AIX-MARSEILLE UNIVERSITÉ

MASTER BIO-INFORMATIQUE

Parcours type : Développement Logiciel et Analyse des Données

2023-2024

Etude de la longévité Canine : de la collecte d'échantillons Cani-DNA à la détection de Variations Structurales

---

Canine Longevity : From Cani-DNA sample collection to Structural Variation Detection

Numéro étudiant : 21225791

Equipe : Génétique du Chien

IGDR : Institut de Génétique et Développement de Rennes

CNRS- Université de Rennes- ERL Inserm

Université de Rennes igdr A Canine Genetics Team cnrs Faculté des Sciences Aix*Marseille Université

REMERCIEMENT

Pour commencer, je voudrais remercier Catherine ANDRÉ, responsable de l'équipe "Génétique du chien" à l'IGDR, pour m'avoir accueillie dans son équipe et m'avoir permis de réaliser ce stage.

Je tiens aussi à remercier mes superviseurs, Richard GUYON et Thomas DERRIEN, pour m'avoir encadrée et soutenue tout au long de ce stage.

Je remercie également toutes les personnes qui m'ont apporté leur aide au cours de ce travail : Victor et Louis pour leurs conseils en bioinformatique, Edouard pour son assistance dans la réparation de l'ordinateur, et Armel pour ses explications concernant l'utilisation de l'outil Jotform.

Pour finir, je remercie l'ensemble de l'équipe "Génétique du Chien" pour leur gentillesse, leur soutien et leur accueil chaleureux tout au long de mon stage.

Je tiens à remercier également Géraldine et Sylvain, pour la mise en place de ma convention de stage.

RÉSUMÉ

Le projet Genomic of Longevity in Dogs (GOLDogs) vise à explorer les bases génétiques de la longévité chez les chiens, en identifiant les variations génomiques associées à la différence d'espérance de vie entre les races des chiens. Ce rapport présente le travail réalisé lors d'un stage de six mois au sein de l'équipe "Génétique du Chien" de l'Institut de Génétique et de Développement de Rennes. Le stage avait deux objectifs principaux : Le premier est de faciliter et améliorer la collecte et l'enregistrement des données des échantillons canins dans le Centre de Ressources Biologiques Cani-DNA. Pour l'atteindre, un formulaire en ligne a été créé, améliorant ainsi la collecte des informations dans la base de données Cani-DNA, leur vérification et optimisant leur enregistrement . Le deuxième objectif est d'effectuer une analyse bioinformatique préliminaire des données de séquençage longues lectures pour le projet GOLDogs. Pour cet objectif, le pipeline NanoSeq, implémenté avec Nextflow, a été utilisé pour analyser ce type de données. Minimap2 a été choisi comme aligneur, et DeepVariant a été utilisé pour la détection des variants génomiques courts. En ce qui concerne la détection des variants structuraux (SVs), l'outil CuteSV a été privilégié en raison de sa rapidité et de sa sensibilité accrues par rapport à Sniffles. L'analyse des SVs en communs a démontré que Bedtools identifie plus de variants communs grâce à des critères de chevauchement flexibles, tandis que Bcftools utilise des critères stricts pour une plus grande précision de séquence. Il a été observé que le type de séquenceur, la profondeur de séquençage et la méthodologie d'extraction de l'ADN influencent significativement les résultats de l'analyse des variants structuraux. Le séquençage avec PromethION est préférable pour sa capacité à produire des lectures de haute qualité et à large couverture. Ces observations soulignent l'importance de choisir des techniques adaptées pour obtenir des résultats précis et fiables pour la suite du projet GOLDogs.

Mots clés : Longévité, Génome, Séquençage Nanopore, Variants Génomiques et structuraux.

ABSTRACT

The Genomic of Longevity in Dogs (GOLDogs) project aims to explore the genetic basis of longevity in dogs, by identifying the genomic variations associated with the difference in lifespan between dog breeds. This report presents the work carried out during a six-month internship in the 'Dog Genetics' team at the Institut de Génétique et Développement de Rennes. The first goal was to facilitate and improve the collection and storage of data from canine samples. To achieve this, an online form was created, improving the collection of information in the Cani-DNA database. The second objective is to carry out a preliminary bioinformatic analysis of long-read sequencing data for the GOLDogs project. For this purpose, the NanoSeq pipeline, implemented with Nextflow, was used to analyze this type of data. Minimap2 was chosen as an aligner, and DeepVariant was used for detection of short genomic variants. For the detection of structural variants (SVs), the CuteSV tool was chosen for its increased speed and sensitivity compared with Sniffles. Analysis of SVs in common showed that Bedtools identifies more common variants thanks to flexible overlap criteria, while Bcftools uses strict criteria for greater sequence accuracy. Sequencer device, sequencing depth and DNA extraction methodology were found to significantly influence the results of structural variant analysis. Sequencing with PromethION is preferred for its ability to produce high-quality, broad-coverage reads. These observations underline the importance of choosing suitable techniques to obtain accurate and reliable results in genomic studies.

Key words : Longevity, Nanopore Sequencing, Genomic Variations & Structural variants.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

ABBREVIATIONS/ACRONYMES

BAM/SAM	Binary/Sequence Alignment Map
BND	Break-end
CIGAR	Compact Idiosyncratic Gapped Alignment Report
CRB	Centre de Ressources Biologiques
DEL	Deletion
DUP	Duplication
GOLDogs	Genomic of Longevity in Dogs
GWAS	Genome-Wide Association Study
IGF1	Insulin-like Growth Factor-1
INS	Insertion
INV	Inversion
Kb	Kilobases
LINEs	Long Interspersed Nuclear Elements
ONT	Oxford Nanopore Technologies
pb	Paire de bases
QC	Quality Control
SINEs	Short Interspersed Nuclear Elements
SNP/SNV	Single Nucleotide Polymorphisms/Variant
SVs	Structural Variants
VCF	Variant Call Format
VS code	Visual Studio Code
WGS	Whole Genome Sequencing

1. INTRODUCTION

1.1 Présentation de l'équipe et des projets

Mon stage s'est déroulé au sein de l'équipe "Génétique du Chien" de l'Institut de Génétique et de Développement de Rennes (IGDR CNRS-Université de Rennes). Cette équipe se concentre sur le modèle spontané du chien, pour étudier les composantes génétiques des maladies rares et/ou complexes chez l'être humain, comme les cancers, des maladies génétiques neurosensorielles, dermatologiques,. Etant donné que les chiens ont des maladies similaires/homologues à l'homme et qu'ils partagent le même environnement, ils représentent un modèle unique pour étudier ces maladies. De plus, la structure de la population canine est composée de plus de 400 races officiellement reconnues qui représentent autant d'isolats génétiques caractérisés par une forte homogénéité intra-race et une forte hétérogénéité inter-races. Cette caractéristique unique dûe à la sélection exercée par l'homme depuis la domestication du chien facilite l'analyse des relations génotype/phénotype et rend le modèle canin particulièrement intéressant pour étudier les composantes génétiques ou épi-génétiques de maladies ou de traits d'intérêt tels que la longévité (le poids, la taille, la morphologie, la couleur des robes)[1]. Pour ce faire, l'équipe Génétique du chien a mis en place le Centre de Ressources Biologiques (CRB) Cani-DNA , une biobanque unique en France, en collaboration avec le laboratoire Antagene, les quatre Écoles Nationales Vétérinaires, et un réseau national de vétérinaires praticiens.

Le CRB Cani-DNA rassemble des échantillons de près de 23,000 chiens de toutes races, qu'ils soient atteints de maladies génétiques ou non, accompagnés de leurs données généalogiques, phénotypiques et cliniques. Ces prélèvements (essentiellement sanguins), collectés avec l'autorisation des propriétaires, par des vétérinaires praticiens lors de consultations, constituent une ressource précieuse pour la recherche scientifique, avec un panel représentatif de plus de 300 races de chiens et plus d'une centaine de modèles de maladies génétiques homologues à celles de l'humain. Les ADN et ARN ainsi extraits de ces prélèvements sont conservés et mis à disposition des chercheurs dans le cadre des projets de recherche biomédicale. Les collections du CRB Cani-DNA ont ainsi été constituées au fil des années (> 20 ans) grâce à la volonté des propriétaires, éleveurs et vétérinaires impliqués dans une démarche de science participative inscrite dans le cadre du parcours de soin du chien.

1.2 Introduction au projet GOLDogs

Le projet Genomic of Longevity in Dogs (GOLDogs) vise à explorer les bases génétiques de la longévité chez les chiens, en identifiant les variations génomiques associées à la différence de durée de vie entre les races de chiens. Plus généralement, le projet GOLDogs a pour objectif de produire un catalogue exhaustif des variations génomiques canines des races les plus représentées.

L'idée de ce projet venait de la relation inverse entre la taille des chiens et leur longévité, les races de petite taille vivent généralement plus longtemps que les races de grande taille[1]. Par exemple, les Chihuahuas ont une espérance de vie relativement longue entre 12 et 20 ans, tandis que les Bergers Allemands ont une espérance de vie moyenne de 9-13 ans[2].

Pour identifier les régions génomiques associées à la différence de longévité canine, des études d'association (GWAS ou Genome Wide Association Studies) sont combinées à des séquençage de génome complets de 100 chiens appartenant à 25 races. L'approche GWAS devrait permettre d'identifier des corrélations entre les différences de longévité et des SNPs (Single Nucleotide Polymorphisms) dont les fréquences alléliques diffèrent dans les populations cas/contrôle, ici jeunes versus âgées. Parallèlement, le séquençage complet de génome (WGS ou Whole Genome Sequencing) permet l'identification des variations génomiques plus complexes impactant la longévité canine.

Un haplotype spécifique du gène IGF1 (Insulin-like Growth Factor 1) est un déterminant majeur de la petite taille chez les chiens. Cet haplotype est commun à toutes les petites races et presque absent chez les races géantes[3]. L'étude menée par Plassais et *al.*,(2022) a identifié un variant unique dans un ARN antisens non codant qui interagit avec le gène IGF1. Étant donné la relation inverse entre taille et longévité, il est suggéré que les variations de ce gène IGF1 pourraient également influencer la longévité.

1.3 Les variations génomiques et leur importance entre les différentes races canines

Les variations génomiques ont un rôle important dans la diversité des races canines. Ces variations peuvent influencer la taille, la morphologie, la prédisposition à des maladies et la longévité des chiens. Ces changements sont divers, allant des mutations ponctuelles et petites insertions/délétions (InDels) à des variations structurales plus complexes. Ils peuvent être classés en trois catégories principales basées sur leur taille. La première catégorie

comprend les SNVs ou SNPs, qui sont des mutations ne touchant qu'un seul nucléotide. La deuxième catégorie englobe les variants courts (short variants), tels que les InDels, petites insertions/délétions de taille inférieure à 50 nucléotides. Enfin, la troisième concerne les variants structuraux, dont la taille dépasse 50 nucléotides à plus d'un million de paires de bases. Jusqu'à récemment, la recherche de variations génomiques s'est principalement focalisée sur les variations d'un seul nucléotide (SNV)[4,5] et les InDels, car elles sont plus faciles à identifier avec les technologies de séquençage de 2nd génération produisant des fragments d'ADN courts et algorithmes de séquençage actuels.[6] Cependant, le développement récents des technologies de séquençage de 3eme génération (cf. paragraphe 1.4) a mis en évidence la présence d'un plus grand nombre de variations structurales (SVs) et leur importance dans les variations génétiques inter-individuelles en complément des SNVs et des InDels courts. Par exemple, Halo *et al*. (2021) ont récemment montré que le nombre de variants structuraux entre 2 races de chiens (boxer et dogue allemand) était très important avec plus de 16 000 délétions et 15 000 insertions. De plus, la majorité de ces SVs étaient liés à la présence retrotransposons polymorphiques (*e.g.* Short interspersed retrotransposable elements (SINEs) de taille ~200nt, et Long interspersed retrotransposable elements (LINEs) de taille ~6000nt) (Fig.1). Il n'est donc pas étonnant que ces variations structurales, au delà des SNVs, jouent un rôle majeur dans la diversité génétique et phénotypique entre les races, ainsi que dans la prédisposition à certaines maladies.[7]

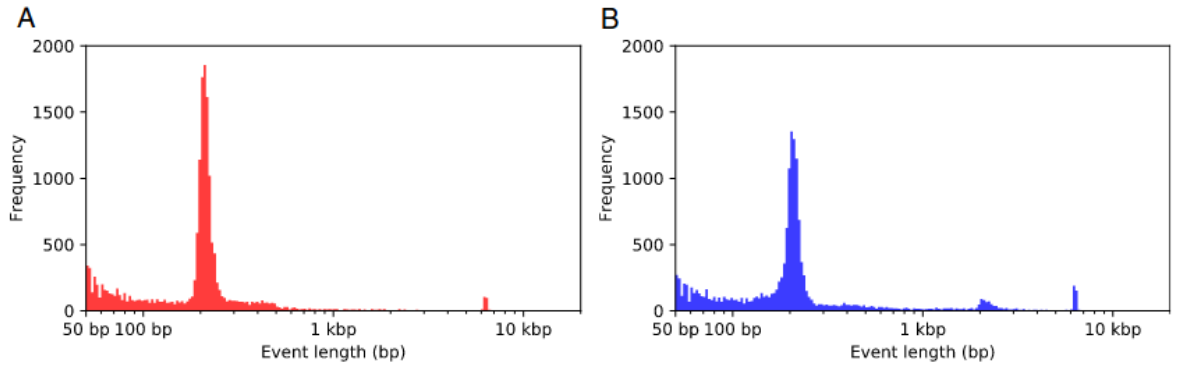

Figure 1. Distribution des tailles des délétions (A) et des insertions (B) entre un Boxer et un Dogue Allemand [7].

Les variants structuraux eux-mêmes comprennent plusieurs types de variants (Fig.2) : les insertions (INS), qui correspondent à l'ajout de fragments nucléotidiques dans l'ADN d'un échantillon par rapport à un génome de référence; les délétions (DEL), qui impliquent la suppression de nucléotides par rapport à la séquence de référence; les duplications (DUP), où des segments d'ADN sont dupliqués, entraînant des copies supplémentaires; les breakends

(BND), qui sont des points de cassure dans l'ADN où une séquence se termine brusquement; et les inversions (INV), qui se caractérisent par le renversement de l'orientation d'un segment d'ADN par rapport à la séquence de référence.

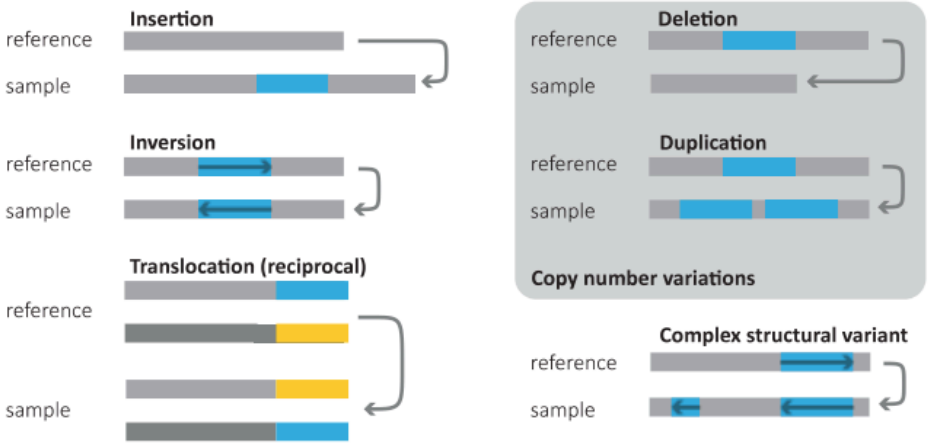

Figure 2. Visualisation des différentes variations structurales (SV) du génome[21].

1.4 Les technologies de séquençage longue lecture

Au cours de la dernière décennie, les technologies de séquençage d'ADN à lecture longue (long-read sequencing) sont devenues de puissants acteurs de la génomique[8]. Elles ont la capacité de générer des lectures de plusieurs centaines de kilobases avec une précision proche de celle des technologies de séquençage à lecture courte (short-read sequencing). Cette longueur de lecture très élevée permet de réduire les lacunes (gaps) dans l'assemblage du génome, en résolvant certaines régions génomiques complexes et répétitives qui posent souvent un problème avec les lectures courtes[9]. De plus, ces longues lectures facilitent grandement la détection et la caractérisation des variants structuraux de grande taille et des régions complexes auparavant inaccessibles tels que les télomères et les centromères. Ces réarrangements complexes sont souvent mal résolus ou manqués avec les lectures courtes en raison du manque de contexte suffisant pour identifier leurs points de cassure (ou breakpoint). Les longues lectures permettent donc de détecter ces variants de façon plus précise en couvrant la totalité des régions réarrangées et en fournissant les séquences exactes au niveau des points de cassure[10].

Il existe deux principales technologies de séquençage de longues lectures, commercialisées par les sociétés PacBio (Pacific Biosciences) et ONT (Oxford Nanopore

Technologies). La technologie proposée par ONT est basée sur la détection d'un courant ionique différentiel lors du passage d'une molécule d'ADN (ou d'ARN) unique à travers un nanopore biologique ancré dans une membrane synthétique[11]. Cette technologie a été développée sur plusieurs séquenceurs différents, notamment le MinION, le GridION et le PromethION. Le MinION et le GridION permettent de produire de 10 à 20 Gb de données par puce (flow cell= unité de séquençage), alors que le PromethION offre un rendement cinq fois supérieur, produisant entre 50 et 100 Gb par flow cell.

1.5 Contexte du stage

Pour ses projets de recherche, l'équipe génétique de chien suit un workflow général, commençant par la collecte des échantillons de chiens et leur enregistrement dans la base de données Cani-DNA puis leur stockage à -20°C ou -80°C dans les équipements dédiés. Après la mise à disposition des échantillons vient la caractérisation moléculaire qui englobe le séquençage et/ou le génotypage des échantillons, suivie par l'analyse bio-informatique pour l'identification des variations génomiques liées au(x) trait(s) d'intérêt et enfin, la validation fonctionnelle.

Dans le cadre de mon stage, j'ai travaillé sur deux axes principaux :

I. Collecte et enregistrement des données dans Cani-DNA

   Mon premier objectif était de faciliter la collecte des données d'identification associées aux échantillons à travers la mise en place d'un formulaire en ligne complémentaire du formulaire papier (cf.Annexe.1) utilisé actuellement en routine pour le recueil, le traitement et l'enregistrement des données dans la base Cani-DNA. Cela inclut également la mise en place d'un processus de vérification des données saisies dans le formulaire.

II. Analyse bioinformatique sur le projet "Longévité"

   Mon second objectif constituait le point de départ du projet GOLDogs. Il s'agissait de réaliser une analyse préliminaire des données de séquençage longue lecture issues de ce projet et de comparer les performances de différents outils bioinformatiques dédiés à l'annotation de variations structurales (SV).

Les résultats de ces analyses préliminaires servent à tester et valider les protocoles d'extraction et d'analyse bioinformatique les plus efficaces. Elles constituent une base essentielle pour le projet GOLDogs sur un petit nombre d'échantillons et seront utilisées pour déterminer les méthodes optimales à appliquer à un ensemble plus large de 100 échantillons.

2. MATERIELS ET METHODES

Durant mon stage, la diffusion des codes et des programmes utilisés ainsi que des versions se faisait via un GitLab privé de l'équipe : GitLab Bioinfog. Les scripts qui ne contiennent pas d'informations confidentielles sont également disponibles sur mon GitHub : Objectif 1, Objectif 2.

Objectif 1: Mise en place d'un formulaire en ligne

1. Utilisation de l'interface JotForm pour la création du formulaire en ligne

L'une des tâches qui m'a été assignée consistait à améliorer le processus de collecte de données associées à un prélèvement en créant un formulaire en ligne à destination des propriétaires, éleveurs ou vétérinaires qui envoient des échantillons au CRB Cani-DNA. Ce formulaire actuellement sous forme papier est crucial pour recueillir les informations d'identité nécessaires sur les vétérinaires, les propriétaires de chiens et les chiens eux-mêmes, afin de faciliter la collecte et l'enregistrement des données dans Cani-DNA. Pour accomplir cette tâche, j'ai utilisé, la plateforme en ligne JotForm permettant de créer des formulaires personnalisés.

2. Description du processus de création et de personnalisation du formulaire JotForm

J'ai conçu le formulaire en ligne avec trois sections principales : les informations sur le vétérinaire, les informations sur le propriétaire du chien et les informations sur le chien. Chacune de ces sections contient des champs spécifiques pour collecter les détails d'identification, tels que les noms, prénoms, adresses e-mail, numéros de téléphone, etc.

La conception du formulaire a été soigneusement pensée pour rendre la saisie des données aussi conviviale et intuitive que possible pour les utilisateurs, qu'il s'agisse de

vétérinaires ou de propriétaires de chiens. Des instructions claires ont été fournies pour guider les utilisateurs tout au long du processus de remplissage du formulaire (cf.Résultats).

Une fois le formulaire rempli en ligne, j'ai configuré JotForm pour que les réponses soient automatiquement enregistrées dans un fichier Excel (`jotform_output.xlsx`), ce qui facilite la gestion et le traitement ultérieurs des données par le personnel dédié du CRB Cani-DNA. De plus, des fonctionnalités de validation des données ont été mises en place pour déceler d'éventuelles erreurs lors de la saisie des informations par les opérateurs.

3. Le langage de programmation Python pour la gestion des données issues du formulaire

J'ai développé deux scripts en utilisant Python (version 3.10).

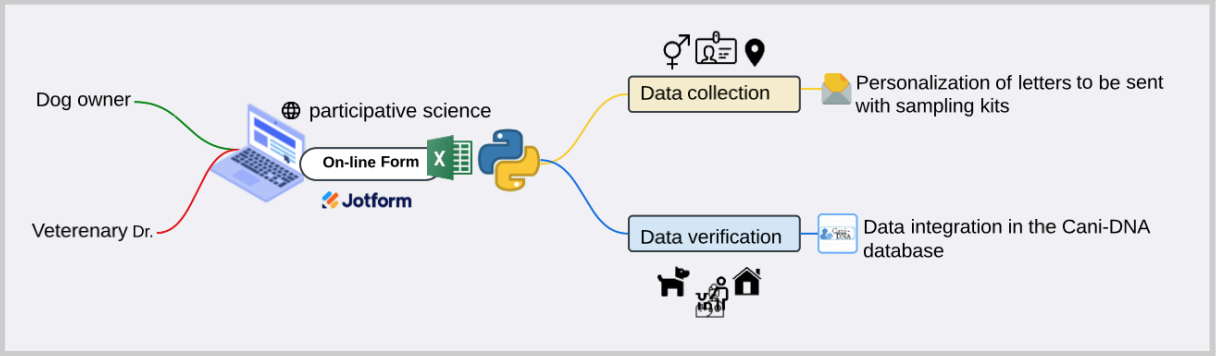

Figure 3. Les étapes clés pour atteindre l'objectif 1 : Mise en place d'un formulaire en ligne, automatisation de la collecte des données et vérification en vue de leur intégration dans Cani-DNA.

Le premier script (`letters_adaptor.py`) sert à formater les données collectées à partir du formulaire en ligne afin d'automatiser l'impression des adresses des destinataires des kits de prélèvements et la création de lettres personnalisées.

Il a été conçu pour récupérer les données pertinentes, telles que la civilité, le nom, le prénom, l'adresse et l'établissement du propriétaire ou du vétérinaire, à partir du fichier Excel généré par JotForm. Ces informations sont ensuite utilisées pour créer des lettres personnalisées adressées à chaque propriétaire ou vétérinaire par publi-postage à partir d'un fichier type.

Les lettres personnalisées sont un élément important du processus, car elles sont envoyées aux propriétaires ou vétérinaires avec les kits de prélèvement. Ces lettres

permettent de préciser aux destinataires le projet de recherche auquel le prélèvement de leur chien contribuera ainsi que les modalités de prélèvement et d'acheminement des prélèvements. La récupération des données d'identification permettra également d'automatiser leur enregistrement dans la base de données Cani-DNA à réception des échantillons.

Le deuxième script a été conçu pour vérifier les données entrées dans le formulaire, ainsi que pour assurer un contrôle qualité rigoureux avant l'enregistrement des données dans la base de données Cani-DNA.

4. Vérification des données du formulaire

Le script effectue une série de vérifications sur les données entrées dans le formulaire en ligne, contenant des informations soumises via des formulaires et une base de données sous Excel. Il commence par importer les librairies nécessaires et utilise la fonction `install_packages()` pour s'assurer que tous les packages requis sont installés, sinon il les installe.

Une fonction `check_existence_and_similarity()` vérifie si les noms et prénoms des vétérinaires et propriétaires dans le formulaire (fichier JotForm) existent déjà dans la base de données, et traite également les cas où les noms et prénoms pourraient être inversés.

Un processus de standardisation des données a également été appliqué. Ce processus utilise le package `"unidecode"` pour normaliser les chaînes de caractères en supprimant les accents et en convertissant les caractères en minuscules tout en prenant en considération les caractères spéciaux. Cette étape de standardisation garantit que les comparaisons ne sont pas affectées par les variations de formatage ou les différences d'encodage entre le formulaire et la base de données. Le script effectue également une vérification spécifique pour chaque animal, notamment l'existence de l'animal dans la base de données en comparant le nom usuel, la date de naissance, la puce et ou son tatouage, afin de ne pas entrer 2 fois le même chien.

En ce qui concerne les numéros de puce des chiens, une vérification spécifique a été ajoutée pour s'assurer que le nombre de caractères correspond au format standard de 15

chiffres. De même, pour les affixes (les élevages), en plus de les standardiser pour une comparaison précise, une autre vérification a été appliquée à l'aide du package "`difflib`". Cette vérification utilise `SequenceMatcher()` pour comparer les affixes (noms des élevages) du formulaire avec la base de données. Ce processus fonctionne sur le même principe qu'un alignement de séquences nucléotidiques et calcule un pourcentage d'identité entre chaînes de caractères. Ainsi, même en cas de légères variations dans l'orthographe ou si l'utilisateur commet des erreurs de frappe, le script peut identifier des similarités potentielles entre les affixes et signaler ces correspondances dans le fichier (`.log`) généré. En cas d'erreur ou de divergence entre les données du formulaire et celles de la base de données, le fichier (`.log`) alerte le personnel chargé de l'enregistrement des données.

Objectif 2: Analyse des données de séquençage longue lecture pour le projet GOLDogs

1. Données expérimentales

Les échantillons d'ADN canins utilisés dans cette étude ont été soumis à un séquençage longs fragments par France Génomique, l'infrastructure nationale pour les grands projets de génomique en France, située à Évry. Deux ensembles distincts (batch) ont été produits par France Génomique. La profondeur de séquençage est exprimée en X, représente le nombre moyen de lectures uniques couvrant le génome canin.

1.1 Premier lot/batch

Le premier ensemble comprend des séquences à faible profondeur du génome entier (low-pass WGS Whole Genome Sequencing) provenant de six chiens. Au total, il y a 12 échantillons, chaque chien étant représenté par deux échantillons distincts (réplicats), identifiés par des barcodes (AA-1/AA-2, AB-1/AB-2..., etc). Ces échantillons sont issus de races de chiens différentes et ont été séquencés sur la plateforme GridIon d'Oxford Nanopore Technologies. Chaque chien est associé à un identifiant spécifique, utilisé pour référencer les données (Tab.1).

L'extraction d'ADN a été réalisée avec le protocole Circulomics CBB, tandis que la préparation des banques d'ADN a suivi le protocole 1D Native barcoding DNA, avec l'utilisation des kits spécifiques EXP-NBD 104 et SQK-LSK109 (ONT). Ces kits assurent un

étiquetage (barcoding) précis des échantillons en vue du séquençage. Pour la purification de l'ADN et l'élimination des fragments de petites tailles, le kit standard Circulomics SRE a été utilisé pour la majorité des échantillons.

1.2 Deuxième lot/batch

Le deuxième batch comprend des séquences provenant de trois chiens, avec un total de six échantillons. Parmi ces échantillons, quatre réplicats techniques ont été réalisés pour un même chien afin de tester différents protocoles d'extraction d'ADN et de déterminer celui offrant le meilleur rendement. Ces échantillons ont été séquencés sur l'appareil PromethION et seulement un échantillon le BB-1 a été séquencé avec le séquenceur GridION. Les échantillons BB-2 et BB-3 ont bénéficié d'un protocole spécifique pour l'extraction d'ADN, comprenant Circulomics CBB + SRE + colonne Qiagen.

Tableau 1. Informations des échantillons d'ADN canins, incluant les identifiants, les races, les quantités produites et les lots.

ID Échantillon	Race	Quantité (Go)	Batch/Lot
AA	Bouvier Bernois	0,83	1er
AB	Bouvier Bernois	1,4	1er
AC	GBS (Grand Bouvier Suisse)	1,4	1er
AD	Leonberg	2,5	1er
AE	Beauceron	1,4	1er
AF	Berger des Pyrénées	2,2	1er
AL	Dogue Allemand	0,53	2nd
BB-1	Boxer	7,2	2nd
BB-2	Boxer	25	2nd
BB-3	Boxer	72	2nd
BB-4	Boxer	12	2nd
BE	Retriever du Labrador	3,1	2nd

1.3 Génome de référence

Le génome de référence utilisé pour cette analyse est CanFam4 (ou UU_CFam_GSD_1.0) qui constitue la référence d'assemblage de haute qualité du génome du chien domestique (*Canis lupus familiaris*) utilisé par la communauté scientifique travaillant sur la génomique du chien. Publié en 2021, il s'agit de l'une des versions les plus récentes et les plus complètes du génome canin disponible. CanFam4 est un assemblage chromosomique complet du génome canin obtenu par un séquençage à très haute profondeur (100X) d'un chien Berger Allemend (German Shephered)[12]. Comprenant les chromosomes autosomiques (38) ainsi que les chromosomes sexuels (X & Y)[13].

2. Analyse bioinformatique

2.1 Environnement de développement

2.1.1 Genouest pour l'infrastructure de calcul et de stockage

Toutes les analyses bioinformatiques ont été menées en utilisant les ressources de GenOuest, une plateforme technologique hébergée à l'INRIA/IRISA de l'Université Rennes 1. Son cluster, composé d'environ quarante machines, exploite le planificateur de tâches SLURM (Simple Linux Utility for Resource Management) pour une gestion efficace des calculs, avec une puissance totale de 2960 threads CPU et 25 To de mémoire RAM (vérifié le 29/04/2024). Chaque utilisateur bénéficie d'un répertoire personnel "`/home`" limité à 120 Go et d'un répertoire temporaire "`/scratch`" avec une capacité maximale de 250 Go. L'authentification se fait via SSH garantissant une confidentialité et une sécurité des données. GenOuest propose une gamme étendue d'outils bioinformatiques, tels que nextflow, fastqc, samtools, minimap2, bedtools, conda, déjà préinstallés et disponibles pour une utilisation immédiate. De plus, les utilisateurs ont la possibilité d'installer d'autres logiciels dans des environnements personnalisés Conda ou Mamba pour répondre aux besoins spécifiques selon les projets de recherche.

Certaines équipes de recherche bénéficient d'un espace total dédié de 910 To, parmi lesquelles l'équipe Génétique du Chien qui dispose d'un espace spécifique accessible via "`/groups/dog`", offrant une capacité de stockage de 72 To.

2.1.2 Visual Studio Code (VS code)

J'ai utilisé Visual Studio Code (VS Code) comme éditeur de code. Il s'adapte à tous les systèmes d'exploitation (Windows, macOS et Linux) et propose une interface colorée qui facilite la lecture du code. VS Code offre également la possibilité d'accéder au serveur SSH de GenOuest via l'extension SSH Remote. De plus, grâce à des extensions spécifiques, il permet de lire des fichiers PDF, HTML et CSV. Il prend en charge plusieurs langages de programmation, ce qui en fait un outil extrêmement polyvalent pour divers besoins de développement.

2.1.3 Langages de programmation utilisés

La majorité de mes scripts ont été écrits en Python V3.10.6, en appelant ses nombreuses bibliothèques comme NumPy, Pandas, Matplotlib, Seaborn pour manipuler et visualiser les données. J'ai également utilisé R V.4.3.2 pour certains tests statistiques et pour la visualisation sophistiqués grâce à ses outils graphiques avancés comme ggplot2. En ce qui concerne la manipulation des fichiers textes, j'ai utilisé le langage awk. Le pipeline nanoseq V.3.1.0. est codé en Nextflow V.23.10.0. Enfin, j'ai utilisé des commandes SLURM, notamment la commande sbatch pour soumettre mes analyses en parallèle de manière optimisée sur les ressources de calcul disponibles.

2.1.4 Gestionnaires de code source Github et Gitlab

GitHub/Gitlab sont des plateformes de développement collaboratives utilisant Git pour le contrôle de version. Elles permettent aux développeurs de stocker, gérer et partager leur code, facilitant ainsi la collaboration à travers des fonctionnalités comme les pull requests et les issues.

2.2 Nextflow/nanoseq pour l'analyse des données de séquençage longue lecture

2.2.1 Nextflow

Nextflow est un système de gestion de workflows principalement utilisé pour les analyses de données bioinformatiques complexes et gourmandes en calcul. Son modèle de programmation orienté flux de données repose sur deux concepts clés: les Process et les Channels. Un Process définit une tâche à exécuter, comme un script, une commande ou un outil, avec des entrées et des sorties définies par des Channels. Ces Channels permettent la communication entre les différents Process, contrôlant ainsi le flux de données du pipeline de

manière déclarative. L'un des principaux atouts de Nextflow est sa portabilité et sa reproductibilité indépendamment de l'environnement. Les workflows peuvent être déployés sur divers environnements (ordinateur local, cluster, cloud) sans modification du code source grâce à l'intégration native avec les conteneurs logiciels comme Docker et Singularity.

La parallélisation est implicite, basée sur les entrées/sorties des Process, et les pipelines peuvent s'exécuter sur différentes plateformes de calcul haute performance. Nextflow offre également la reprise du workflow, le suivi des résultats intermédiaires, l'intégration avec GitHub et un langage de domaine spécifique fluide pour définir les pipelines.

2.2.2 Nf-core/nanoseq

Nf-core est une communauté visant à rassembler un ensemble de pipelines d'analyse bioinformatique construits avec Nextflow. Nf-core s'engage à produire des workflows standardisés, conviviaux et bien documentés, testés et validés par la communauté avant leur mise à disposition (Nano-seq, RNA-seq…, etc.). Ces workflows répondent à des normes strictes de qualité et de reproductibilité, permettant aux utilisateurs de réaliser des analyses de données efficaces et fiables, avec l'assurance de résultats de qualité. Actuellement, nf-core propose 105 pipelines, offrant ainsi aux chercheurs une palette d'outils pour mener des analyses standardisées et reproductibles. Les dépendances sont automatiquement gérées via des conteneurs ou des environnements virtuels. Nf-core est soutenu par une large communauté de contributeurs du monde entier avec un canal Slack actif et des événements réguliers (Hackathon Mars 2024 à Rennes).

nf-core/nanoseq est le pipeline d'analyse développé par nf-core pour les données de séquençage d'ADN/ARN produites par les technologies nanopore. Ce pipeline permet d'effectuer l'appel de base, le démultiplexage, le contrôle qualité, l'alignement, ainsi que l'analyse en aval des données Nanopore DNA. Les lignes verte et orange sont dédiées à l'analyse de données transcriptomique, tandis que la branche bleue, que nous avons suivie dans notre projet, concerne l'ADN (Fig.4).

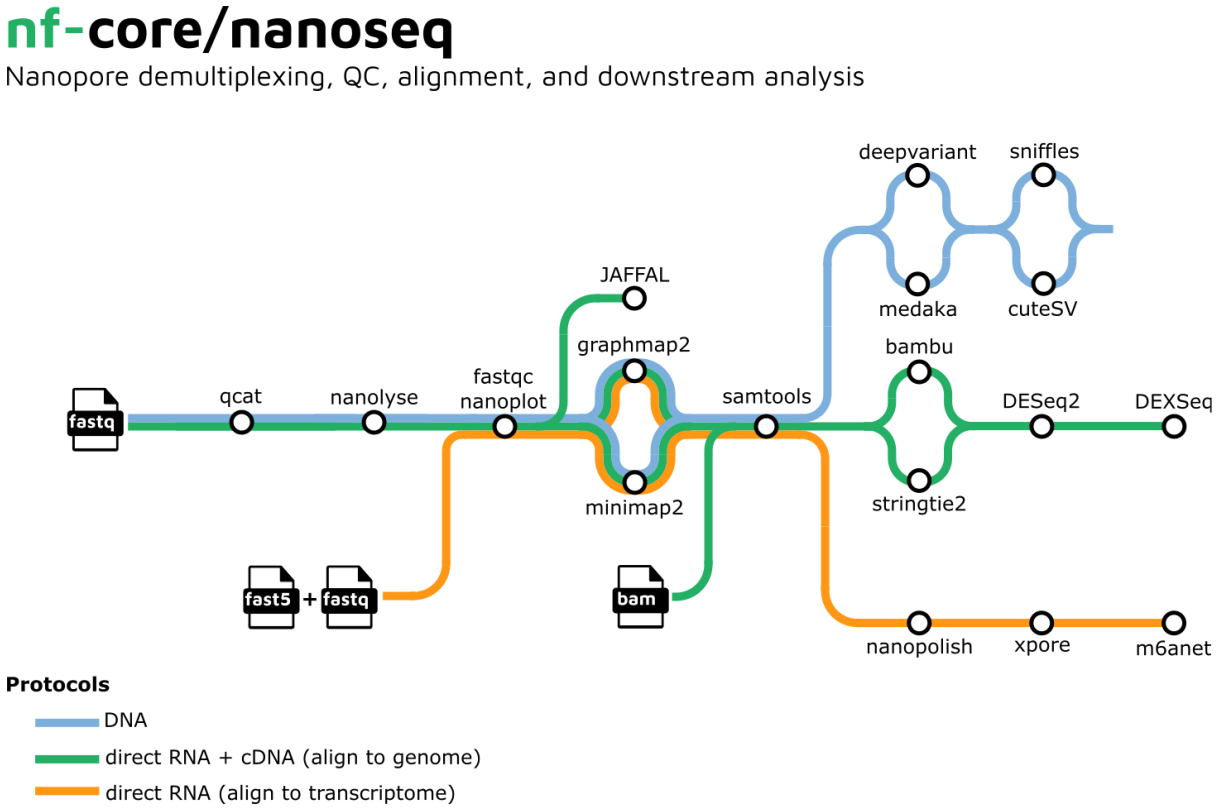

Figure 4. Représentation schématique (Metro map) du pipeline nanoseq conçu par nf-core.

Utilisation

La version (`nanoseq/3.1.0`) a été utilisée par défaut pour les échantillons du premier batch de petite taille. En revanche, une version téléchargée et modifiée localement a été employée pour les échantillons du deuxième batch afin d'ajuster certains paramètres par défaut pour qu'ils supportent les données de grande taille en termes de ressources et de temps.

Le fichier d'entrée requis pour le pipeline est spécifié avec l'option (`--input`) suivie du chemin du fichier (.csv) correspondant contenant les informations sur les échantillons à analyser, réparties en six colonnes : les groupes d'échantillons, le nombre de réplicats, les barcodes utilisés pour identifier les échantillons, les chemins vers les fichiers (`.fastq`) d'entrée, le chemin vers le fichier de génome de référence pour l'alignement et le fichier (.gtf). Le pipeline exige également de connaître le type de données d'entrée via l'option (`--protocol`). Dans notre cas, le protocole défini est DNA.

Certaines étapes du pipeline ont été ignorées avec les options (`--skip_quantification`) et (`--skip_demultiplexing`). Cela signifie que l'analyse ne tiendra pas compte de la quantification des échantillons et du démultiplexage des séquences d'ADN, ces opérations ayant été réalisées préalablement par les séquenceurs.

L'appel des variants est activé avec l'option (`--call_variants`). Deux outils spécifiques sont utilisés pour l'appel des variants : “cutesv” et “sniffles” pour les variants structuraux (spécifiés avec `--structural_variant_caller`) et “deepvariant” pour les autres variants (spécifié avec `--variant_caller`). Le profil d'exécution est également exigé par le pipeline, dans notre cas "singularity" a été utilisé en spécifiant l'option "--profile singularity". Cela garantit que les dépendances du pipeline sont exécutées dans des conteneurs Singularity. Les ressources de calcul sont spécifiées pour l'exécution du pipeline pour les données volumineuses du deuxième batch. Les options sont : (`--max_cpus`) (24 cpu), (`--max_memory`) (256 G) pour la mémoire maximale et (`--max_time`) (48 heures) pour la durée maximale d'exécution.

2.4 Contrôle qualité des données (FastQC, nanoplot, mapping QC, variant calling QC)

2.4.1 Contrôl de qualité FastQC

L'étape FastQC a été ignorée pendant l’analyse car elle utilise l’encodage de qualité (Sanger/Illumina 1.9) qui est inadapté pour nos données séquencées par la technologie nanopore. Les scores de qualité Nanopores sont codés avec des caractères ASCII de 33 à 126 (plus la valeur est élevée, meilleure est la qualité attendue)[14]. Par conséquent, j'ai basé mon analyse sur NanoPlot, mieux adapté pour traiter et évaluer la qualité des séquences de longues lectures.

2.4.2 Contrôle de qualité des Lectures Longues (Nano QC)

Le pipeline utilise NanoPlot (version=1.41.0) pour le contrôle qualité des données. Il permet de créer diverses représentations graphiques à partir de fichiers (`.fastq`)[15]. Ses principales fonctionnalités incluent des tracés de la longueur des lectures, des scores de qualité (Phred) et d'autres métriques clés. Le score de qualité Phred est une mesure de confiance basée sur le taux d'erreur estimé et est calculé comme -10 x log (Pe), où Pe est la probabilité d'erreur estimée (eg: une erreur de 1 sur 100 donnera un score Q de 20 et une erreur de 1 sur 1 000 donnera un score Q de 30). Les scores de qualité par base sont stockés

avec la séquence de base dans les fichiers FASTQ (4ème ligne) générés par les algorithmes d'appel de base.

Nanoplot offre également des représentations comparant différentes statistiques telles que la longueur des lectures (reads) versus leur score de qualité, ainsi que des histogrammes, des nuages de points, des graphiques en forme d'hexagone ou de noyau de densité.

2.4.3 Contrôle qualité d'alignement (Mapping QC)

Dans la metro map, deux choix pour les outils d'alignement ont été proposés : soit Minimap2 (outil par défaut), soit Graphmap2. Je me suis basée sur la documentation, et j'ai utilisé Minimap2 dans mon analyse car il est largement reconnu pour sa rapidité, sa capacité de gérer des données volumineuses ainsi que sa précision dans l'alignement de longues séquences d'ADN, ce qui en fait un excellent choix pour l'alignement de génomes [16]. Minimap2 prend en entrée (`input`) des fichiers de type (`.fastq` ou `.fasta`) et génère en sortie des fichiers de type (`.sam`). Les fichiers (`.sam`) à l'aide de l'outil Samtools sont par la suite compressés en fichiers de types (`.bam`). Afin de contrôler la qualité des alignements à partir des fichiers (`.bam`), j'ai utilisé Samtools pour compter les identifiants des lectures alignées et non alignées sur le génome de référence. Dans les fichiers SAM et BAM, la chaîne CIGAR (Concise Idiosyncratic Gapped Alignment Report), décrit comment chaque séquence lue est alignée à la séquence de référence. La chaîne CIGAR spécifie les opérations telles que les matches, insertions et délétions, fournissant des détails essentiels pour l'analyse et la visualisation des alignements de séquences d'ADN. Pour calculer la profondeur de séquençage, j'ai utilisé (`Samtools depth`), et pour calculer la couverture, j'ai utilisé (`Samtools coverage`).

2.4.4 Contrôle qualité de l'appel de variants (Variant calling QC)

2.4.4.1 Détecteurs des SNVs/SNPs (Singles Nucleotides Variants) et InDels

L'appel de variants s'est fait par l'option (`--call_variants`), l'identification des SNVs se fait par l'option (`--variant_caller`) suivie du nom d'outil de détection. La métro map propose deux outils : Medaka et Deepvariant.

2.4.4.2 Détecteurs des SVs (Variants Structuraux)

Le pipeline nf-core/nanoseq propose pour la détection des variants structuraux deux outils: CuteSV et Sniffles. Ils sont conçus pour détecter les variations structurales dans les séquences génomiques à partir de fichiers (`.bam`) triés et indexés.

CuteSV fonctionne sous trois étapes principales (Fig.5)[8]. Tout d'abord, une phase de découverte des signatures de SV où chaque type de SVs (délétions, insertions, inversions, duplications et translocations) se dispose d'une signature spécifique identifiable à partir des chaînes CIGAR, ainsi que les discordances d'alignement à travers les lectures non contigues[17]. Ensuite, une deuxième étape de regroupement des signatures des SVs en clusters en fonction de leur position génomique et de leur type, il utilise une méthode heuristique de clustering et de raffinement pour identifier avec précision les allèles altérés, en combinant les signatures détectées de plusieurs lectures et en éliminant les faux positifs affinant ainsi les clusters. Les lectures d'insertion ou de délétion situées à une distance maximale de 100 à 200 bases peuvent être regroupées. Les points de rupture ne sont regroupés que si leur similarité ne dépasse pas un ratio de différence de 0,3 à 0,5, ce qui réduit le risque de fusionner des événements distincts. Les faux positifs, souvent dus à des alignements de faible qualité (MAPQ inférieur à 20) ou des discordances mineures, sont éliminés si moins de 10 lectures les soutiennent ou si leur taille est inférieure à 30 bases.

Enfin, une dernière phase de détection et de génotypage des SVs, où CuteSV génère les ensembles de SV et attribue les génotypes correspondants, enregistrant les résultats finaux dans un fichier (`.vcf`) (Variant call format) contenant des informations détaillées sur les SV détectées.

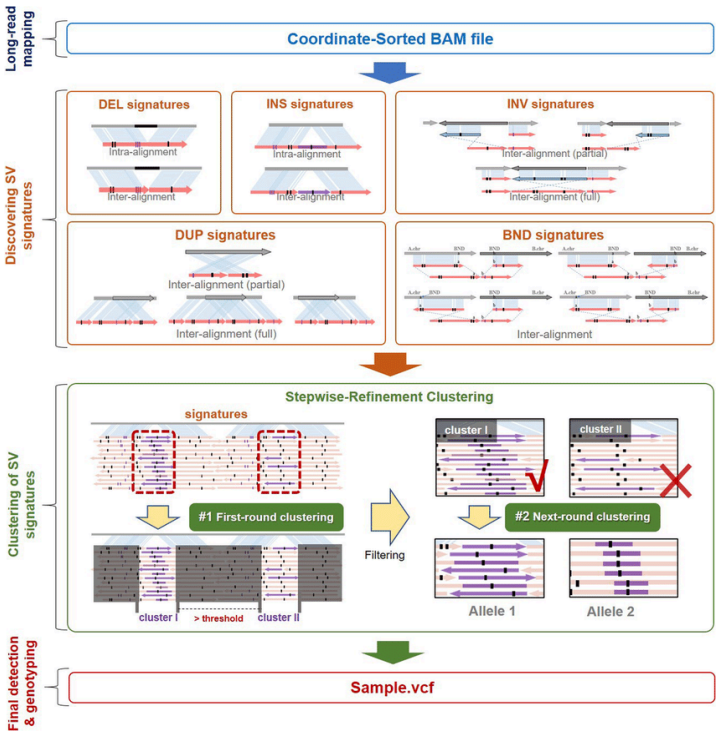

Figure 5. Processus détaillé de l'algorithme CuteSV pour la détection des variations structurales [8].

Sniffles quant à lui, fonctionne selon un algorithme qui suit plusieurs étapes (Fig.6). Une étape d'estimation des paramètres nécessaires tels que la taille des variants attendus (longueur minimale des SVs : 100 bases), la position, le type et la couverture. Ensuite, une étape d'analyse pour l'alignement sert à identifier les segments de séquences montrant des décalages ou des incohérences. Après cette étape, l'algorithme analyse les signaux de cassure (splittés), qui indiquent des réarrangements potentiels. Et puis il regroupe les points de rupture en clusters qui montrent une cohérence et une fréquence suffisantes pour être considérés comme des SVs valides. Enfin, une dernière étape de filtre en fonction du support des lectures (minimum 10 lectures) et de la confiance afin d'éliminer les faux positifs[17]; (SVs avec un alignements de faible qualité < 20, ou de taille < 30 bases.

Sniffles par rapport aux autres outils, se caractérise par sa capacité à détecter des types de SVs mal étudiées, tels que les duplications en tandem inversées (INVDUP) ou les inversions flanquées d'indels (INVDEL)[18].

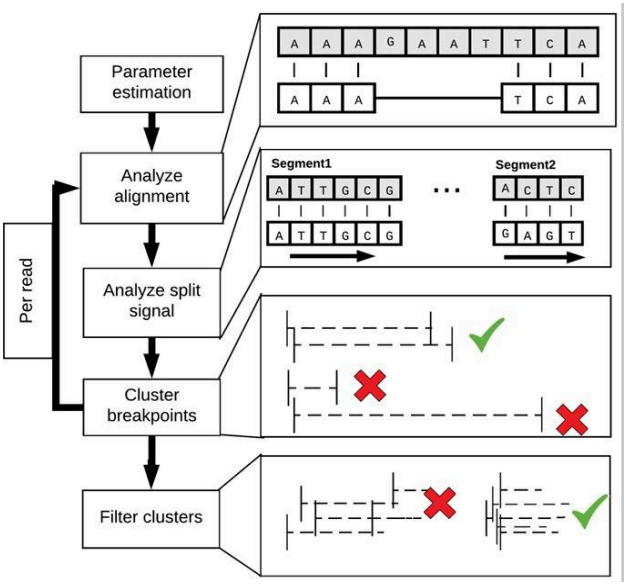

Figure 6. Aperçu des principales étapes mises en œuvre dans Sniffles [18].

2.4.5 Identification des Variants Structuraux en communs

Afin d'identifier et de manipuler les intersections entre l'ensemble des variants issus de fichiers (`.vcf`) générés par CuteSV et Sniffles, j'ai utilisé l'outil **Bcftools**. (`Bcftools isec`) fonctionne de la manière suivante: D'abord il compare plusieurs fichiers entre eux pour identifier les variants structuraux en communs (intersections) et uniques (différences) entre ces fichiers. Les fichiers doivent être triés et indexés auparavant. Ensuite, après avoir fournis en entrée les fichiers (`.vcf`) avec leurs indexes (.tbi), l'outil commence par lire les informations sur les variants, y compris les positions chromosomiques, les allèles de référence et alternatifs, les identifiants de variants, et d'autres annotations. Une étape de comparaison se met en place, où l'algorithme utilise une approche de balayage synchronisé (synchronized sweep) pour comparer les variants. Il lit les variants de chaque fichier en parallèle et compare leurs positions chromosomiques. Pour chaque position, il vérifie si un variant est présent dans tous les fichiers à comparer. Cela se fait en avançant de manière synchronisée dans les fichiers pour chaque position chromosomique.

Les résultats des comparaisons sont filtrés en fonction des options spécifiées. L'option `-n=2` que j'ai spécifié permet d'indiquer le nombre minimum de fichiers dans lesquels un variant doit apparaître pour être inclus dans le résultat.

J'ai également ajouté l'option `-c all`, pour comparer tous les aspects des variants lors de l'intersection. Cette option compare les positions des variants; que les allèles alternatifs ALT

correspondent ou non, les variants sont considérés comme identiques s'ils ont la même position[19].

Une autre méthode d'identification des SVs en commun a été utilisée. **Bedtools intersect**, est un outil qui permet de rechercher les chevauchements entre deux ou plusieurs fichiers de type (`.bam/.vcf/.bed`), il lit les fichiers d'entrée ligne par ligne pour extraire les informations sur les chromosomes et les positions de début et de fin des SVs. Pour un accès rapide aux régions génomiques, il faut que les fichiers soient indexés. Ensuite, pour chaque chromosome, l'outil effectue une comparaison afin de trouver s'il y a des intersections. Pour chaque région potentiellement chevauchante trouvée, l'algorithme vérifie si le chevauchement est réel en comparant les positions de début et de fin des deux régions.

3. RÉSULTATS

3.1. Mise en place du formulaire en ligne

Le formulaire de renseignement créé à partir de l'outil Jotform a été mis en ligne et est accessible depuis le site internet de l'IGDR à l'adresse suivante :

https://igdr.univ-rennes.fr/crb-cani-dna/kit-de-prelevement.

La mise en place d'un bouton radio au bas du formulaire permet de recueillir le consentement du propriétaire à l'utilisation des données d'identité en conformité avec le Règlement sur la protection des données personnelles (RGPD).

Jotform a été configuré pour qu'un mail de notification soit envoyé au CRB chaque fois qu'un propriétaire a complété le formulaire en ligne. Les données soumises peuvent ensuite être récupérées en batch en format xlsx depuis l'espace Jotform afin d'exécuter les scripts permettant la création des lettres personnalisées d'une part, et la vérification de l'existence ou non des informations relatives aux propriétaire/vétérinaire/chien dans la base de données de Cani-DNA d'autre part.

3.2. Analyse des données de séquençage longue lecture pour le projet GOLDogs

3.2.1 Résultats de contrôle qualité des lectures longues (Nano QC)

Le tableau ci-dessous (Tab.2) présente une comparaison des valeurs de qualité issues de l'analyse de contrôle qualité NanoQC pour deux échantillons représentatifs AD et BB-2 des premier et deuxième batch de séquençage respectivement.

Tableau 2. Tableau comparatif des métriques de qualité pour deux échantillons représentatifs du premier et deuxième batch.

Synthèse générale	Premier batch (GridION)	Deuxième batch (PromethION)
Mean read length (bp)	5148,1	5216,1
Mean read quality	14,6	18,5
Median read length (bp)	2294	3370
Median read quality	15,1	19
Number of reads	587 677	4 800 935
Read length N50 (bp)	11 815	9249
STDEV read length (bp)	7347,9	5564,7
Total bases (bp)	3 025 434 310	25 042 137 499

Les deux échantillons représentatifs de faible profondeur de séquençage, et de profondeur élevée séquencés par GridION et PromethION respectivement, ont une longueur moyenne des lectures d'environ 5 kb; peu importe le type de séquenceur la longueur est similaire. Néanmoins, on observe que la qualité moyenne des lectures dans le deuxième batch séquencé par PromethION est nettement meilleure, avec un Phred score de 18,5 correspond à une probabilité qu'une base soit incorrecte d'environ 1,41% (Figure 7). Par rapport au batch séquencé par GridION avec un Phred score de 14,6 qui correspond à une probabilité d'erreur d'environ 3,47%.

La qualité médiane des lectures augmente de 15,1 à 19, indiquant que la majorité des séquences dans le batch PromethION sont de meilleure qualité. En ce qui concerne le nombre de lectures dans le batch PromethION (n = 4 800 935), on peut remarquer qu'il est considérablement plus élevé que dans le batch GridION (n = 587 677), indiquant une profondeur de séquençage beaucoup plus élevée et une couverture plus importante. Parallèlement, le nombre total des bases séquencées dans le batch PromethION est beaucoup plus élevé avec 25 042 137 499 pb que dans le batch GridION avec 3 025 434 310 pb, montrant une nette amélioration en termes de quantité de données générées.

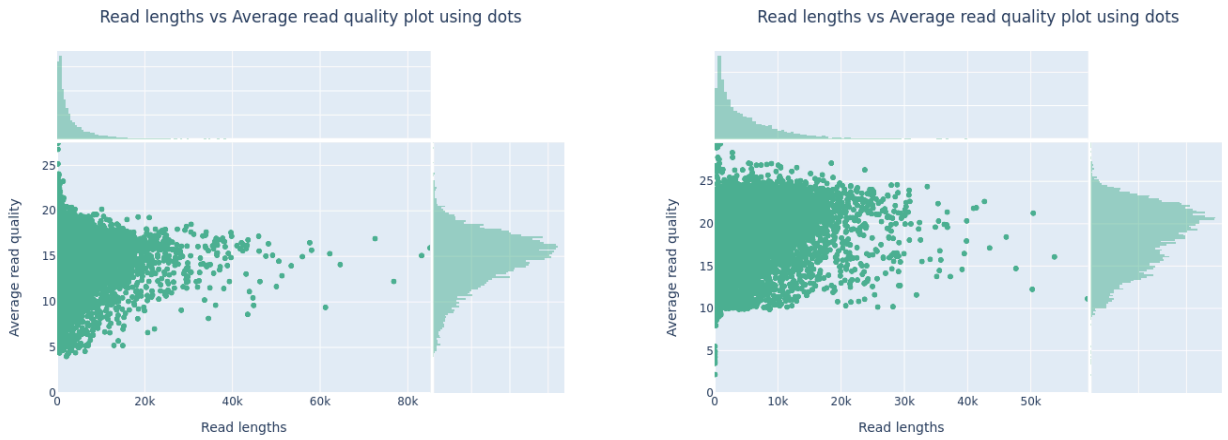

Figure 7. Diagrammes de dispersion (Nanoplot) des longueurs et de la qualité moyenne pour les échantillons AD (premier batch) à gauche et BB-2 (deuxième batch) à droite, montrant la distribution des longueurs des lectures et leur qualité moyenne à l'aide d'histogrammes en haut et à droite. Chaque point représente une lecture.

On observe une concentration des points pour des longueurs des lectures de taille inférieure à 50 kb pour les deux échantillons. Les scores moyen de qualité varient de 5 à 25 pour l'échantillon AD, et de 10 à plus de 25 pour l'échantillon BB-2.

3.2.2 Résultats de contrôle qualité d'alignement (MappingQC)

Après l'étape de contrôle-qualité des longues lectures ONT, le pipeline nanoseq inclut l'étape d'alignement des lectures sur le génome de référence canin (canFam4, cf Mat/Met). À partir des fichiers d'alignements produits par minimap2, nous avons représenté la distribution des lectures alignées et non alignées sur le génome de référence pour les échantillons du premier batch (Fig.8) séquencés par GridION et du deuxième batch (Fig.9) séquencés principalement par PromethION.

Les graphiques représentent à la fois la distribution d'alignements et le nombre des lectures générées par chaque échantillon.

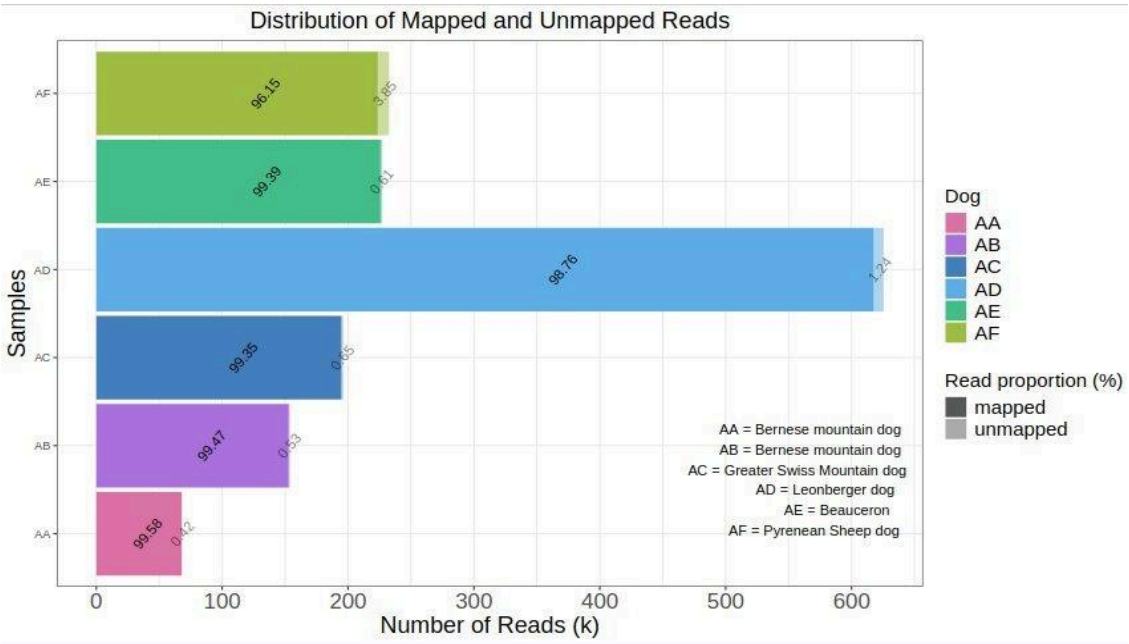

Figure 8. Distribution des lectures alignées et non-alignées sur le génome de référence pour les échantillons du premier batch, avec des couleurs différentes pour chaque chien. Les lectures non alignées sont représentées par une couleur moins intense que les lectures alignées.

D'après la (Fig.8), on constate que la majorité des lectures ont été bien alignées avec un pourcentage moyen d'alignement de (98%) pour tous les échantillons. On note cependant que les échantillons AD et AF ont un pourcentage des lectures non alignées un peu plus élevé de 1,24% et de 3,85% probablement lié à une variabilité génomique avec le génome de référence ou à des problèmes techniques liés à ces échantillons lors d'extraction d'ADN ou lors du séquençage. Le nombre de lectures générées varie entre les différents échantillons de 60 000 à 600 000 lectures.

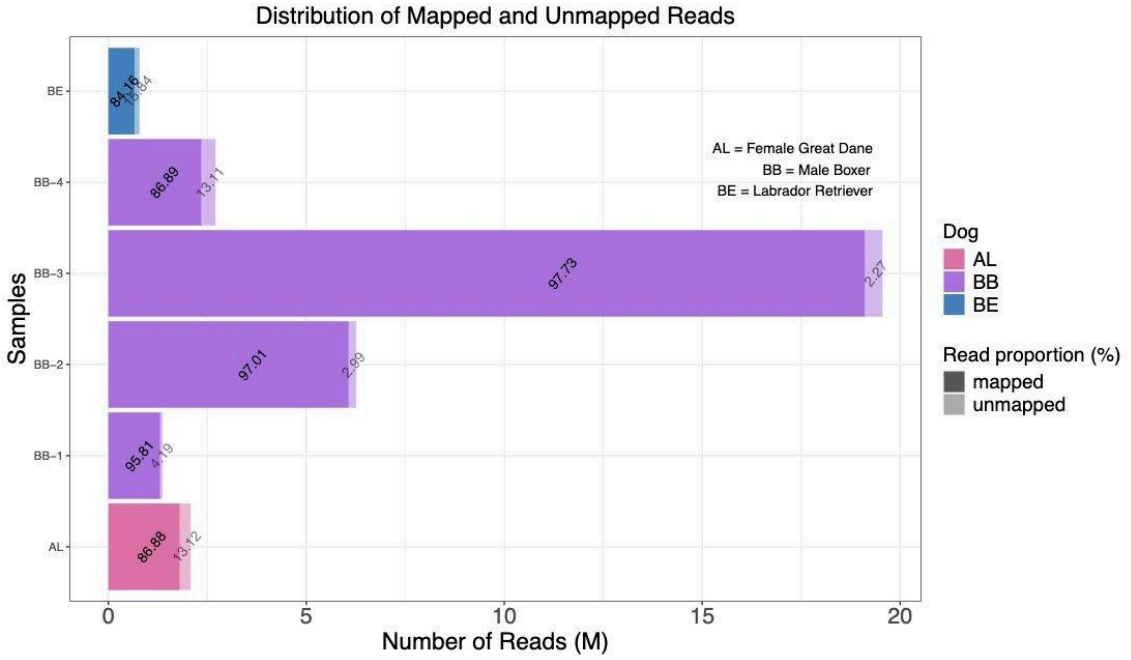

Figure 9. Distribution des lectures alignées et non-alignées sur le génome de référence pour les échantillons du deuxième batch, avec des couleurs différentes pour chaque chien. Les lectures non alignées sont représentées par une couleur moins intense que les lectures alignées. Les barres qui ont la même couleur représentent des réplicats techniques provenant du même chien.

Le graphique de la (Fig.9) représente les échantillons séquencés principalement par promethION, sauf l'échantillon BB-1 qui a été séquencé par GridION. L'ensemble des lectures ont été bien alignées avec un pourcentage moyen de (91%). On remarque que pour un même individu (entre les réplicats de l'échantillon BB), il y a une grande variabilité du nombre de lectures générées. Ceci est principalement lié à des protocoles d'extraction d'ADN différents par les collaborateurs de France Génomique. PromethION a produit une quantité de lectures plus importante que GridION allant jusqu'à 19 millions de lectures. Cette différence peut s'expliquer par les techniques utilisées pour l'extraction de l'ADN. Selon les collègues de France Génomique, l'échantillon BB-3 avec le meilleur rendement en termes de quantité de lectures est celui pour lequel ils ont utilisé un protocole d'extraction avec une étape de purification sur colonnes Qiagen.

3.2.3 Résultats de contrôle qualité des variants (Variant calling QC)

3.2.3.1 Identification des variants de petites tailles (SNVs et InDels)

Le nombre de variants courts identifiés par l'outil Deepvariant varie entre les différents échantillons. Pour les échantillons du premier batch, le nombre de variants ayant passé tous les critères de qualité varie de 155 788 variants détectés dans l'échantillon AA à 900 763 variants détectés dans l'échantillon AD. Toutefois, pour le deuxième batch, sur l'ensemble des six échantillons, seulement trois d'entre eux (BB-1, BB-4 et BE) ont pu être analysés pour les variants courts, ceci à cause de problèmes de stockage du serveur de notre équipe. L'échantillon BB-4 et BE ont respectivement 3 871 539 et 1 130 641 SNVs. Ces échantillons ont été séquencés par PromethION, fournissant ainsi un nombre plus important de lectures.

3.2.3.2 Évaluation du temps d'exécution de pipeline nanoseq

Le temps d'exécution de pipeline a été évalué, afin de comparer les deux outils de détection des variants structuraux CuteSV et Sniffles. Les échantillons varient en taille et en profondeur de séquençage, fournissant un aperçu des capacités de chaque outil dans diverses conditions.

Pour un échantillon de 25 Go et une profondeur de 11X, le pipeline utilisant CuteSV a complété l'analyse en 3 heures et 17 minutes, tandis que Sniffles a pris 21 heures et 16,5 minutes. Pour un échantillon de 12 Go et une profondeur de 5X, CuteSV a terminé l'analyse en environ 8 heures et 38 minutes, contre 15 heures et 40 minutes pour Sniffles. Pour un échantillon de 2,2 Go et une profondeur de 1,6X, CuteSV a pris 5 heures et 33 minutes, tandis que Sniffles a complété l'analyse en 3,9 heures.

3.2.3.3 Identification des variants structuraux

Le tableau ci-dessous (Tab.3) résume les informations concernant les noms des échantillons, les races des chiens, la profondeur de séquençage, et le nombre de variants structuraux identifiés par les deux méthodes CuteSV et Sniffles.

Tableau 3. Nombre de variants structuraux identifiés par CuteSV et Sniffles selon la profondeur de séquençage.

Echantillons	Séquenceurs	Profondeur de séquençage (X)	Nombre de SVs (CuteSV)	Nombre de SVs (Sniffles)
AA	GridION	1.24	13	5
AB	GridION	1.40	14	11
AC	GridION	1.42	16	12
AD	GridION	1.76	25	15
AE	GridION	1.41	13	8
AF	GridION	1.64	27	14
AL	PromethION	4.86	1391	917
BB-1	GridION	3.60	378	236
BB-2	PromethION	11.38	2 4467	18 231
BB-3	PromethION	33.13	78 269	ND
BB-4	PromethION	5.18	1890	1306
BE	PromethION	1.99	59	33

Comme nous l'avons vu précédemment, le nombre de séquences générées et donc la profondeur de séquençage varie considérablement entre les échantillons du premier batch séquencés par GridION et ceux du deuxième batch séquencés par promethION. Ainsi, l'échantillon AA a la profondeur de séquençage la plus faible (1,24X) alors que l'échantillon BB-3 possède une profondeur maximale de (33X). Les échantillons BB-1, BB-2,BB-3 et BB-4 proviennent d'une même source (quatre aliquots d'un prélèvement de sang de Boxer). Les variations observées au niveau de la profondeur de séquençage et du nombre de variants structuraux détectés chez le même individu reviennent à la variabilité technique et méthodologique lors de la préparation des échantillons et du séquençage.

De façon générale, le séquenceur PromethION produit des lectures plus nombreuses et donc une couverture de séquençage plus élevée que le séquenceur GridION. Cela est lié au fait que le nombre de pores disponibles sur les flowcells/membrane de séquençage PromethION est 5 fois plus important que pour les flowcells GridION/MinION.

Pour la plupart des échantillons, CuteSV semble être plus sensible, capturant un plus grand nombre (environ 1,6 fois plus) de variants structuraux que Sniffles. Cependant, les échantillons avec des profondeurs de séquençage plus faibles montrent des différences moins marquées entre les deux méthodes.

On remarque également que les échantillons avec une grande profondeur de séquençage ont plus de variants structuraux qu'avec ceux à faible profondeur. Ceci m'a poussé à poser la question de la corrélation entre la profondeur de séquençage et le nombre de SVs détectés.

3.2.4 Corrélation entre la profondeur de séquençage et le nombre de variants structuraux (SV)

L'hétérogénéité des profondeurs de séquençage causée par les différentes méthodes utilisées et par la qualité de l'extraction de l'ADN semble influencer l'identification des variants structuraux (SV). Les capacités de lecture qui définissent la profondeur de séquençage ou la couverture varient d'un séquenceur à un autre. Afin d'étudier l'impact de la profondeur de séquençage sur le nombre de variants détectés, j'ai mis en place un test statistique. L'hypothèse nulle suggère qu'il n'y a pas de relation entre la profondeur de séquençage et le nombre de variants identifiés pour les 12 échantillons. Le test statistique choisi est le test non paramétrique de Spearman, nous nous attendons à observer une relation monotone où le nombre de variants détectés augmentera avec la profondeur de séquençage.

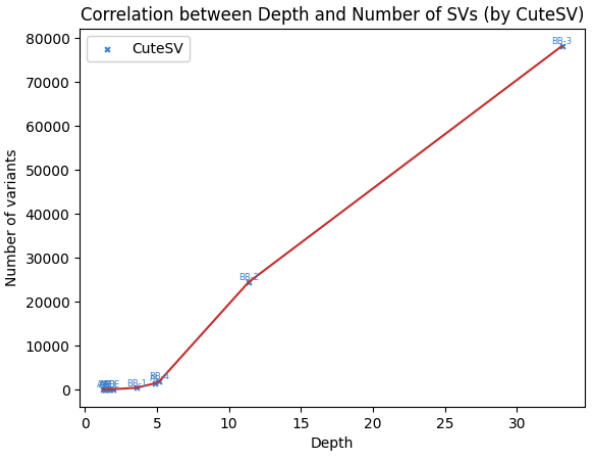

Figure 10. Corrélation entre la profondeur de séquençage (Depth en X) et le nombre de variants détectés par CuteSV.

Le test de corrélation de Spearman a donné un coefficient de corrélation de rang (rho) de 0,98, et une P-valeur de 2.022 e-08. Le coefficient est proche de 1 et la p-valeur est bien inférieure au seuil significatif alpha de 0,05. Ce résultat signifie que la corrélation observée est statistiquement significative, et il est très peu probable que cette corrélation soit due au hasard. Dans ce cas l'hypothèse nulle est rejetée et il existe bien une corrélation positive entre les deux variables. Autrement dit, plus la profondeur de séquençage est importante, plus le nombre de SVs augmente.

3.2.3.4 Analyse des types de variants structuraux identifiés par CuteSV et Sniffles

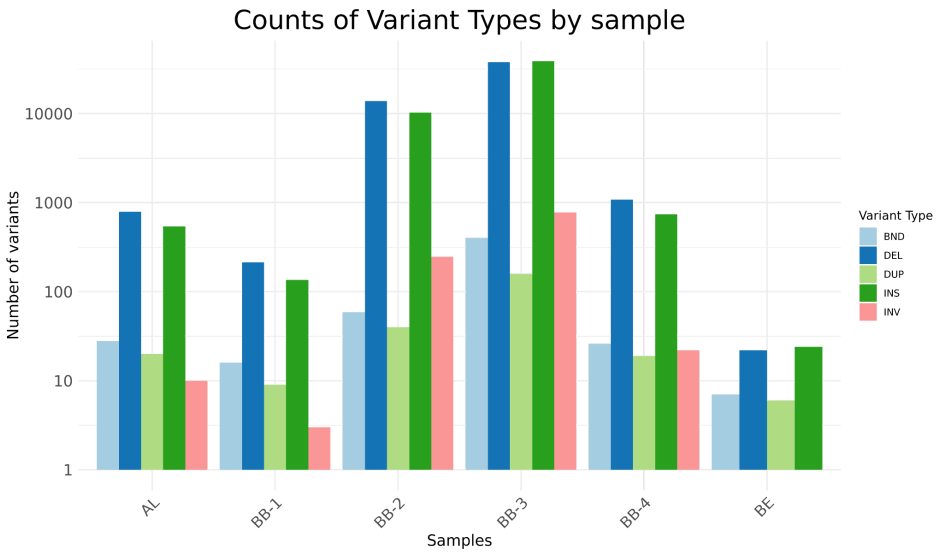

Figure 11. Répartition des types de variants structuraux détectés par CuteSV dans les échantillons du deuxième batch.

En analysant le graphique, on observe que l'échantillon BB-3 présente le plus grand nombre de SVs détectés avec environ 80 000 SVs. Parmi ces variants, les délétions et les insertions représentent la majorité des types de variants et ceci pour l'ensemble des échantillons.

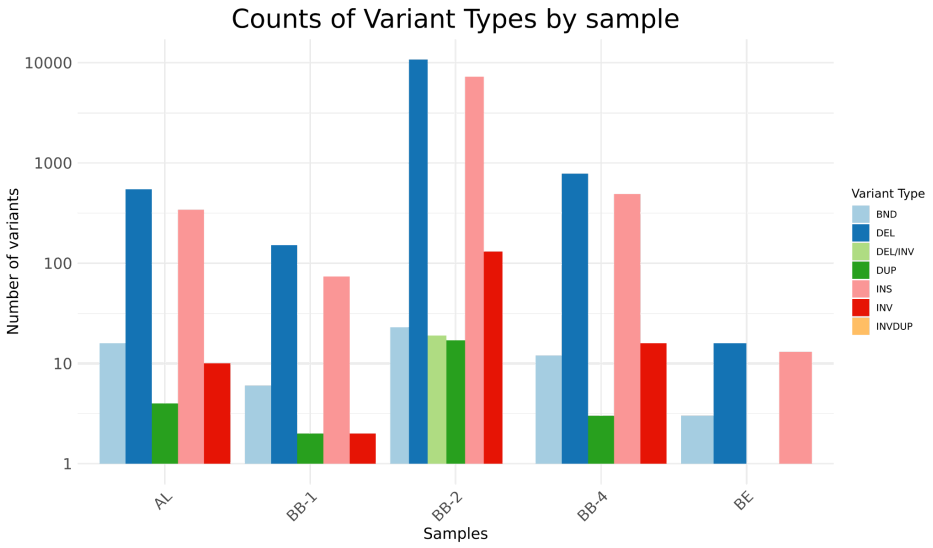

Figure 12. Répartition des types de variants structuraux détectés par Sniffles dans les échantillons du deuxième batch.

On observe également que Sniffles a détecté moins de SVs en terme de nombre, mais il a pu détecter d'autres types de SVs plus complexes que CuteSV n'a pas pu identifier tels que les combinaisons de deux types de modifications génomiques (variants nested) proches. Ainsi, seul Sniffles a permis de détecter les DEL/INV qui sont des délétions suivies d'une inversion dans la région adjacente ou impactée et les INVDUP, définie comme une inversion suivie d'une duplication de la séquence inversée ou d'une séquence adjacente.

3.2.3.5 Analyse des variants structuraux en communs identifiés par CuteSV et Sniffles

L'identification des SVs en commun entre les deux outils s'est faite par deux méthodes distinctes : d'une part avec l'utilisation de Bcftools isec et d'autre part avec l'outil Bedtools intersect (cf. Mat/Met). Pour simplifier l'analyse, je vais prendre un seul exemple pour comparer les résultats de ces outils (Fig.15). Pour l'échantillon BB-2 du deuxième lot, CuteSV a pu identifier 24 467 SVs contre Sniffles qui a identifié 18 231 SVs. Le nombre de SVs en commun trouvé par Bcftools est 4269 SVs (= 10% du total) contre 11 599 SVs (= 27.2% du total) trouvé par Bedtools. Cette tendance se confirme pour tous les échantillons.

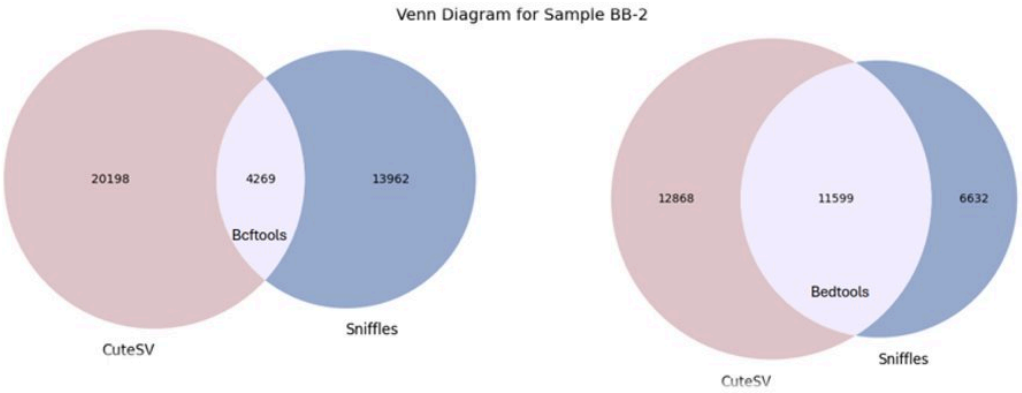

Figure 13. Diagramme de Venn comparant les variants structuraux (SVs) identifiés par CuteSV et Sniffles pour l'échantillon BB-2, ainsi que les SVs en commun identifiés par Bcftools et Bedtools.

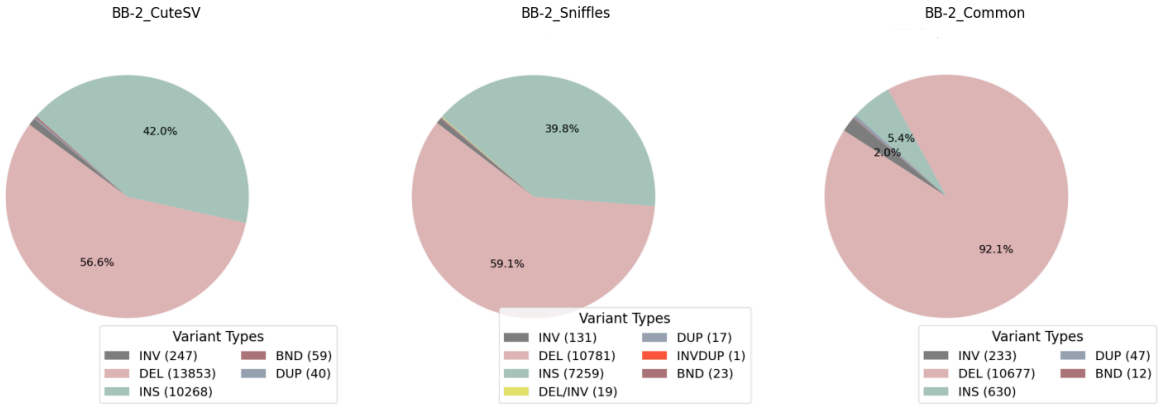

Figure 14. Diagrammes circulaires comparant les types de variants structuraux identifiés par CuteSV, Sniffles, et les variants communs identifiés par Bedtools.

Les résultats obtenus par CuteSV, montrent que les délétions constituent plus de la moitié des variants détectés (56,6%), suivies par les insertions (42%). Les autres types de variants (inversions, duplications et translocations) sont très minoritaires, chacun représentant moins de 1%.

Les résultats obtenus par Sniffles montrent également une prévalence des délétions, qui représentent 59,1% des variants, suivies par les insertions avec 39,8%. Les autres types de variants, y compris les DEL/INV, sont très rares.

Les SVs en commun entre CuteSV et Sniffles identifiés par Bedtools, présentent une prédominance de délétions (92,1% des SVs communs). Ils sont fortement enrichis par les délétions. Les insertions sont beaucoup moins fréquentes, représentant seulement 5,4%. Les

inversions, duplications et breakends sont très minoritaires, avec des proportions respectives de 2%, 0,4% et 0,1%.

Seulement 47,4% (11 599 sur 24 467) des variants détectés par CuteSV sont communs avec ceux détectés par Sniffles, et 63,6% (11 599 sur 18 230) des variants détectés par Sniffles sont communs avec ceux détectés par CuteSV.

3.2.3.6 Comparaison des variants structuraux entre races canines de tailles différentes

Malgré le nombre restreint des échantillons, et leur hétérogénéité en termes de couverture et de méthodes d'extraction. Une comparaison entre races a été faite. J'ai comparé les SVs identifiés pour deux races de différentes tailles (l'échantillon AL Dogue Allemand considéré comme une race géante) et (l'échantillon BB Boxer considéré comme race moyenne plus petite que le Dogue Allemand).

D'abord, la position du gène (Insulin-like Growth Factor-1) IGF1 a été repérée sur le génome de référence canFam4 (UU_CFam_GSD_1.0), situé sur le chromosome 15 position (41,392,877-41,567,967).

Un seul SV a été trouvé à l'intérieur de la région du gène IGF1. Il s'agit d'une délétion de 79 pb, située sur le chromosome 15 entre les positions 41,402,325 et 41,402,404, précisément dans un intron entre deux exons. Cette variation structurale est présente uniquement dans l'échantillon BB et absente dans l'échantillon AL.

3.2.3.7 Analyse des tailles des variants structuraux identifiés par CuteSV

Comme évoqué dans l'introduction, des collaborateurs américains (Halo et al.), ont montré que les variations structurales entre 2 races de chiens (dogue allemand et boxer) étaient dominées par la présence d'insertions et de délétions d'éléments transposables de type SINE (~200pb) et LINE (~6000 pb). Pour confirmer cette observation dans nos données de séquences, nous avons construit l'histogramme ci-dessous de la distribution des tailles des SVs selon leurs types pour l'échantillon BB-3 (Boxer) avec le génome de référence (Berger allemand) (Fig.15).

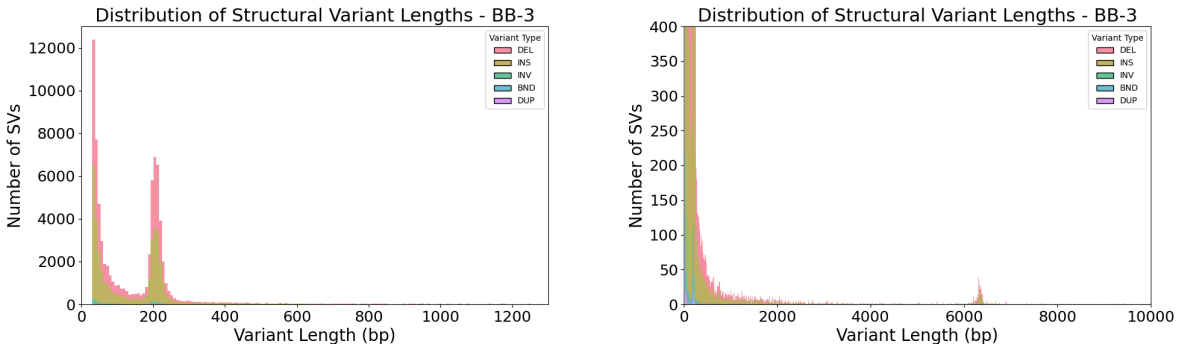

Figure 15. Distribution des longueurs des variants structuraux (pb) pour l'échantillon BB-3. Graphique de gauche : Distribution sur une échelle allant jusqu'à 12 000 SVs avec types de variants en couleurs (délétion, insertion, inversion, duplication, translocation). Graphique de droite : Distribution jusqu'à 400 SVs pour une meilleure visibilité des variations moins fréquentes.

On observe que la majorité des délétions et insertions identifiés ont une longueur d'environ 50 pb mais aussi deux pics dans les environs de 200 pb (Fig 15) et 6000 pb. Enfin, nous avons visualisé les différents SVs grâce au programme IGV (Fig. 16) qui nous a permis de confirmer la véracité des prédictions des outils bioinformatiques d'annotation de SVs.

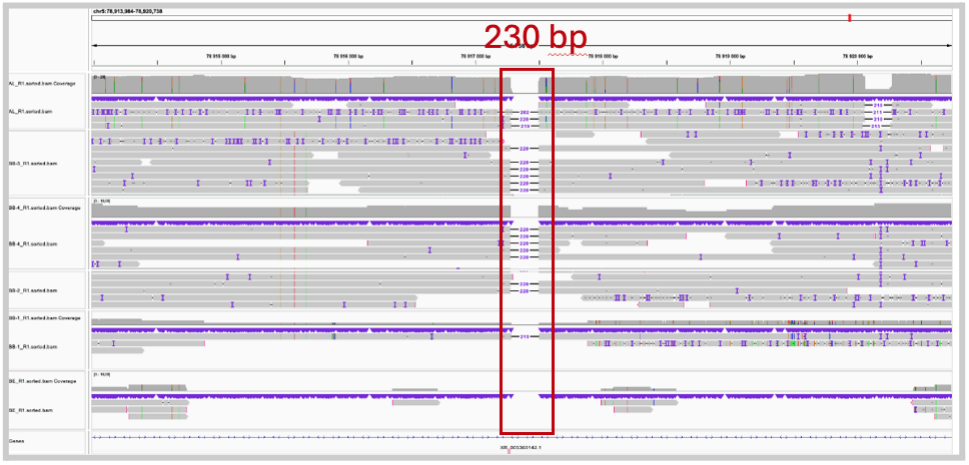

Figure 16. Visualisation IGV d'une délétion de 230 pb de longueur.

4. DISCUSSION

4.1 Discussion des résultats obtenus par rapport aux objectifs du stage

Il est essentiel de rappeler les objectifs initiaux de mon stage: le travail a reposé sur deux objectifs principaux: l'amélioration de la collecte et l'enregistrement des données dans le CRB Cani-DNA, ainsi que l'analyse bioinformatique de données de séquençage longues lectures, étude pilote du projet GOLDogs.

Pour la première partie, l'objectif était de faciliter et d'optimiser le recueil des informations sur les prélèvements qui entrent dans Cani-DNA. Un formulaire en ligne a été conçu afin d'améliorer l'efficacité et la précision des enregistrements dans la base de données Cani-DNA. Le processus de vérification des données saisies a pour but de garantir la fiabilité des informations collectées. La mise en ligne du formulaire a pour but d'une part de faciliter le traitement des données par le personnel de Cani-DNA mais également d'inciter les propriétaires, éleveurs et vétérinaires à participer aux projets de recherche de l'équipe Génétique du chien dans une démarche de science participative. Une telle initiative aux États-Unis via le site Darwin's Ark permet à tous propriétaires de devenir acteur de la recherche biomédicale et de faire progresser la compréhension des mécanismes génétiques chez les animaux de compagnie et les humains. Les prélèvements ainsi collectés viendront enrichir les collections d'échantillons de la bio-banque Cani-DNA et apporteront les ressources génétiques nécessaires à l'étude de la longévité chez le chien.

Pour la deuxième partie, mon travail a servi de point de départ (ou d'étude pilote) de projet GOLDogs. Une analyse préliminaire a été réalisée à partir des premières données de séquençage longue lecture réalisées par France Génomique sur 9 chiens. Les résultats obtenus lors de ces analyses permettent de définir les meilleures méthodes d'extraction et de séquençage à appliquer pour le projet GOLDogs et guide le choix des méthodes en fonction des résultats en termes de (nombre de lectures, profondeur de séquençage, couverture, nombre de variants identifiés…, etc).

Les échantillons séquencés avec PromethION présentent une qualité moyenne des lectures nettement supérieure au GridION, ainsi qu'une profondeur de séquençage et une quantité de données générées beaucoup plus importantes. Ces différences suggèrent que le

séquenceur PromethION est plus performant pour les projets nécessitant une bonne précision et une couverture génomique importante.

En ce qui concerne l'alignement, et en se basant sur la documentation, il est préférable d'utiliser **Minimap2**, car il est reconnu pour sa rapidité, sa capacité de gérer des données volumineuses ainsi que sa précision dans l'alignement de longues séquences d'ADN. La variabilité dans le nombre de lectures alignées et non alignées sur le génome de référence, particulièrement pour l'échantillon BB du 2nd batch, est due aux différences de méthodologies appliquées pour l'extraction d'ADN et pour le séquençage. On le voit surtout sur l'échantillon BB-3 séquencé avec PromethION et en utilisant une colonne Qiagen ce qui souligne l'impact des techniques d'extraction et de purification d'ADN sur le rendement du séquençage.

Pour la détection des SNVs/SNPs et Indels, je recommande l'utilisation de **Deepvariant**, car cet outil utilise des réseaux de neurones profonds (deep neural networks) pour identifier les variations génétiques. Selon Shafin et *al.*, il donne une précision de 99,8% pour les SNPs et 97,5% pour les Indels[20]. En comparaison, **Medaka** est développé par ONT, mais son auteur principal, Chris Wright, ne le recommande plus pour les séquences diploïdes, ce qui justifie l'utilisation de Deepvariant pour la détection des petits variants génomiques.

Dans notre analyse, nous avons observé une variation importante entre les résultats des deux batchs. Deepvariant a identifié plus de SNVs et d'indels dans les échantillons du 2nd batch séquencés principalement par PromethION, probablement en lien avec une profondeur de séquençage plus élevée par rapport à celle obtenue avec le GridION.

L'analyse de temps d'exécution du pipeline Nanoseq a comparé CuteSV et Sniffles pour la détection des variants structuraux sur des échantillons de tailles et de profondeurs de séquençage variées. CuteSV a une meilleure performance globale, complétant l'analyse en 3 heures et 17 minutes pour un échantillon de 25 Go et une profondeur de 11X, vs 21 heures et 16 minutes pour Sniffles (soit 7 fois plus rapide), qui a fini le processus avec des erreurs et a généré en plus des fichiers VCF mal formatés. Pour des échantillons plus petits, Sniffles s'est avéré plus rapide, mais ses difficultés avec des volumes de données plus élevés font de lui un choix moins pertinent. Cela est mentionné dans l'étude de Jiang et al., (2020), considérant en termes de performance, CuteSV comme l'outil le plus rapide qui utilise moins de mémoire

RAM. Sa capacité à presque doubler la vitesse avec l'ajout de threads CPU (Central Processing Unit) le rend particulièrement efficace pour les analyses à grande échelle. Ces caractéristiques font de CuteSV un outil idéal pour les tâches d'analyse de données à grande échelle.

En ce qui concerne les variants structuraux, et pour l'ensemble des échantillons des deux batchs, **CuteSV** a identifié un nombre de variants plus important que **Sniffles**, que ce soit à des faibles profondeurs de séquençage ou à des profondeurs de séquençage plus élevées. Cela souligne l'importance de la profondeur de séquençage et de la sensibilité des outils pour une identification précise des variants structuraux. L'étude de Jiang et al., (2020) confirme que CuteSV offre de meilleurs rendements de détection des SVs, avec une sensibilité accrue même pour les ensembles de données à faible couverture sans perte de précision.

L'importance de la profondeur de séquençage pour la détection des SVs a été mise en évidence en étudiant la corrélation entre la profondeur de séquençage et le nombre de variants structuraux détectés. En utilisant le test non paramétrique de Spearman, nous avons révélé une corrélation positive entre ces 2 facteurs. Les résultats montrent que le nombre de variants détectés augmente avec la profondeur de séquençage suggérant qu'une couverture plus élevée améliorerait l'identification des SVs jusqu'à probablement atteindre un plateau pour une certaine profondeur où tous les variants possibles auraient été détectés. Pour confirmer cette hypothèse, il faudrait envisager une approche de simulation des lectures à différentes profondeurs (1X, 10X, 50X, 100X) et ensuite une vérification du nombre de variants détectés et voir s'il atteint effectivement un plateau à partir d'une certaine profondeur de séquençage.

L'analyse comparative des types de variants structuraux identifiés par CuteSV et Sniffles révèle des différences notables dans les performances de ces outils. Les délétions et les insertions représentent le plus grand nombre de SVs détectés. Sniffles a détecté moins de SVs que CuteSV en terme de nombre total. Mais il a pu identifier d'autres types de SVs plus complexes représentant des combinaisons de modifications génomiques, telles que les DEL/INV (délétion suivie d'une inversion) et INVDUP (inversion suivie de duplication). D'après Sedlazeck et *al*., (2018) Sniffles par rapport aux autres outils, se caractérise par sa capacité à détecter des types de SVs complexes, tels que les duplications en tandem inversées (INVDUP). Cette particularité de Sniffles souligne son avantage pour les analyses nécessitant une détection précise de ces variants. Les SVs en commun détectés par CuteSV et Sniffles

sont enrichis par les délétions, suggérant que celles-ci sont faciles à détecter et à valider par les deux outils, tandis que la détection d'autres types de SVs peut être plus sensible aux algorithmes spécifiques à chaque méthode. Ces résultats soulignent l'importance de choisir l'outil de détection des SVs en fonction des besoins et des questions scientifiques. Pour des analyses nécessitant l'identification de SVs complexes, Sniffles offre des avantages significatifs. et pour les autres cas, CuteSV semble être le mieux adapté.

L'analyse des variants structuraux communs qui a été effectuée par les deux méthodes **Bcftools isec** et **Bedtools intersect** a donné des résultats différents. Par exemple, pour l'échantillon BB-2 du deuxième batch, CuteSV a identifié 24 467 SVs, tandis que Sniffles en a identifié 18 231. Les SVs communs trouvés par Bcftools étaient 4 269, contre 11 599 trouvés par bedtools, montrant une tendance similaire pour tous les échantillons. **Bedtools intersect** a pu identifier un plus grand nombre de variants communs en raison de ses critères de chevauchement plus larges et flexibles. Il considère un variant comme commun si les variants dans les fichiers se chevauchent avec un minimum de 1 paire de base. Ce qui peut inclure des variants qui ne correspondent pas exactement en termes de position mais chevauchent partiellement. Alors que **Bcftools isec** offre une correspondance plus stricte basée sur des critères bien spécifiques impliquant les positions exactes, ceci peut avoir pour effet de réduire le nombre de variants identifiés comme communs mais avec une plus grande précision en termes de séquence et de signification biologique.

Les résultats de comparaison des variants structuraux entre le Dogue Allemand et le Boxer montre que l'absence de SV dans la région de l'IGF1 pour le Dogue Allemand et sa présence dans l'échantillon de Boxer peut être cohérente avec l'étude de Sutter et *al.* (2007) concernant le rôle de l'IGF1 dans la détermination de la taille des chiens. Toutefois, cette différence pourrait également être due à l'hétérogénéité entre les deux échantillons, notamment la profondeur de séquençage, avec une profondeur de 11.38X pour le Boxer (BB) contre 4.86X pour le Dogue Allemand (AL). Pour confirmer ces résultats, il serait nécessaire de comparer davantage de races de tailles différentes et avec une profondeur de séquençage plus importante.

L'analyse de la longueur des variants structuraux montre un pic notable autour de 200 pb et un autre pic, moins important autour de 6000 pb. Ces résultats sont cohérents avec les résultats récemment publiés par Halo *et al*., (2021) qui ont trouvé des augmentations

significatives de variants structuraux à environ 200 pb et 6000 pb entre les assemblages CanFam3.1 et Zoey (Boxer *vs* Dogue Allemand) .

Pour déterminer si les pics observés sont consistants au sein d'une diversité génétique plus large et de confirmer les tendances détectées, il sera pertinent d'étendre l'analyse à un plus grand nombre de données, comme celles du projet GOLDogs avec 25 races. Cela permettra également de mieux comprendre l'impact des éléments mobiles (SINEs, LINEs) sur la variabilité génomique des différentes races de chiens, et ainsi de tirer des conclusions plus solides sur leurs rôles dans l'évolution des canidés.

4.2 Les défis rencontrés et les solutions proposées

Parmi les limites de cette analyse, nous notons 1/ le nombre restreint d' échantillons et 2/ la variabilité de types de séquençage et de méthodes d'extraction et de purification d'ADN. Cela rend les données non homogènes et complique l'interprétation de leurs résultats. Pour comparer efficacement les outils de détection des variants structuraux, et d'identifier des SVs impliqués dans la longévité canine, il sera nécessaire d'homogénéiser les conditions de prélèvements et d'extraction des ADN longs fragments des échantillons et d'augmenter leur nombre.

L'un des problèmes que j'ai rencontré est lié au stockage et à la saturation des disques du serveur Genouest. Même si l'équipe dispose de 72 To, un espace de seulement 4 To était disponible à mon arrivée pour plusieurs bio-informaticiens de l'équipe, rendant le travail plus fastidieux. Cela m'a obligé à lancer les tâches indépendamment, une par une, pour éviter de saturer le disque. Pour des échantillons de grande taille, comme ceux de 72 Go, l'exécution du pipeline nécessite énormément d'espace de stockage. Ainsi, le répertoire (work) créé par nextflow et qui conserve les traces du pipeline en cas de changement de paramètres ou de présence d'erreurs, peut facilement atteindre 2 To. Cela m'a obligée à supprimer toutes les traces de pipeline pour libérer de l'espace de stockage au fur et à mesure et pouvoir analyser tous les échantillons.

Pour un petit ensemble de données, le stockage peut ne pas avoir d'effet direct sur l'analyse, mais pour un grand ensemble de données, il est recommandé de prévoir un espace de stockage adapté.

Lors de l'exécution de pipeline nanoseq, spécifiquement en utilisant Sniffles comme variant caller. J'ai rencontré une erreur à la fin du processus indiquant que le contenu des fichiers (`.vcf`) générés par Sniffles présentait des problèmes, notamment des informations

manquantes ou mal définies empêchant ainsi leur tri par `Bcftools sort` les rendant donc incompatibles avec les outils d'indexation (`tabix ou bcftools index`). L'erreur suggère que le format de certaines lignes était incorrectement interprété. Plusieurs utilisateurs de ce pipeline ont rencontré le même problème et ont partagé leurs expériences sur le Slack de nf-core/nanoseq. Cette erreur est spécifique à la version de Sniffles utilisée par nf-core/nanoseq actuelle. Une correction est prévue prochainement avec la nouvelle mise à jour (3.2.0) du pipeline. Pour résoudre ce problème, j'ai récupéré les fichiers VCF incorrectement formatés issus de Sniffles à partir du répertoire (/work), je les ai reformatés, ensuite je les ai triés et indexés moi-même pour pouvoir les manipuler.

5. CONCLUSION

Durant mon stage, deux objectifs ont été visés : l'amélioration de la collecte et de l'enregistrement des données des échantillons canins, ainsi que la réalisation d'une analyse bioinformatique préliminaire/pilote pour le projet GOLDogs.

Pour le premier objectif, un formulaire en ligne a été créé, améliorant ainsi la vérification et optimisant la collecte et l'enregistrement des informations dans la base de données Cani-DNA.

Pour le deuxième objectif, l'analyse préliminaire/pilote des données de séquençage long read a démontré que le séquençage avec PromethION offre une qualité et une profondeur de séquençage meilleures par rapport à GridION, faisant de PromethION le choix idéal pour produire un ensemble plus large de données de séquençage. Pour cette analyse, le pipeline NanoSeq, implémenté avec Nextflow, a été appliqué, utilisant Minimap2 comme outil d'alignement et DeepVariant pour la détection des SNVs/SNPs et Indels, ces choix ont été basés seulement sur la documentation. Minimap2 s'est avéré performant pour l'alignement des données volumineuses issues du séquençage long read. Tandis que DeepVariant s'est avéré le plus précis pour les SNVs/SNPs et Indels, surpassant Medaka pour les données diploïdes. Concernant les variants structuraux, CuteSV a montré une supériorité en termes de rapidité et de sensibilité, détectant plus de variants que Sniffles, surtout pour les échantillons de grande taille et à haute profondeur de séquençage. L'analyse des tailles de SVs a révélé des résultats cohérents avec les travaux précédents, renforçant l'hypothèse de l'importance des SVs de type insertion/délétion d'éléments transposables dans l'étude de la génomique canine.

Enfin, il a été observé que la profondeur de séquençage et la méthodologie d'extraction d'ADN influencent les résultats de l'analyse des variants structuraux, démontrant l'intérêt du benchmarking en termes de méthodes de collecte et de préparation des ADN

6. PERSPECTIVES

Le développement des technologies de séquençage à longues lectures implique le développement en parallèle d'outils bioinformatiques capables d'analyser ces données, notamment dans le cadre de l'annotation des variations structurales. Ainsi, de nombreux outils existent actuellement pour annoter des SVs à partir de données longues lectures (séquençage ONT ou PacBio). Dans un souci de reproductibilité, nous avons opté pour l'utilisation du pipeline nf-core nanoseq qui implémente deux outils d'annotation de SVs : CuteSV et Sniffles. Une piste d'amélioration potentielle serait d'utiliser le programme EPI2ME (développé par ONT) pour faciliter les analyses bioinformatiques de routine. EPI2ME permet aux utilisateurs d'exécuter les pipelines Nextflow dans une application de bureau avec une interface graphique facile à utiliser. Il est conçu pour une installation et une utilisation facile afin de rendre l'analyse des données de séquençage aussi accessible que possible. EPI2ME utilise Sniffles2 au lieu de Sniffles, ce qui nous permettrait de bénéficier des dernières avancées technologiques de ce logiciel.

Pour pouvoir comparer les différents outils de détection des variants structuraux, il est essentiel de connaître le nombre exact de variants réels et leur position. Avec notre ensemble de données, composé d'échantillons de différentes races séquencés pour la première fois, il est difficile d'évaluer l'exactitude et les performances des outils de détection. C'est pour cette raison qu' il serait intéressant de simuler des données de séquençage afin de réaliser une étude comparative fiable des *SV caller*. Nous pourrions utiliser NanoSim, un simulateur de lecture rapide et évolutif qui capture les caractéristiques spécifiques des données ONT. NanoSim est capable de simuler des lectures de transcriptome ONT (ADNc/ARN direct) ainsi que des lectures génomiques. En simulant des lectures à partir d'un génome de référence et en créant des variants synthétiques, NanoSim offre une approche prometteuse pour étudier les variants structuraux et évaluer les performances des outils de détection (précision, sensibilité, etc.) selon différents critères (niveau de qualité des séquences, profondeur de séquençage, …).

En comparant CuteSV et Sniffles, les résultats ont montré que ces deux outils détectent principalement des délétions et des insertions. Pour une détection plus efficace des

inversions et des duplications, il est recommandé d'utiliser l'outil NanoVar lorsque ces types de SV revêtent une importance particulière.

RÉFÉRENCES

1. Greer KA, Canterberry SC, Murphy KE. Statistical analysis regarding the effects of height and weight on life span of the domestic dog. *Res Vet Sci*. 2007;82(2):208-214. doi:10.1016/j.rvsc.2006.06.005

2. Teng KT yun, Brodbelt DC, Pegram C, Church DB, O'Neill DG. Life tables of annual life expectancy and mortality for companion dogs in the United Kingdom. *Sci Rep*. 2022;12(1):6415. doi:10.1038/s41598-022-10341-6

3. Sutter NB, Bustamante CD, Chase K, et al. A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. *Science*. 2007;316(5821):112-115. doi:10.1126/science.1137045

4. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. doi:10.1038/nature15394

5. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704-712. doi:10.1038/nature08516

6. Dierckxsens N, Li T, Vermeesch JR, Xie Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol*. 2021;22(1):342. doi:10.1186/s13059-021-02551-4

7. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. *Annu Rev Med*. 2010;61(Volume 61, 2010):437-455. doi:10.1146/annurev-med-100708-204735

8. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597-614. doi:10.1038/s41576-020-0236-x

9. Marx V. Method of the year: long-read sequencing. *Nat Methods*. 2023;20(1):6-11. doi:10.1038/s41592-022-01730-w

10. Séquençage long-read : mécanismes, promesses et applications en Santé. Biotech.info. Accessed May 29, 2024. https://biotechinfo.fr/article/sequencage-long-read-mecanismes-promesses-et-applications-en-sante/

11. Séquençage à "longues lectures ONT." France Génomique. Accessed May 29, 2024. https://www.france-genomique.org/expertises-technologiques/genome-entier/sequencage-a-longues-lectures-ont/

12. Chao Wang et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity.file:///home/aitmahfoud/T%C3%A9l%C3%A9chargements/s42003-021-01698-x.pdf.

13. GUYONNET Alexandre. Etude bibliographique des anomalies génétiques impliquées dans les lymphomes non hodgkiniens canins.file:///home/aitmahfoud/T%C3%A9l%C3%A9chargements/2012lyon023.pdf. Published 2012.

14. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*. 2021;16(10):e0257521. doi:10.1371/journal.pone.0257521

15. NanoPlot - UFRC. Accessed May 30, 2024. https://help.rc.ufl.edu/doc/NanoPlot

16. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma Oxf Engl*. 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191

17. Gaever BV. Long read sequencing for the detection of cryptic structural variation in patients with intellectual disability and congenital anomalies.

18. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single molecule sequencing. *Nat Methods*. 2018;15(6):461-468. doi:10.1038/s41592-018-0001-7

19. Ubuntu Manpage: bcftools - utilities for variant calling and manipulating VCFs and BCFs. Accessed June 9, 2024. https://manpages.ubuntu.com/manpages/focal/en/man1/bcftools.1.html

20. Shafin K, Pesout T, Chang PC, et al. Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. Published online March 5, 2021:2021.03.04.433952. doi:10.1101/2021.03.04.433952

ANNEXES

Annexe.1 *PDF du formulaire en papier avant sa transformation en formulaire en-ligne (https://partage.univ-rennes1.fr/service/home/~/?auth=co&loc=fr&id=2489&part=2.2)*

Tableau 4. Outils et versions utilisés pour le deuxième objectif.

Outils	Description	Version	Github
Nextflow	Gestionnaire de workflows	23.10.0	https://github.com/nextflow-io/nextflow
nf-core/nanoseq	Workflow pipeline	3.1.0	https://github.com/nf-core/nanoseq/tree/3.1.0
Nanoplot	Outil de visualisation pour les données de séquençage et les alignements à lecture longue	bioconda::nanoplot=1.41.0	https://github.com/wdecoster/NanoPlot
multiqc	Outil de création des rapports avec des parcelles interactives pour plusieurs analyses bioinformatiques	bioconda::multiqc=1.14	https://github.com/MultiQC/MultiQC
Minimap2	Aligneur	2.17-r941	https://github.com/lh3/minimap2
deepvariant	Outil de détection et d'appel des SNPs et SNVs	1.4.0	https://github.com/google/deepvariant
bedtools	Outil pour manipuler et analyser des données génomiques au format BED	2.29.2	https://github.com/arq5x/bedtools2
samtools	Outil pour manipuler et analyser des données génomiques au format	1.13	https://github.com/samtools

	SAM/BAM		
CuteSV	Outil de détection et d'appel de variations structurales	1.0.12	https://github.com/tjiangHIT/cuteSV
Sniffles	Outil de détection et d'appel de variations structurales	1.0.12	https://github.com/fritzsedlazeck/Sniffles
bcftools	Outil pour manipuler et analyser des données génomiques au format VCF	1.9	https://github.com/samtools/bcftools

Tableau 5. Outils et versions utilisés pour le premier objectif.

Librairie	**Usage**	**Github**
Difflib (sequencemacher())	Comparer des séquences, afin de trouver les similarités et différences entre elles.	https://github.com/python/cpython/blob/main/Lib/difflib.py
openxls	Lire et écrire des fichiers Excel (format .xlsx).	https://github.com/troldal/OpenXLSX
unidecode	Convertir des caractères non-ASCII (comme les caractères accentués, les caractères de langues non-latines, etc.) en une représentation ASCII approximative.	https://github.com/unicode-org/icu
gspread	Interagir avec Google sheets.	https://github.com/burnash/gspread
Pandas	Manipulation et analyse des données	https://github.com/pandas-dev/pandas

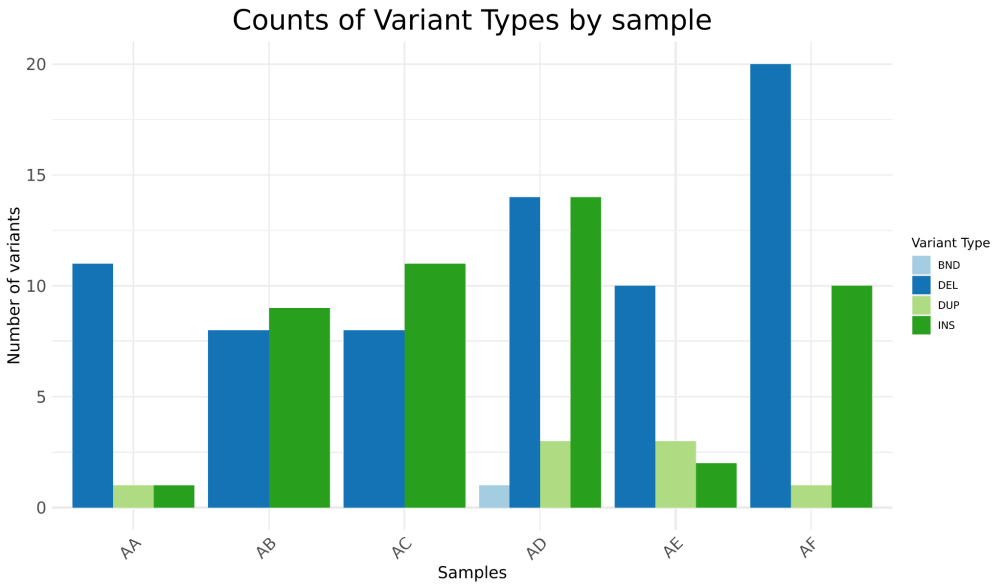

Figure 20. Répartition des types de variants structuraux détectés par CuteSV dans les échantillons du premier lot.

1st batch Sniffles

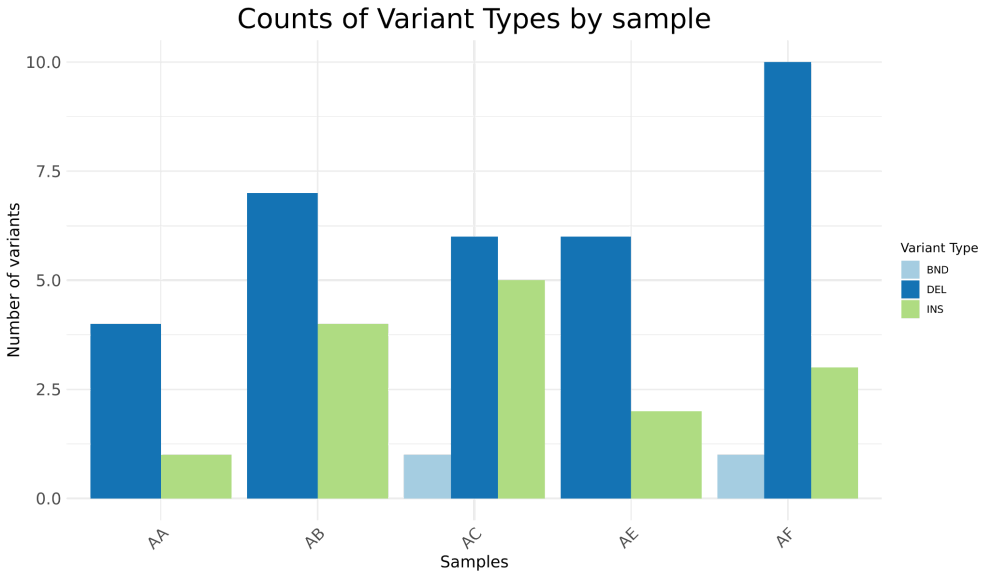

Figure 21. Répartition des types de variants structuraux détectés par Sniffles dans les échantillons du premier lot.