdes à la fois chez les mâles et les femelles.

Les contigs X, quant à eux, sont deux fois plus nombreux chez les femelles que chez les mâles. Ainsi, la profondeur des contigs X est deux fois supérieure chez les femelles que chez les mâles. De même, la valeur de données manquantes pour les contigs X est nettement supérieure chez les mâles que chez les femelles, de par la profondeur réduite (moins de chance d'échantillonner un SNP donné). Les femelles, présentant deux X par paires de chromosomes sexuels, présentent un taux d'hétérozygotie plus important que pour les mâles qui sont eux hémizygotes pour les contigs X.

Les contigs Y, n'étant présents que chez les mâles, la profondeur devrait donc être nulle chez les femelles. De même, les contigs Y étant absents chez les femelles, la valeur de données manquantes des femelles est de un pour les femelles et au contraire proche de zéro pour les mâles. Les contigs Y, n'étant présents qu'en un seul exemplaire par paire de chromosomes sexuels, sont donc homozygotes pour les mâles.

Bien que la grande majorité des contigs présentent les critères expliqués ci-dessus, ceux encadrés en noir sur la Figure 4.2, s'en éloignent grandement. Ces contigs aux profils anormaux chez les autosomaux sont en réalité des contigs X ayant étaient mal assignés. Les autres contigs aux profils anormaux chez les X et les Y sont bien quant à eux des contigs sexuels mais des contigs sexuels comprenant des PAR. Parmi eux, six, en rouge dans la Table 4.2, présentent une région X spécifique ainsi qu'une PAR avec une limite entre ces deux régions bien visibles (plus d'explication sur les limites de PAR dans la section 4.5).

La présence de ces régions PAR communes aux contigs X et Y modifie les profils attendus : les contigs X avec PAR présentent notamment un taux de profondeur équivalent pour les mâles et les femelles tandis que les contigs Y avec PAR présentent de la profondeur chez les femelles (cf. Figure 4.2).

Le contig X numéro 265, corrigé dorénavant en contig Y, avait quant à lui été assigné aux contigs X par l'alignement de sa PAR sur l'un des chromosomes X du premier génome de référence. L'ensemble des contigs nouvellement assignés sont listés dans la Table 4.2.

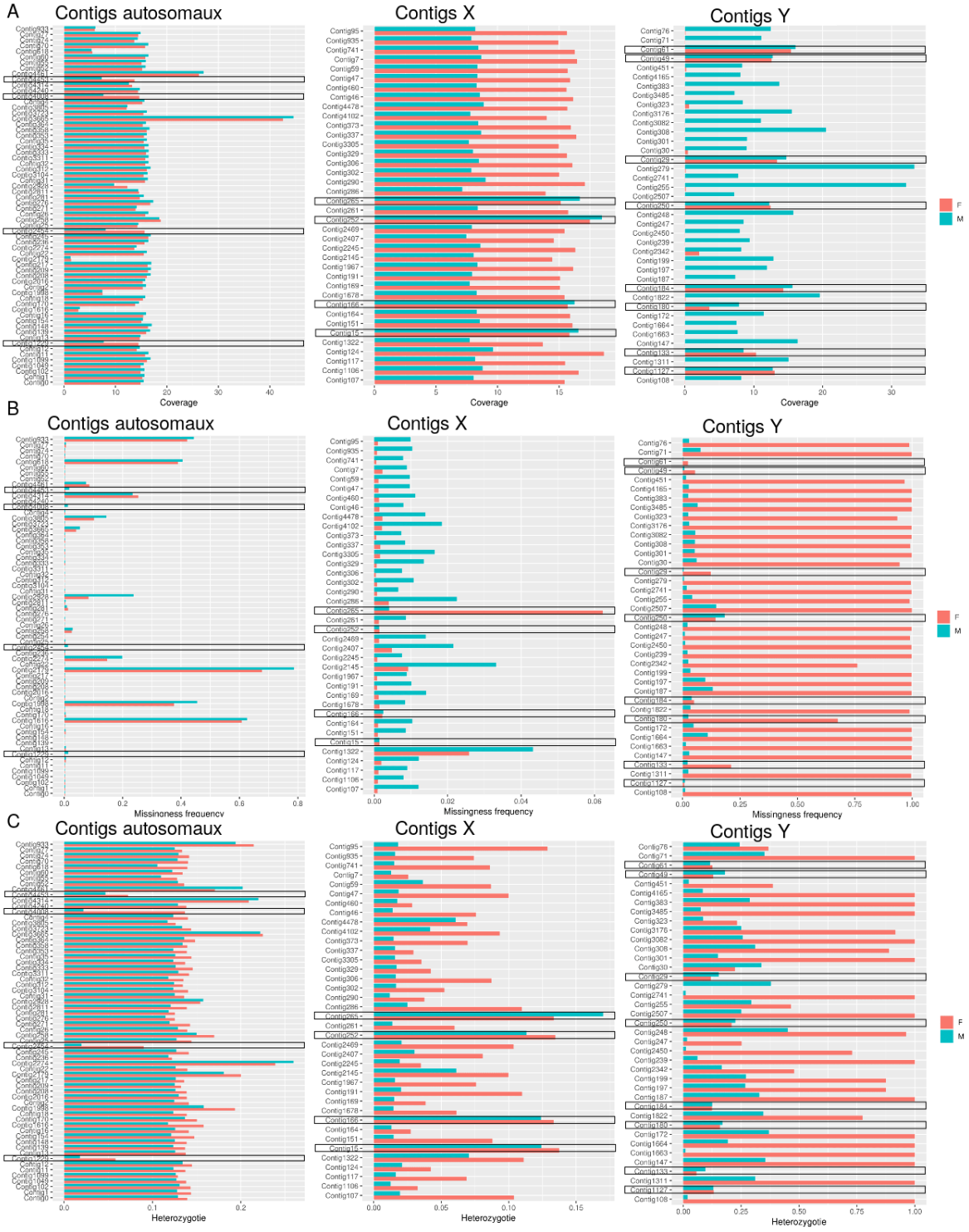

FIGURE 4.2 – Profils des contigs préalablement catégorisés selon A) la profondeur, B) la valeur de données manquantes et C) l'hétérozygotie des mâles en bleu et des femelles en rouge

TABLE 4.2 – Ré-assignation des contigs catégorisés avec en rouge les contigs présentant les limites des PAR

Contigs	**1ère assignation**	**2nde assignation**
Contig 1229	Autosome	X
Contig 2454	Autosome	X
Contig 4008	Autosome	X
Contig 4453	Autosome	X
Contig 15	X	X (PAR)
Contig 166	X	X (PAR)
Contig 252	X	X (PAR)
Contig 265	X	Y (PAR)
Contig 1127	Y	Y (PAR)
Contig 133	Y	Y (PAR)
Contig 180	Y	Y (PAR)
Contig 184	Y	Y (PAR)
Contig 29	Y	Y (PAR)
Contig 49	Y	Y (PAR)
Contig 61	Y	Y (PAR)
Contig 250	Y	Y (PAR)

4.3 Mise en évidence des contigs avec PAR

Les matrices de données manquantes, d'hétérozygotie et de profondeur constituent un grand ensemble de données pouvant être difficiles à interpréter. Nous avons réalisé une ACP sur ces matrices en ne conservant que les deux premières Composantes Principales (CP) des données de profondeur des contigs préalablement identifiés (cf. Figure A 4.3). Chaque point de l'ACP correspond à un contig. Les contigs X sont colorés en rouge, les Y en vert et les autosomaux en bleu. L'ACP révèle trois «branches», chacune d'entre elles représentant les catégories X, Y et autosomaux. L'ACP capte la différence de profondeur entre mâles et femelles observée sur les profils des contigs. L'absence de différence de profondeur entre mâles et femelles crée le groupe des autosomaux. La différence du simple au double entre mâles et femelles des contigs X forme la cohorte des X. La différence de profondeur entre les femelles de valeur nulle et les mâles constitue l'ensemble des Y.

Les branches des X et des Y sont pures, contenant uniquement les contigs de leurs catégories. La branche des contigs autosomaux contient l'ensemble des contigs autosomaux et également des contigs X et Y. Ces contigs sexuels n'appartenant pas à leurs catégories dans l'ACP sont les contigs identifiés via les profils comme présentant des PAR. Ces contigs présentant des PAR ayant une région semblable aux contigs autosomaux se classent dans la catégorie des autosomaux. Cette appartenance des contigs PAR dans la branche des autosomaux confirme la présence de PAR dans ces contigs.

Nous avons ensuite cherché à résumer les CP en deux dimensions visualisables sur un même graphique en utilisant la UMAP. Une première UMAP sur les données de profondeur des contigs préalablement identifiés a été réalisée (cf. Figure B 4.3). Tout comme avec l'ACP, un point de la UMAP correspond à un contig et le même code couleur a été respecté. La UMAP forme trois groupes distincts, le groupe des X, des Y et des autosomaux comprenant les PAR. Contrairement à l'ACP, la UMAP produit des groupes mixtes où le groupe des Y comprend quatre autosomaux et le groupe des X un autosome et un Y. Ces cinq contigs ont tous été assignés au chromosome 7. Il faudra par la suite regarder de plus près ces contigs afin de comprendre cette assignation.

Afin d'obtenir des groupes par catégorie comme dans la UMAP tout en gardant l'exactitude de ces groupes comme dans l'ACP et d'augmenter le nombre de composantes principales à plus de deux, nous avons décidé d'appliquer une UMAP sur les dix premières composantes principales de l'ACP (cf. Figure C 4.3).

La seconde UMAP réalisée sur les dix premières CP de l'ACP de la profondeur présente là encore trois groupes, un groupe X, un groupe Y et un groupe avec les autosomaux et PAR. Cette fois-ci, les groupes X et Y de la UMAP ne présentent que les contigs de leurs catégories.

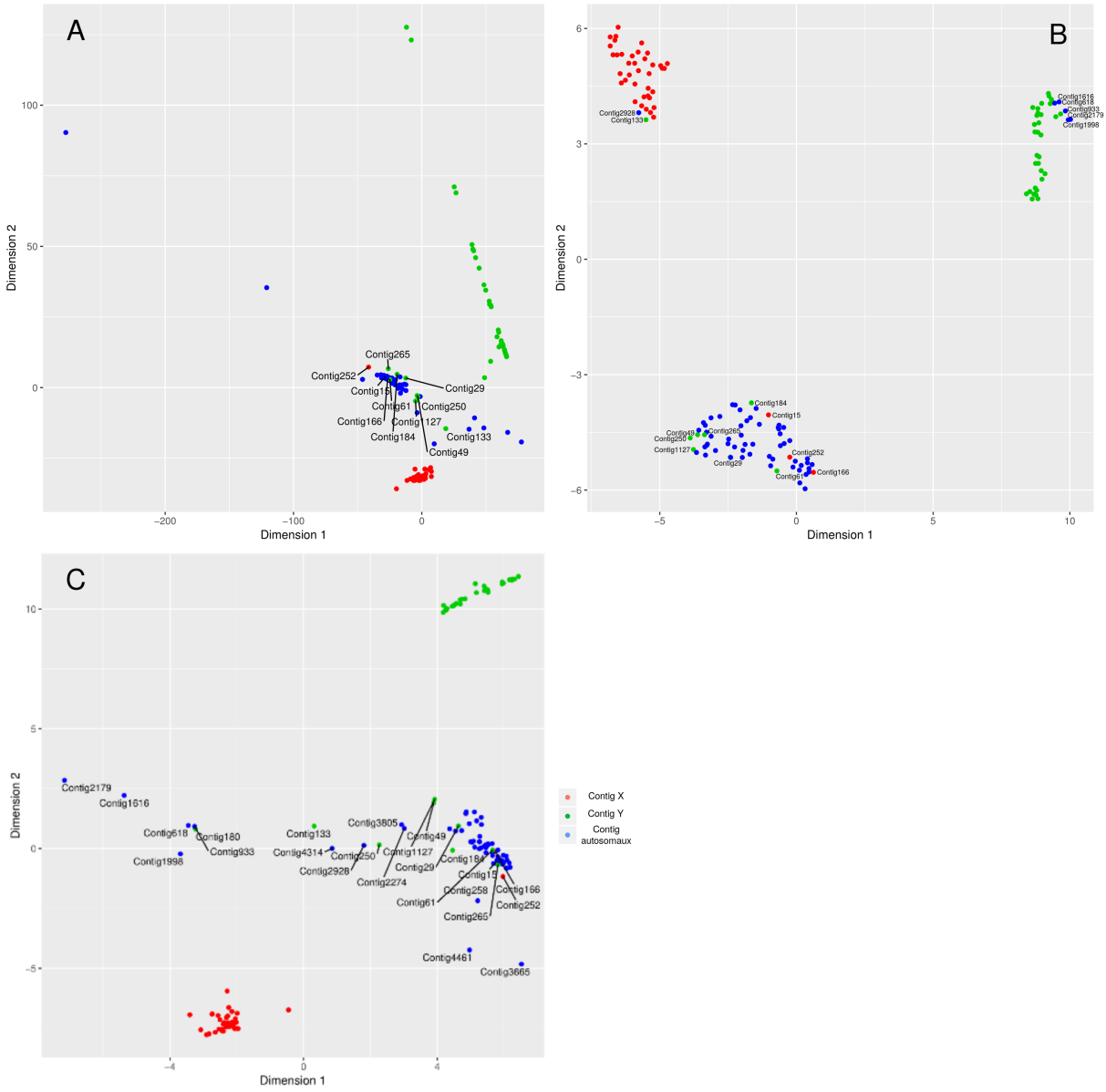

FIGURE 4.3 – A) ACP sur les données de profondeur B) UMAP sur les données de profondeur C) UMAP sur les 10ers CP de l'ACP sur les données de profondeur des contigs préalablement catégorisés

4.4 Classification des contigs

L'objectif est dorénavant la classification des contigs non catégorisés. En sélectionnant uniquement ceux dont la taille est supérieure à quarante kilo-bases avec un nombre de SNP de plus de cinquante, cinq cent quatre-vingt-trois contigs restent à être catégorisés.

Afin de rendre plus robuste la classification des contigs, nous avons exploité les trois critères de profondeur, de données manquantes et d'hétérozygotie : une matrice sur les contigs déjà catégorisés comprenant les dix premières CP de chacune des trois ACP de chaque critère est tout d'abord réalisée. Puis, une UMAP sur cette matrice est ensuite effectuée (cf. Figure A 4.4, Figure C.1 de l'annexe C et Figure D.1 de l'annexe D).

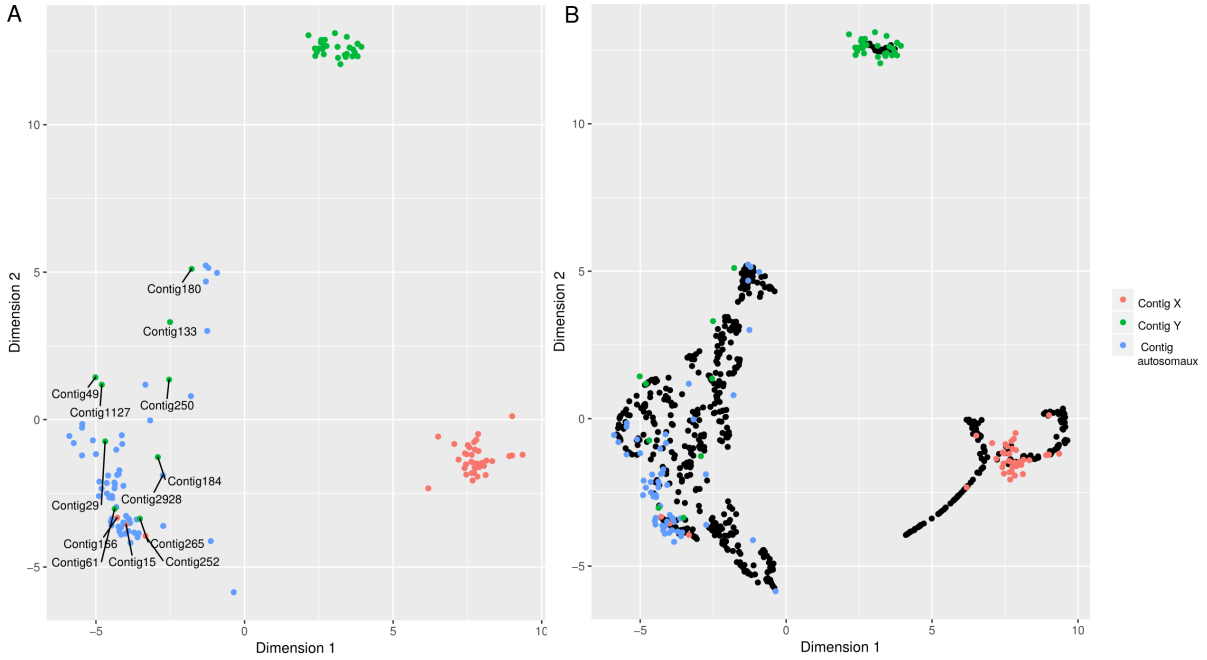

FIGURE 4.4 – UMAP mixte sur les 10ers CP des ACP de profondeur, d'hétérozygotie et des données manquantes A) des contigs connus, B) des contigs connus avec la projection des cinq cent quatre-vingt-trois contigs sélectionnés après filtres sur leur taille et leur nombre de SNP

Les trois critères permettent la distinction des trois groupes : X, Y et autosomaux qui comprend également les PAR. Les groupes sont tous constitués des contigs de leurs catégories.

Nous avons ensuite réalisé une projection des contigs non catégorisés (en noir) sur cette même UMAP (cf. Figure B 4.4). Les contigs projetés se positionnent dans les trois groupes préalablement constitués. La Table 4.3 attribue le nombre de contigs nouvellement identifiés aux différentes catégories.

TABLE 4.3 – Nombre de contigs par catégories

Catégories	**Nombre de contigs**
Contigs X	142
Contigs Y	32
Contigs autosomaux	409
Total	583

La Table E.1 en annexe E donne l'ensemble des noms des contigs nouvellement identifiés pour les catégories X, Y et autosomaux.

4.5 Nouvelles PAR identifiées

En élaborant les profils par catégories des contigs nouvellement catégorisés, certains contigs ont révélé des profils atypiques pour leurs catégories. Ces contigs sont considérés comme étant des PAR potentielles. Parmi eux, le contig numéro 33, appartenant aux groupes des X sur la UMAP réalisée sur les trois critères de profondeur, de données manquantes et d'hétérozygotie (cf. Table E.1 en annexe E), présente un profil anormal pour cette catégorie : la profondeur des femelles est moins du double de la profondeur des mâles.

Afin de confirmer la présence d'une région pseudo-autosomale dans ce contig, nous avons analysé la profondeur de celui-ci. La profondeur normalisée des mâles et des femelles a été prélevée par position tout le long du contig. En abscisse sont représentées les positions des bases du contig numéro 33 et en ordonnée la valeur de profondeur normalisée pour chaque position, en rouge pour les femelles, en bleu pour les mâles (cf. Figure A 4.5).

La profondeur des femelles du début du contig juqu'à cinq cent kilo-bases y est deux fois supérieure à celle des mâles comme attendu pour un contig X confirmant ainsi la catégorisation des contigs de la UMAP de la Figure B 4.4. La profondeur sur la seconde moitié du contig ne montre que très peu de différences entre mâles et femelles. Cette seconde moitié du contig 33 est une région commune aux mâles et aux femelles appelée PAR. Le contig 33 présente donc une région X spécifique ainsi qu'une PAR dont la limite est nettement visible autour de cinq cent kilo-bases. Parmi les contigs comprenant une PAR, huit ont présenté une limite nette entre une région sexuelle et une PAR dont six parmi ceux préalablement identifiés (cf. contig en rouge de la Table 4.2) et deux nouvellement catégorisés : le contig 33 et le contig 267.

Nous observons un grand pic de profondeur autour de quatre cent-cinquante kilo-bases aussi bien pour les mâles que pour les femelles. Afin de comprendre à quoi est dû ce pic, cette région est observée plus en détail à l'aide de l'outil IGV (cf. Figure B 4.5). Nous avons réalisé l'observation de cette région sur deux individus : une femelle sur la première ligne et un mâle sur la seconde ligne. Les bases de lecture correspondant à la référence sont affichées en gris. Les bases de lecture qui ne correspondent pas ont un code couleur : l'adénine (A) est représentée en vert, la cytosine (C) en bleu, la guanine (G) en jaune et la thymine (T) en rouge. Les insertions sont indiquées par un I violet. IGV révèle un grand nombre d'insertions sur la séquence d'intérêt. Une des explications à ce pic de profondeur est la présence de répétitions, comprenant un grand nombre d'insertions, qui entraîne un problème d'assemblage dans ce génome amélioré.

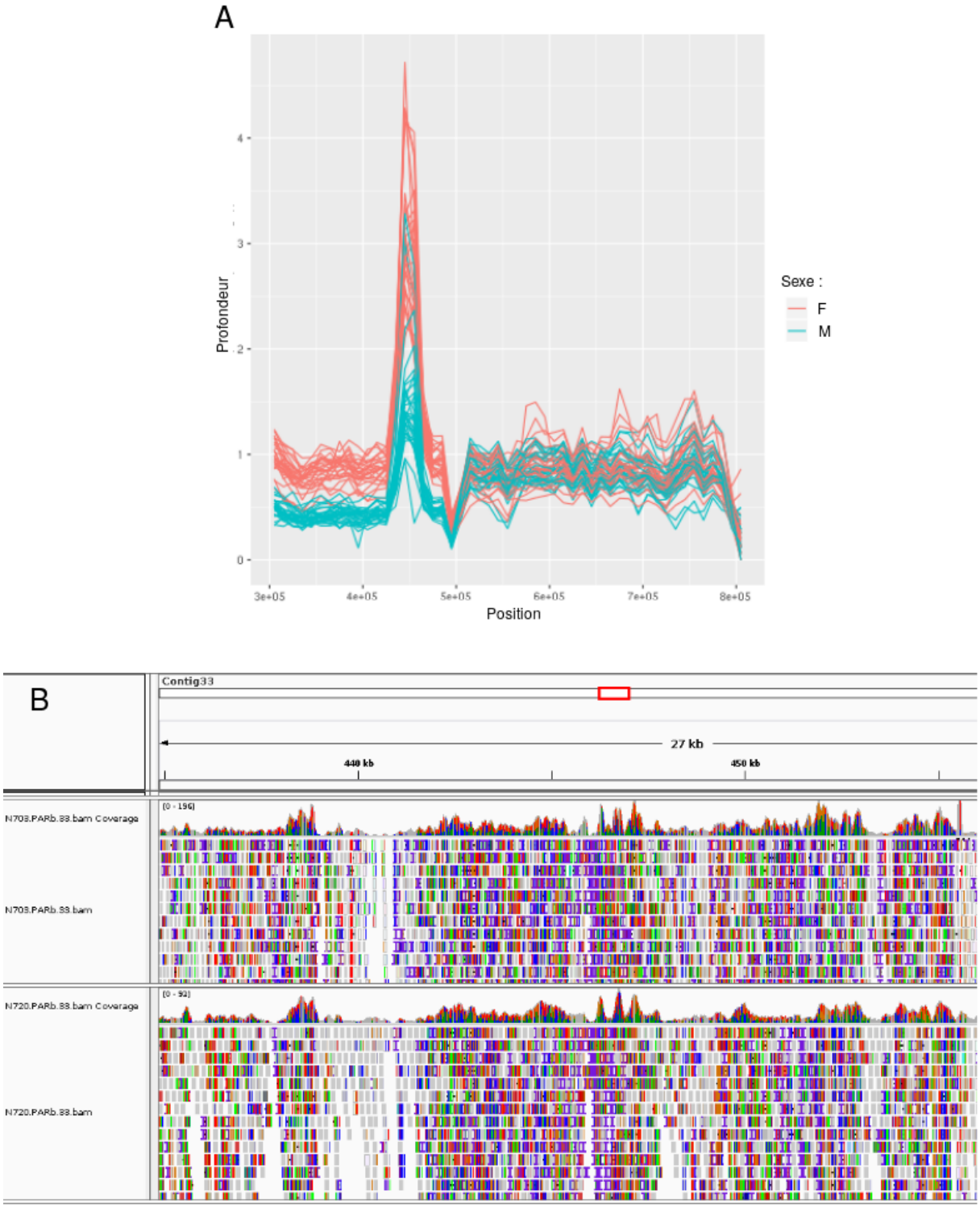

FIGURE 4.5 – A) Profondeur par position le long du contig 33 avec en rouge la profondeur des femelles et en bleu celle des mâles B) Visualisation de l'alignement de la région entre 435 kilo-bases et 455 kilo-bases du contig 33 avec le code couleur suivant pour les bases de lecture qui ne correspondent pas : l'adénine en vert, la cytosine en bleu, la guanine en jaune et la thymine en rouge. Les insertions sont indiquées par un I violet

5 Conclusion

Mon stage a permis de répondre aux objectifs fixés. L'élaboration de profils basés sur les critères d'hétérozygotie, de données manquantes et de profondeur pour chacune des catégories X, Y et autosomaux a été effectuée permettant la validation de l'attribution des contigs à leur catégorie pour certains et la ré-attribution pour cinq d'entre eux. Basés sur ces profils, douze contigs ont révélé la présence de PAR dont six avec une limite visible entre les régions X ou Y et la PAR. L'ACP réalisée sur la profondeur a pu valider l'assignation des contigs préalablement identifiés ainsi que les contigs possédant des PAR via la présence de trois «branches». Chaque branche de l'ACP correspond à une des trois catégories avec notamment l'appartenance des PAR au groupe des autosomaux. La UMAP mixte réalisée sur les trois critères d'hétérozygotie, de données manquantes et de profondeur a permis la classification de cinq cent quatre-vingt-trois contigs inconnus : cent-quarante-deux appartenant aux groupe des X, trente-deux au groupe des Y et quatre-cent-neuf au groupe des autosomaux. Les profils de ces contigs nouvellement catégorisés ont révélé des contigs possédant des PAR. Parmi ces contigs, le contig numéro 33 a été validé comme étant bien un contig sexuel constitué d'une PAR.

La mauvaise assignation des contigs préalablement effectuée est le reflet d'une mauvaise qualité du génome initial ayant servi a l'assignation. Bien que le génome de référence ait été considéré comme de meilleure qualité, le pic de profondeur du contig 33 révèlent encore un problème d'assemblage même dans un génome de meilleure qualité.

L'approche utilisée ici pour la caractérisation des contigs, basée sur l'hétérozygotie, les données manquantes et la profondeur, n'est pas uniquement valide pour le génome de l'ornithorynque. En effet, cette approche est applicable à l'ensemble des être vivants présentant le système de détermination sexuelle XY ou ZW.

La méthode utilisée ici peut être plus informative et précise en réalisant les profils des catégories X, Y et autosomaux par population : Barnard (North New South Wales (NNSW)), North QueensLanD (NQLD), Shoalhaven, Tasmanie et Wingecarribee (Central New South Wales (CNSW)).

La suite du travail sera de s'intéresser plus particulièrement aux contigs présentant des PAR afin d'étudier ces régions d'intérêt. L'ensemble du travail a pour but la compréhension de la méiose dans cet organisme. L'etude du fonctionnement de celle-ci permettra notamment la compréhension à plus grande echelle de la méiose et de l'évolution des chromosomes sexuels dans le vivant.

Bibliographie

[1] H. C. Martin, E. M. Batty, J. Hussin, P. Westall, T. Daish, S. Kolomyjec, P. Piazza, R. Bowden, M. Hawkins, T. Grant, C. Moritz, F. Grutzner, J. Gongora, and P. Donnelly, "Insights into platypus population structure and history from whole-genome sequencing," *Molecular Biology and Evolution*, vol. 35, no. 5, pp. 1238–1252, Mar. 2018. [En línea]. Disponible : https://doi.org/10.1093/molbev/msy041

[2] "Genome analysis of the platypus reveals unique signatures of evolution," *Nature*, vol. 453, no. 7192, pp. 175–183, May 2008. [En línea]. Disponible : https://doi.org/10.1038/nature06936

[3] "IV. a description of the anatomy of the ornithorhynchus paradoxus," *Philosophical Transactions of the Royal Society of London*, vol. 92, pp. 67–84, Jan. 1802. [En línea]. Disponible : https://doi.org/10.1098/rstl.1802.0006

[4] A. K. Enjapoori, T. R. Grant, S. C. Nicol, C. M. Lefèvre, K. R. Nicholas, and J. A. Sharp, "Monotreme lactation protein is highly expressed in monotreme milk and provides antimicrobial protection," *Genome Biology and Evolution*, vol. 6, no. 10, pp. 2754–2773, Sep. 2014. [En línea]. Disponible : https://doi.org/10.1093/gbe/evu209

[5] W. Rens, P. C. OBrien, F. Grutzner, O. Clarke, D. Graphodatskaya, E. Tsend-Ayush, V. A. Trifonov, H. Skelton, M. C. Wallis, S. Johnston, F. Veyrunes, J. A. Graves, and M. A. Ferguson-Smith, "The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian z," *Genome Biology*, vol. 8, no. 11, p. R243, 2007. [En línea]. Disponible : https://doi.org/10.1186/gb-2007-8-11-r243

[6] F. Veyrunes, P. D. Waters, P. Miethke, W. Rens, D. McMillan, A. E. Alsop, F. Grutzner, J. E. Deakin, C. M. Whittington, K. Schatzkamer, C. L. Kremitzki, T. Graves, M. A. Ferguson-Smith, W. Warren, and J. A. M. Graves, "Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes," *Genome Research*, vol. 18, no. 6, pp. 965–973, May 2008. [En línea]. Disponible : https://doi.org/10.1101/gr.7101908

[7] F. Grützner, W. Rens, E. Tsend-Ayush, N. El-Mogharbel, P. C. M. OBrien, R. C. Jones, M. A. Ferguson-Smith, and J. A. M. Graves, "In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird z and mammal x chromosomes," *Nature*, vol. 432, no. 7019, pp. 913–917, Oct. 2004. [En línea]. Disponible : https://doi.org/10.1038/nature03021

[8] J. A. Shendure, G. J. Porreca, G. M. Church, A. F. Gardner, C. L. Hendrickson, J. Kieleczawa, and B. E. Slatko, "Overview of DNA sequencing strategies," *Current Protocols in Molecular Biology*, vol. 96, no. 1, pp. 7.1.1–7.1.23, Oct. 2011. [En línea]. Disponible : https://doi.org/10.1002/0471142727.mb0701s96

[9] A. Rhoads and K. F. Au, "PacBio sequencing and its applications," *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 5, pp. 278–289, Oct. 2015. [En línea]. Disponible : https://doi.org/10.1016/j.gpb.2015.08.002

[10] A. Limasset, B. Cazaux, E. Rivals, and P. Peterlongo, "Read mapping on de bruijn graphs," *BMC Bioinformatics*, vol. 17, no. 1, Jun. 2016. [En línea]. Disponible : https://doi.org/10.1186/s12859-016-1103-9

[11] H. Li and R. Durbin, "Fast and accurate long-read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, Jan. 2010. [En línea]. Disponible : https://doi.org/10.1093/bioinformatics/btp698

[12] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. D. and, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Jun. 2011. [En línea]. Disponible : https://doi.org/10.1093/bioinformatics/btr330

[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. D. and, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Jun. 2009. [En línea]. Disponible : https://doi.org/10.1093/bioinformatics/btp352

[14] A. Rimmer, , H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. McVean, and G. Lunter, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications," *Nature Genetics*, vol. 46, no. 8, pp. 912–918, Jul. 2014. [En línea]. Disponible : https://doi.org/10.1038/ng.3036

[15] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage : key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, Jan. 2014. [En línea]. Disponible : https://doi.org/10.1038/nrg3642

[16] G. Lunter and M. Goodson, "Stampy : A statistical algorithm for sensitive and fast mapping of illumina sequence reads," *Genome Research*, vol. 21, no. 6, pp. 936–939, Oct. 2010. [En línea]. Disponible : https://doi.org/10.1101/gr.111120.110

[17] A. R. Quinlan, "BEDTools : The swiss-army tool for genome feature analysis," *Current Protocols in Bioinformatics*, vol. 47, no. 1, pp. 11.12.1–11.12.34, Sep. 2014. [En línea]. Disponible : https://doi.org/10.1002/0471250953.bi1112s47

[18] P. Danecek and S. A. McCarthy, "BCFtools/csq : haplotype-aware variant consequences," *Bioinformatics*, vol. 33, no. 13, pp. 2037–2039, Feb. 2017. [En línea]. Disponible : https://doi.org/10.1093/bioinformatics/btx100

[19] I. T. Jolliffe and J. Cadima, "Principal component analysis : a review and recent developments," *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016. [En línea]. Disponible : https://doi.org/10.1098/rsta.2015.0202

[20] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell

data using UMAP," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, Dec. 2018. [En línea]. Disponible : https://doi.org/10.1038/nbt.4314

[21] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov, "Integrative genomics viewer (IGV) : high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, Apr. 2012. [En línea]. Disponible : https://doi.org/10.1093/bib/bbs017

[22] K. Ito and D. Murphy, "Application of ggplot2 to pharmacometric graphics," *CPT : Pharmacometrics & Systems Pharmacology*, vol. 2, no. 10, p. e79, Oct. 2013. [En línea]. Disponible : https://doi.org/10.1038/psp.2013.56

[23] N. M. Roslin, L. Weili, A. D. Paterson, and L. J. Strug, "Quality control analysis of the 1000 genomes project omni2.5 genotypes," Sep. 2016. [En línea]. Disponible : https://doi.org/10.1101/078600

Annexes

A Participation à la journée de la recherche de l'ICM

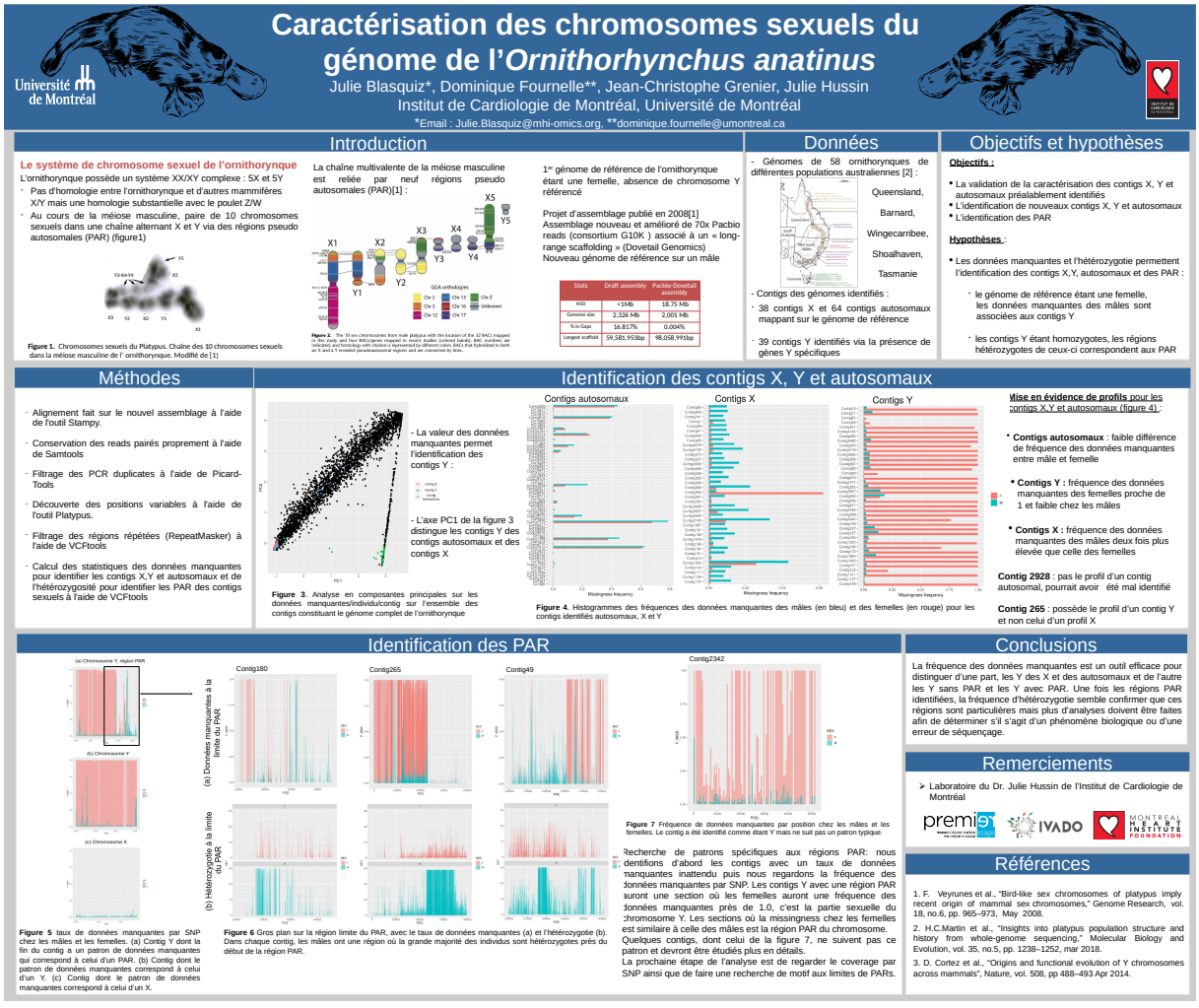

FIGURE A.1 – Poster présenté lors de la journée de la recherche de l'institut de cardiologie de Montréal

B Filtrage des contigs non catégorisés

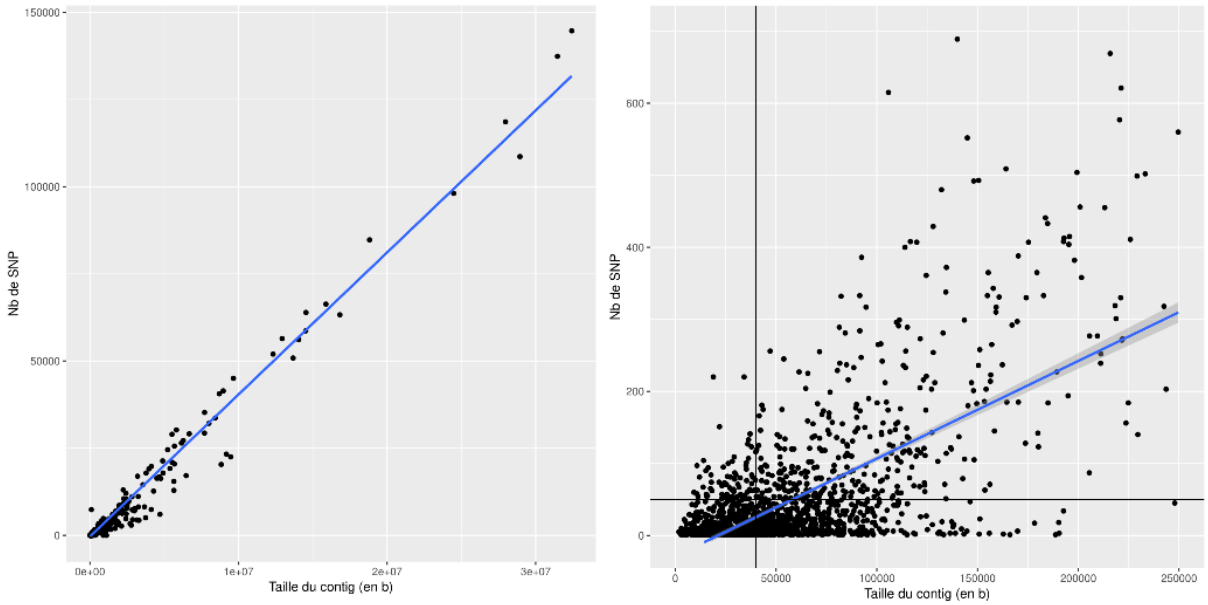

FIGURE B.1 – Sélection des contigs en fonction de leur taille et du nombre de SNP A) L'ensemble des contigs, B) Zoom sur le début du graphe

C ACP et UMAP sur les données de données manquantes

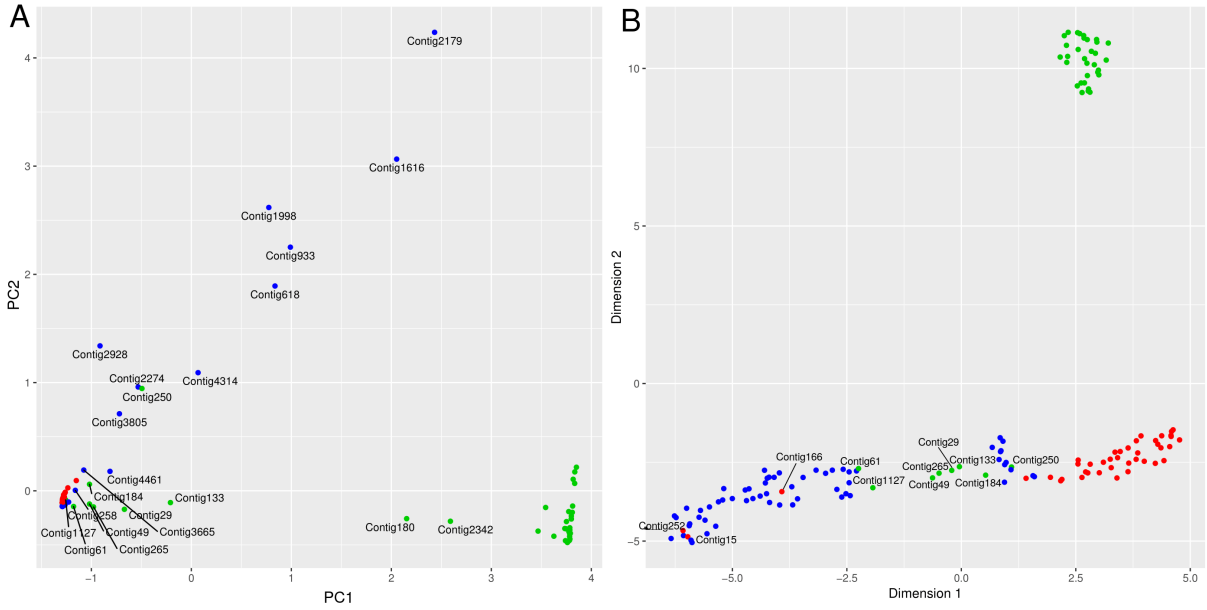

FIGURE C.1 – A) ACP sur les données de données manquantes des contigs identifiés B) UMAP sur les 10^{ers} CP de l'ACP des données de données manquantes des contigs identifiés

D ACP et UMAP sur l'hétérozygotie

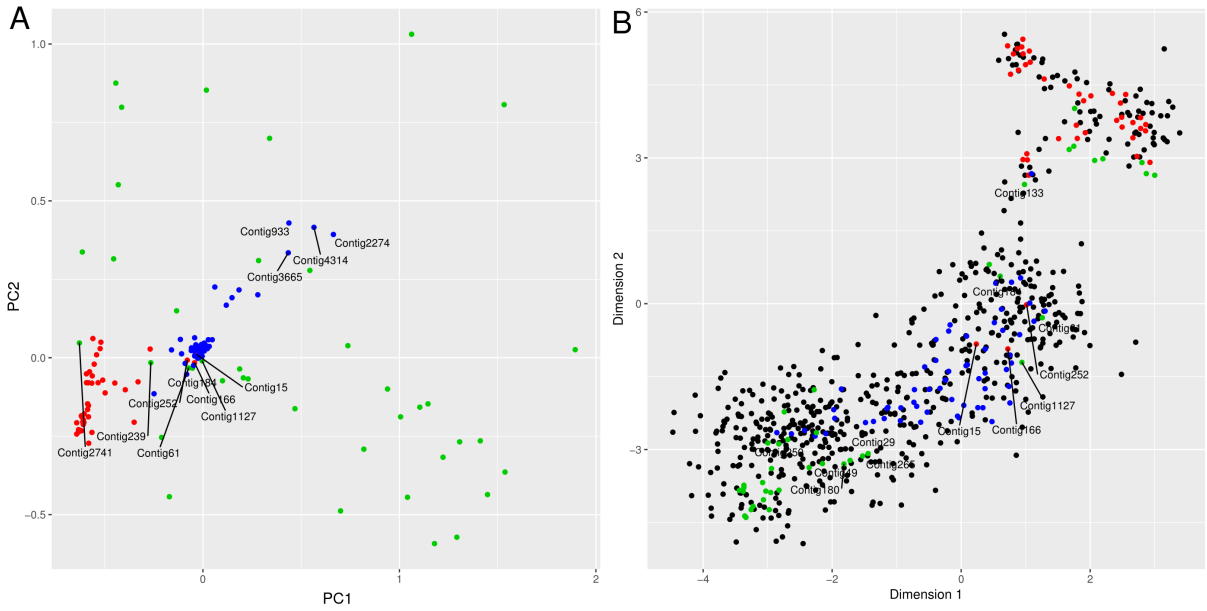

FIGURE D.1 – A) ACP sur l'hétérozygotie des contigs identifiés B) UMAP sur les 10ers CP de l'ACP de l'hétérozygotie des contigs identifiés

E Catégorisation des contigs inconnus

Table E.1 – Contigs par catégories

Contig	**Noms des contigs**
Contigs X (142 contigs)	Contig1076 Contig109 Contig1108 Contig1115 Contig114 Contig1159 Contig118 Contig123 Contig1240 Contig1256 Contig128 Contig1316 Contig1352 Contig1362 Contig136 Contig1372 Contig1393 Contig140 Contig1541 Contig1564 Contig1598 Contig1622 Contig1625 Contig163 Contig168 Contig1710 Contig171 Contig174 Contig175 Contig1807 Contig1819 Contig188 Contig1924 Contig194 Contig1951 Contig195 Contig2001 Contig214 Contig216 Contig2270 Contig2294 Contig2320 Contig233 Contig238 Contig2449 Contig245 Contig246 Contig249 Contig257 Contig259 Contig263 Contig269 Contig270 Contig275 Contig2820 Contig2823 Contig283 Contig288 Contig2899 Contig293 Contig3005 Contig3009 Contig304 Contig3079 Contig309 Contig311 Contig3154 Contig321 Contig327 Contig3298 Contig3337 Contig33 Contig3425 Contig3462 Contig3479 Contig3509 Contig3543 Contig354 Contig3591 Contig3617 Contig363 Contig3653 Contig370 Contig375 Contig376 Contig380 Contig381 Contig3845 Contig384 Contig386 Contig3874 Contig3905 Contig3915 Contig3967 Contig3990 Contig4024 Contig411 Contig4184 Contig43 Contig4440 Contig4472 Contig4495 Contig44 Contig50 Contig519 Contig574 Contig575 Contig584 Contig608 Contig63 Contig651 Contig685 Contig704 Contig779 Contig78 Contig80 Contig836 Contig86 Contig90 Contig947 Contig94 Contig952 Contig967 Contig96 Contig985 Contig987
Contigs Y (32 contigs)	Contig1022 Contig1028 Contig1135 Contig1360 Contig1468 Contig1488 Contig167 Contig2073 Contig2136 Contig2272 Contig229 Contig2651 Contig273 Contig280 Contig2851 Contig2966 Contig3086 Contig318 Contig3227 Contig3464 Contig365 Contig3763 Contig390 Contig4018 Contig4041 Contig4076 Contig4298 Contig4313 Contig4331 Contig591 Contig699 Contig903

TABLE E.1 – Contigs par catégories

Contigs	**Noms des contigs**
Contigs autosomaux (409 contigs)	Contig101 Contig1034 Contig1036 Contig106 Contig1072 Contig1086 Contig1092 Contig10 Contig1103 Contig112 Contig113 Contig1140 Contig1147 Contig115 Contig1165 Contig116 Contig1191 Contig119 Contig1204 Contig120 Contig121 Contig1226 Contig122 Contig1235 Contig1242 Contig1247 Contig1258 Contig125 Contig1262 Contig1267 Contig126 Contig129 Contig1306 Contig130 Contig131 Contig132 Contig1355 Contig135 Contig1370 Contig137 Contig138 Contig1397 Contig1410 Contig141 Contig142 Contig1432 Contig143 Contig1447 Contig144 Contig146 Contig1477 Contig1498 Contig149 Contig14 Contig150 Contig1511 Contig152 Contig153 Contig156 Contig1573 Contig157 Contig158 Contig159 Contig161 Contig162 Contig1635 Contig165 Contig1676 Contig1677 Contig1704 Contig1727 Contig1728 Contig1734 Contig173 Contig1761 Contig1765 Contig176 Contig177 Contig178 Contig1792 Contig17 Contig1828 Contig1831 Contig1834 Contig1837 Contig183 Contig1847 Contig1857 Contig185 Contig1866 Contig186 Contig1903 Contig190 Contig1910 Contig192 Contig1953 Contig198 Contig1995 Contig19 Contig200 Contig201 Contig2027 Contig202 Contig203 Contig2050 Contig205 Contig207 Contig20 Contig210 Contig211 Contig2141 Contig2174 Contig2175 Contig2178 Contig2186 Contig218 Contig219 Contig21 Contig220 Contig2210 Contig2216 Contig221 Contig224 Contig225 Contig227 Contig228 Contig2298 Contig2299 Contig2307 Contig230 Contig2320 Contig2327 Contig2329 Contig235 Contig2365 Contig2386 Contig2387 Contig2399 Contig23 Contig240 Contig2414 Contig2424 Contig243 Contig2443 Contig2473 Contig2496 Contig24 Contig251 Contig253 Contig2550 Contig2567 Contig256 Contig2583 Contig260 Contig2628 Contig2659 Contig266 Contig2678 Contig2679 Contig267 Contig268 Contig2709 Contig2719 Contig2742 Contig274 Contig2752 Contig2781 Contig278 Contig27 Contig282 Contig2844 Contig2854 Contig285 Contig2872 Contig287 Contig2887 Contig289 Contig28 Contig2910 Contig292 Contig294 Contig295 Contig296 Contig2975 Contig297Contig2982 Contig2984 Contig298 Contig2998 Contig299 Contig3001 Contig3008 Contig3023 Contig303 Contig3069 Contig307 Contig3084 Contig3094 Contig3105 Contig3123 Contig3125 Contig3133

TABLE E.1 – Contigs par catégories

Contigs	**Noms des contigs**
Contig autosomaux (409 contigs)	Contig313 Contig3145 Contig314 Contig316 Contig317 Contig3196 Contig319 Contig3201 Contig320 Contig3224 Contig322 Contig3231 Contig3232 Contig3244 Contig325 Contig3267 Contig3269 Contig3270 Contig3275 Contig328 Contig3301 Contig330 Contig331 Contig332 Contig3358 Contig336 Contig3371 Contig338 Contig3400 Contig3405 Contig3411 Contig341 Contig3431 Contig343 Contig344 Contig345 Contig346 Contig3472 Contig347 Contig348 Contig349 Contig34 Contig3507 Contig3508 Contig350 Contig3518 Contig3521 Contig3522 Contig3523 Contig3553 Contig357 Contig3586 Contig359 Contig360 Contig3614 Contig362 Contig3636 Contig3646 Contig3671 Contig367 Contig368 Contig3693 Contig36 Contig3702 Contig3703 Contig371 Contig372 Contig3731 Contig3737 Contig3749 Contig374 Contig3761 Contig377 Contig378 Contig3791 Contig3793 Contig37 Contig3801 Contig3804 Contig3813 Contig3817 Contig3822 Contig3826 Contig382 Contig3835 Contig3847 Contig3854 Contig385 Contig3870 Contig3871 Contig387 Contig3882 Contig38 Contig3906 Contig3909 Contig3910 Contig3923 Contig3925 Contig3943 Contig3960 Contig3968 Contig39 Contig3 Contig4012 Contig4020 Contig403 Contig4042 Contig4071 Contig4081 Contig4094 Contig40 Contig4111 Contig4136 Contig4147 Contig4150 Contig4174 Contig4190 Contig41 Contig4201 Contig4204 Contig4265 Contig4276 Contig42 Contig4332 Contig4342 Contig4344 Contig4361 Contig4368 Contig4384 Contig4401 Contig4450 Contig4473 Contig4477 Contig4497 Contig4499 Contig4500 Contig4505 Contig4506 Contig4553 Contig4557 Contig456 Contig45 Contig488 Contig48 Contig491 Contig495 Contig504 Contig51 Contig53 Contig54 Contig557 Contig56 Contig57 Contig58 Contig590 Contig5 Contig604 Contig62 Contig633 Contig642 Contig644 Contig648 Contig64 Contig65 Contig66 Contig672 Contig67 Contig68 Contig690 Contig693 Contig697 Contig698 Contig69 Contig6 Contig709 Contig722 Contig72 Contig73 Contig75 Contig79 Contig813 Contig81 Contig820 Contig823 Contig82 Contig83 Contig84 Contig857 Contig85 Contig878 Contig87 Contig881 Contig882 Contig88 Contig89 Contig8 Contig913 Contig919 Contig91 Contig924 Contig92 Contig93 Contig977 Contig97 Contig98 Contig991 Contig9