UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 605
Biologie, Santé
Spécialité : *Génétique, génomique, bioinformatique*

Par
Camille KERGAL

Méthode d'apprentissage profond pour l'analyse génomique des cancers canins comme modèles des cancers humains

Thèse présentée et soutenue à Rennes, le 18 octobre 2022
Unité de recherche : Institut de Génétique et Développement de Rennes

Rapporteurs avant soutenance :

Anita BURGUN	Professeure des Universités et Praticienne Hospitalier, INSERM - Unité 1138
Hugues ROEST CROLLIUS	Directeur de Recherche, CNRS - UMR 8197, IBENS

Composition du Jury :

Président :	Hugues ROEST CROLLIUS	Directeur de Recherche, CNRS - UMR 8197, IBENS
Examinateurs :	Christophe ANTONIEWSKI	Directeur de Recherche, ARTbio - IBPS
	Anita BURGUN	Prof. des Universités et Prat. Hospitalier, INSERM - Unité 1138
	Anna-Sophie FISTON-LAVIER	Maîtresse de conférence, ISEM - UMR CNRS - IRD - UM 5554
	Marie-Dominique GALIBERT	Prof. des Universités et Prat. Hospitalier, UR1 - CHU
	Andrea RAU	Chargée de Recherche, INRA - UMR 1313
Dir. de thèse :	Christophe HITTE	Ingénieur de Recherche, CNRS - IGDR
Co-dir. de thèse :	Thomas DERRIEN	Chargé de Recherche, CNRS - IGDR

RÉSUMÉ

Les méthodes d'apprentissage profond (DL) se sont récemment révélées être de puissantes stratégies pour prédire l'activité régulatrice d'une séquence génomique et donc pour, *in fine*, évaluer l'impact des mutations régulatrices sur l'expression des gènes. L'outil Basenji propose une approche DL utilisant des réseaux de neurones convolutifs pour prédire le niveau d'expression de gènes humains. Nous avons adapté ce programme pour entraîner un modèle d'expression génique spécifique au chien et montré que ce modèle de prédiction atteignait des performances similaires à celles observées chez l'homme, avec des corrélations élevées entre les niveaux d'expression réels et ceux prédits (R=0,66). Pour prédire le niveau d'expression de gènes canins, nous démontrons également que l'utilisation du modèle de prédiction canin (approche intra-espèce) aboutit à de meilleures performances que le modèle humain (approche inter-espèce), notamment en lien avec certaines caractéristiques spécifiques aux séquences canines (niveau de GC, d'éléments transposable et conservation évolutive). Le chien étant un modèle naturel pour l'étude des cancers humains, nous avons également exploité ces modèles pour prédire l'impact de mutations non-codantes sur l'expression de gènes impliqués dans les cancers. Nous avons ainsi localisé 1301 mutations communes entre l'homme et le chien, suggérant un rôle fonctionnel dans la régulation de l'expression de gènes impliqués dans les cancers. Finalement, nos modèles et les outils pour les exploiter sont disponibles sur GitHub : https://github.com/ckergal/BLIMP.

Abstract

Deep learning (DL) methods have recently been shown to be powerful strategies for predicting the regulatory activity of a genomic sequence and thus for ultimately assessing the impact of regulatory mutations on gene expression. The Basenji tool proposes a DL approach using convolutional neural networks to predict the expression level of human genes. We adapted this program to train a dog-specific gene expression model and showed that this model achieved similar performance to that observed in humans, with high correlations between real and predicted expression levels (R=0.66). To predict the expression level of canine genes, we show that the canine prediction model (within-species approach) leads to better performances than the human model (cross-species approach), particularly due to some specific features of canine sequences (GC content, transposable elements and evolutionary conservation). As the dog is a spontaneous model for human cancers, we used these models to predict the impact of non-coding mutations on the expression of genes involved in cancers. We identified 1301 common mutations to both humans and dogs, suggesting a functional role in the regulation of the expression of genes involved in cancer. Finally, models and tools to exploit them are available on GitHub : https://github.com/ckergal/BLIMP.

REMERCIEMENTS

Je souhaite tout d'abord remercier l'ensemble des membres du jury. Merci à Anita Burgun et Hugues Roest Crollius pour avoir accepté le rôle de rapporteur. Merci à Christophe Antoniewski, Anna-Sophie Fiston-Lavier et Andrea Rau pour leur implication en tant qu'examinateur.

Merci également aux membres de mon comité de suivi individuel, Frédéric Chalmel, Charles-Henri Lecellier et Yann Le Cunff pour leurs précieux conseils et leurs encouragements au cours de nos deux rencontres.

Lors de cette thèse, j'ai eu la chance d'être encadrée par le brillant trio composé par Thomas Derrien, Marie-Dominique Galibert et Christophe Hitte. Je vous remercie sincèrement de m'avoir donné l'opportunité de réaliser ce projet à vos côtés.

Je remercie les membres du consortium DoGA pour avoir accepté la collaboration nous permettant de disposer des données essentielles à la réalisation des travaux liés à cette thèse. Les différents échanges que nous avons pu avoir étaient toujours très constructifs et l'intérêt porté à ce projet fut très stimulant.

Je souhaite également remercier les différents membres de l'équipe pédagogique de l'UFR SVE avec qui j'ai pu avoir l'occasion d'échanger lors des missions d'enseignement. Je pense notamment à Emmanuelle Becker, Kévin Da Silva, Marine Jacquier et Yann Le Cunff. J'ai beaucoup apprécié cette expérience qui, de plus, permettait de s'évader des conditions de télétravail.

Je pense aussi aux différents membres du comité d'organisation des journées scientifiques de l'Ecole Doctorale. L'organisation de cet évènement était laborieuse, parfois décourageante, à cause des changements incessants dus à la situation que l'on connaît tous, mais souvent très drôle finalement.

Je tiens évidemment à remercier l'ensemble de l'équipe "Génétique du Chien", les personnes qui y sont actuellement mais aussi celles qui ont poursuivi leur chemin d'une

autre manière. Merci à Christophe et Thomas, pour votre patience, votre pédagogie, vos encouragements et votre bienveillance. Merci aussi à toutes les autres personnes que j'ai eu la chance de croiser, membres de l'équipe ou non. J'ai passé d'excellents moments avec vous. Que ce soit pour partager un bureau, un café, une partie de baby-foot, un verre au MeM ou à La Piste, un repas à la cantine ou dans le jardin d'une île du golfe du Morbihan ou même un voyage à Barcelone, j'ai passé trois belles et très drôles années à vos côtés.

Pour finir je souhaite remercier l'ensemble de mes proches, amis et famille. Je sais que je peux compter sur vous pour m'accompagner quelle que soit la voie que je souhaite emprunter.

TABLE DES FIGURES

1.1 Schéma des technologies RNA-seq et CAGE-seq 21
1.2 Schéma simplifié du mode d'action des facteurs de transcription 24
1.3 Mutations du promoteur de *TERT* 25
1.4 Principe d'une étude d'association pangénomique 27
1.5 Intérêt du chien domestique 32
1.6 Schéma simplifié de l'histoire évolutive du chien. 33
1.7 Diversité de l'espèce canine 35
1.8 Principales caractéristiques biologiques du cancer 40
2.1 Répartition du nombre de lectures selon la version d'assemblage 51
2.2 Taux de GC des séquences promotrices des gènes de cancers selon l'espèce 56
2.3 Répartition du taux de conservation des séquences. 57
2.4 Principe de l'encodage one-hot 58
2.5 Schéma des profils CAGE 58
2.6 Convolutions de l'algorithme 60
2.7 Évaluation du modèle de prédiction 62
3.1 Entraînement du modèle canFam3 67
3.2 Entraînement du modèle canFam4 69
3.3 Évaluation des modèles de prédiction 71
3.4 Optimisation des modèles de prédiction canins 73
3.5 Évaluation des modèles de prédiction après optimisation 74
3.6 Visualisation des prédictions d'expression 75
3.7 Principe des approches inter-espèce et intra-espèce 77
3.8 Évaluation des approches inter-espèce et intra-espèce 79
3.9 Impact des caractéristiques génomiques sur la performance de l'approche inter-espèces 82
3.10 Mutagenèse saturée de *TERT* 84
3.11 Distance des mutations par rapport au TSS 86
3.12 Impacts sur le niveau d'expression 87
3.13 Analyse des mutations en fonction des tissus 89

3.14 Répartition des mutations communes par tissus 91
4.1 Score d'impact Basenji humain 101

Liste des abréviations

CAGE Cap Analysis of Gene Expression . . . 21

CDS séquence codante (ou Coding DNA Sequence) . . . 22

CNN réseaux de neurones convolutifs (Convolutional Neural Network) . . . 15

DL apprentissage profond (Deep Learning) . . . 29

ENCODE ENCyclOpedia of DNA Element . . . 21

ETS Erythroblast Transformation Specific . . . 23

FANTOM Functional ANnoTation Of the Mammalian genome . . . 21

GWAS études d'association pan-génomiques (Genome Wide Association Studies) . 22

ISM mutagenèse saturée *in silico* . . . 83

ML apprentissage automatique (Machine Learning) . . . 28

NGS Next Generation Sequencing . . . 19

RNASeq séquençage de l'ARN . . . 20

SNP changements ponctuels de nucléotides (Single Nucleotide Polymorphism) . . . 19

SV variants structuraux . . . 19

TF facteurs de transcription (Transcription Factors) . . . 23

TFBS sites de fixations des facteurs de transcription (Transcription Factor Binding Sites) 23

TGS Third Generation Sequencing . . . 19

TSS site d'initiation de la transcription (Transcription Start Site) . . . 42

WGS séquençages complets de génomes (Whole Genome Sequencing) . . . 38

TABLE DES MATIÈRES

Résumé **3**

Abstract **5**

Remerciements **7**

Table des illustrations **10**

Liste des abréviations **11**

Introduction **15**

1.1 Contexte 15

1.2 Le deep learning et la génomique 18

1.2.1 L'essor de la génomique 18

1.2.2 Les avancées et l'évolution de la bioinformatique 25

1.3 Le chien comme modèle spontané des cancers de l'homme 31

1.3.1 L'intérêt du modèle canin en génétique 31

1.3.2 Les prédispositions du chien à différentes maladies génétiques . . 32

1.3.3 Le projet Dog10K 37

1.3.4 L'oncologie comparée entre l'homme et le chien 39

1.4 Objectifs des travaux de thèse 42

Matériels et méthodes **45**

2.1 Description des données 45

2.1.1 Assemblages de référence 45

2.1.2 Profils d'expression CAGE-seq 47

2.1.3 Panel de gènes impliqués en cancérologie 53

2.2 Algorithme de prédiction 55

2.2.1 Intégration des données 55

2.2.2 Entraînement du modèle de prédiction 59

2.2.3 Evaluation du modèle de prédiction 61

2.2.4 Méthode d'optimisation des modèles de prédictions 61

Résultats **65**

3.1 Prédiction du niveau d'expression des gènes canins 65
3.1.1 Création des modèles 65
3.1.2 Mesure des performances 68
3.1.3 Optimisation des hyperparamètres 70
3.2 Comparaison des approches inter-espèce et intra-espèce 76
3.2.1 Approches inter-espèces 76
3.2.2 Impact des caractéristiques génétiques sur les performances du modèle 78
3.3 Impact des mutations régulatrices sur le niveau d'expression des gènes . . 81
3.3.1 Mutagenèse saturée *in silico* 83
3.3.2 Analyses des mutations prédites comme impactantes 85

Discussion **95**

4.1 Résumé des travaux 95
4.2 Apport de nos travaux à la littérature 97
4.2.1 Modèles de prédictions du niveau d'expression des gènes canins . 97
4.2.2 Développement de l'outil BLIMP 97
4.3 Limites de nos travaux 98
4.3.1 Choix de l'algorithme d'entraînement 98
4.3.2 Validations expérimentales 99
4.3.3 Approches et limites de l'oncologie comparée 100
4.4 Perspectives 102

Bibliographie **105**

Annexes **117**

INTRODUCTION

1.1 Contexte

La recherche en bioinformatique s'approprie à une vitesse croissante les méthodes d'apprentissage automatique pour analyser et interpréter les données génomiques [1]. Ces travaux ouvrent la voie à de nouveaux axes de recherche en biologie en général et en médecine, notamment en matière de cancers. La génomique est désormais un pilier en recherche médicale, pour laquelle l'analyse des génomes des individus permet de renseigner les prédispositions génétiques à certaines maladies, et alors permettre d'agir avant leur développement [2]. Ainsi, dès les années 2000, des équipes ont cherché à analyser les données génétiques en s'appuyant sur les technologies de traitement du langage de l'époque [3]. Mais ce n'est véritablement qu'à partir des années 2010 que l'apprentissage automatique a accéléré ces travaux. De multiples formes d'algorithme ont été mises en œuvre, comme les réseaux de neurones convolutifs (Convolutional Neural Network) (CNN), qui ont connu un certain succès dans l'analyse de l'ADN des tumeurs [4]. Le cancer, pour lequel on compte 157 400 décès en 2018, est la première cause de mortalité en France, et la deuxième dans le monde avec près de 10 millions de morts par an [5].

Un cancer peut toucher des tissus de différentes natures, être diagnostiqué à plusieurs stades et être défini par des cellules de formes et de structures différentes. Ce contexte implique l'existence de nombreux types de cancers mais ils partagent néanmoins le même processus biologique. En effet, les cancers d'origine génétique se développent à partir d'une anomalie conduisant à l'apparition de cellules anormales qui vont proliférer de manière incontrôlée. On entend par anomalie, l'altération d'un ou plusieurs gènes au sein des cellules constituant un organisme vivant. Les causes de ces altérations, favorisant le développement des cancers, sont nombreuses et disparates. Comme exemples nous pouvons citer l'âge, les radiations, l'alimentation, la fumée du tabac, l'exposition à des

agents chimiques comme l'alcool ou bien des agents biologiques comme les papillomavirus humains. L'apparition d'un cancer peut également être dûe à la présence d'une anomalie génétique héritée directement d'un parent ou formée par la combinaison du patrimoine génétique des deux parents [6].

L'aboutissement des techniques et des appareils de mesure dans le domaine de la recherche médicale permet aujourd'hui d'identifier et de reconnaître les anomalies génétiques responsables de l'apparition de cancers. En revanche, dans un objectif d'amélioration et même de personnalisation des préventions et des thérapies, il apparaît nécessaire d'enrichir notre compréhension et de saisir le sens des mécanismes biologiques conduisant à la formation de cellules cancéreuses.

Dans cette perspective, nous pouvons mettre à profit les récents progrès en bioinformatique et en intelligence artificielle dédiés à la recherche en génomique, notamment à l'analyse des données génétiques tumorales, et ainsi améliorer le diagnostic et le traitement des cancers [7]. En effet, les données acquises par le biais de séquençage haut débit de génomes représentent une source d'information notable, à la fois conséquente et précise, méritant d'être analysée par des méthodes avancées d'algorithmique. En effet, les chercheurs ont produit ces dernières années des données de séquences extrêmement importantes [8-11]. Ces travaux nous permettent d'obtenir des outils capables de produire en quantité des profils génétiques pertinents afin de constituer ou compléter les jeux de données indispensables à l'entraînement des modèles. Cependant, le manque de données réelles incarne un problème récurrent pour ce type de recherche. Leur utilisation est particulièrement encadrée par les lois de protection des données personnelles et les coûts de production, notamment pour les espèces non-modèles, restent importants.

L'utilisation de procédés automatiques de modélisation mathématique est indispensable afin de déduire les caractéristiques les plus pertinentes et révélatrices des bases de données établies en cancérologie. Un autre avantage de l'utilisation des méthodes informatiques dans la recherche en génétique des cancers réside dans l'abstraction de barrières éthiques de la recherche sur l'être humain. De fait, la modélisation mathématique des processus biologiques permet de simuler les effets de tests, tels que la simulation de mutations génétiques, leurs impacts sur la biologie de la cellule ou d'essais thérapeuthiques, qu'il ne serait pas envisageable de réaliser de manière expérimentale.

En complément des informations disponibles pour les cancers chez l'homme, il est essentiel d'analyser si ces mêmes cancers peuvent survenir chez d'autres espèces. En effet,

les approches de recherche en pathologie comparée se sont développées ces 40 dernières années, avec le processus de l'étude des maladies humaines, ainsi que des maladies des êtres vivants qui entourent l'homme, qu'il s'agisse d'animaux de production ou de compagnie [12, 13]. Les états pathologiques ont d'abord été observés, puis scientifiquement étudiés, comparés entre eux selon leurs facteurs étiologiques, et également comparés entre les différentes espèces animales, y compris l'homme. Dans le domaine du cancer, l'oncologie comparée, plus récente, a évolué de manière similaire, par l'accumulation de données, qui une fois enregistrées et analysées ont conduit à la conclusion qu'il existe parfois des similitudes, voire même des processus identiques de prédispositions génétiques et de développement tumoral, et d'autres fois une variabilité spécifique entre les données humaines et les tumeurs animales. La recherche scientifique a d'abord permis de bien définir les terminologies de tumeur ou néoplasme comme un groupe de lésions qui se caractérisent par des proliférations, au sein d'un tissu, de manière anormales des cellules génétiquement modifiées. En général, les processus de prolifération vont dominer la capacité et la vitesse de tous les processus intervenant dans la réparation, la régénération ou d'inflammation. Puis, les études comparées sur les tumeurs spontanées ont déterminé des corrélations sur les conditions d'apparition des néoplasmes chez diverses espèces avec des éléments extrinsèques communs à l'homme et à l'animal, tels que la zone géographique, la nutrition, l'âge, l'impact des niveaux hormonaux, les facteurs de pollution et également des facteurs intrinsèques aux tissus et plus généraux liés à l'organisme.

Concernant les facteurs oncogéniques extrinsèques, la recherche utilise des espèces animales appelées « animaux sentinelles », notamment le chien, ayant une sensibilité particulière bien supérieure à celle de l'homme. Le développement des tumeurs est complexe, il inclut les différentes composantes d'un processus biologique soumis à un état et à une évolution permanents, avec des manifestations physiologiques et morphologiques spécifiques particulières, qui évoluent dans le temps. Le processus tumoral est aussi caractérisé par des éléments communs à la majorité des tumeurs, tels que le comportement biologique, l'autonomie, la prolifération incontrôlée, le caractère invasif et métastasant, la formation de nouveaux clones cellulaires et l'irréversibilité. Une autre série de propriétés communes sont les aspects moléculaires et ultra-structuraux, et les aspects histologiques et macroscopique du tissu tumoral. Au cours des années, l'oncologie comparée a prouvé toute sa pertinence comme démarche scientifique, avec l'approfondissement des connaissances sur les maladies cancéreuses chez l'homme et l'animal, soulignant qu'il existe peu de types de tumeurs qui ne soient pas communes à l'homme et à l'animal, et que les traits

communs sont indéniablement plus nombreux et essentiels que ceux qui les différencient.

1.2 Le deep learning et la génomique

1.2.1 L'essor de la génomique

La connaissance du génome

La génomique se rapporte à la science des génomes formés par l'information génétique contenue dans chaque cellule des êtres vivants. L'acide désoxyribonucléique, ou ADN, est une molécule constituée de nucléotides qui permet de stocker l'information génétique sur les chromosomes des cellules. Chaque nucléotide est constitué d'un sucre (le désoxyribose), d'un acide phosphorique ainsi que d'une base azotée parmi l'adénine, la cytosine, la guanine et la thymine, respectivement notées A, C, G et T. Toutes les cellules d'un organisme possèdent a priori le même code ADN. Elles vont être différenciées par un rôle, une forme et une fonction bien distincts grâce aux gènes et à tous les éléments fonctionnels inscrits dans ce patrimoine ADN commun qui vont être exprimés par endroit et éteints à d'autres à des stades de développement différents. La génomique permet donc l'étude de l'ADN d'un organisme en analysant sa structure, son organisation et son fonctionnement.

Cette discipline de la biologie s'est développée à partir des premiers séquençages de génomes. En 1977, le phage phiX174 devient le premier organisme séquencé [14]. Ce virus génétiquement simple est composé de 5386 nucléotides formant 11 gènes. Il faudra ensuite attendre jusqu'en 2001 pour que soient publiés les premiers résultats du projet "Génome Humain" [15]. Puis en 2003, le séquençage du génome humain fut accompli [16], nécessitant l'implication de nombreuses équipes scientifiques à travers le monde et un coût financier estimé à 3 milliards de dollars. Cette première ébauche du génome humain a été assemblée à partir de l'ADN de quelques donneurs. Il ne s'agit donc pas du génome d'un individu en particulier et il n'est pas représentatif de l'humanité. En revanche, il s'agit du premier génome de référence pour l'homme. Compte tenu de la technologie utilisée, la réalisation de cette avancée scientifique précieuse et remarquable a nécessité plus d'une dizaine d'années. Il se trouve que, l'application de la méthode de séquençage Sanger [17], bien qu'étant incontournable pour l'accomplissement de ce

projet, n'offrait pas une rapidité d'exécution pour des génomes aussi conséquents que celui de l'homme. En effet, la séquence du génome humain compte plus de 3 milliards de nucléotides (ou paires de bases) répartis le long de 23 chromosomes.

Enfin, depuis le 1er avril 2022, le consortium T2T (Telomere-to-Telomere) propose la plus récente version de l'assemblage du génome humain et permet d'en exploiter la séquence complète sans gap de séquence [18].

La connaissance d'un génome permet de comparer et identifier les caractéristiques et les différences génétiques entre les individus au sein ou entre populations/espèces et ainsi caractériser ce qui différencie un individu d'un autre. En effet, entre plusieurs individus d'une même espèce, la séquence du génome est variable. Chez l'homme, cette variabilité moyenne entre deux individus sélectionnés aléatoirement s'élève à 0,1% [19]. Cette variabilité est due à des changements ponctuels de nucléotides (Single Nucleotide Polymorphism) (SNP) dans la séquence, et à des variants structuraux (SV) qui sont des variants affectant des fragments de génome d'une taille supérieure à 50 paires de bases, comportant les inversions et les translocations qui ne modifient pas le nombre de bases du génome, tandis que les insertions, les délétions et les duplications augmentent ou diminuent la taille du génome.

Les techniques actuelles

L'investissement scientifique et technique considérable lié au programme de séquençage du génome de l'homme a permis aux laboratoires de recherche d'accumuler l'expérience et le savoir-faire nécessaires à l'expansion des méthodes à d'autres génomes de mammifères. C'est ainsi que furent séquencés les génomes de la souris, du rat, du chien ou encore du chimpanzé pour un coût moins élevé et un temps plus court [20-23]. Puis à partir 2005, l'avènement de technologies plus rapides et moins coûteuses appelées Next Generation Sequencing (NGS) [24], vont permettre de multiplier les grands projets de séquençage. Enfin, à partir des années 2010 sont apparues les méthodes dîtes Third Generation Sequencing (TGS) [25], ou longues lectures. Aujourd'hui, le coût et le temps nécessaires à l'obtention d'une séquence du génome sont extrêmement réduits et permettent la réalisation de nombreuses expériences.

C'est dans ce contexte que s'est inscrit le projet 1000 génomes humains [26], ayant comme objectif de séquencer le génome d'au moins 1000 personnes d'origines ethniques

différentes et ainsi établir un premier catalogue des variations génétiques humaines. Au total, cette initiative internationale a permis d'analyser plus de 2500 génomes et de décrire plus 88 millions de variations [27]. Aujourd'hui, les instances de santé publiques de nombreux pays conçoivent des projets de génomique des populations à grande échelle pour obtenir des informations susceptibles de contribuer à améliorer les soins de santé. Aux États-Unis, le programme de recherche "All of us" du National Institute of Health [28] recueille des données sur au moins un million d'Américains afin de créer une base de données unique pour la médecine de précision et la recherche en génomique. Au Royaume-Uni, la "UK BioBank" [29] a recruté près de 500 000 personnes pour étudier des maladies graves combinant dossiers médicaux, données d'imagerie et de génétique. De nombreux autres exemples existent à différentes échelles, tels que l'étude FinnGen en Finlande [30], en Asie, le GenomeAsia 100K consortium [31] qui se concentre sur la médecine de précision dans la population asiatique, ou en Australie avec le Mission Genomics Health Futures [32] investissement dans la recherche australienne en génomique. En France, le plan Médecine France Génomique [33] se concentre sur la médecine et l'oncologie avec comme objectif de séquencer l'équivalent de 18 000 génomes par an, puis d'interpréter cette information biologique pour la comprendre et la rendre utile à tous les patients concernés.

La valeur de ces initiatives s'avère extraordinaire. Par ailleurs, la diminution continue des coûts et l'amélioration des technologies ne cessent de stimuler la recherche en génomique, qui permet une meilleure compréhension des maladies et propose de meilleures mesures de prévention [2]. Au-delà des initiatives gouvernementales publiques, il existe des sociétés de tests génétiques qui s'adressent directement aux consommateurs, telles que 23andMe et AncestryDNA, pour avoir des bases de données beaucoup plus volumineuses que tous les projets de recherche publique. Les chiffres des données de ces sociétés sont considérables avec pour AncestryDNA plus de 15 millions de personnes, et plus de 10 millions de clients pour 23andMe.

De la génomique à la transcriptomique

Les méthodes de séquençage de nouvelle génération, initialement développées pour faciliter l'étude des génomes, ont également permis de révolutionner l'étude du transcriptome [34]. Le séquençage de l'ARN (RNASeq) permet de connaître l'ensemble des transcrits d'une cellule, d'un tissu ou d'un organisme à un stade de développement donné.

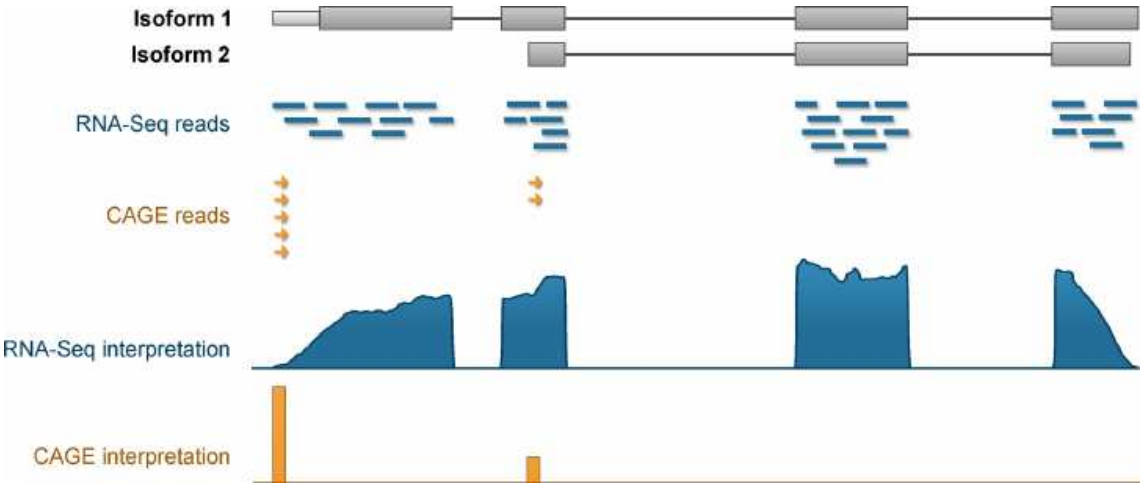

Figure 1.1 – Schéma des technologies RNA-seq et CAGE-seq. La technologie RNA-seq permet de capturer l'expression de la totalité des transcrits lorsque la technologie CAGE-seq se focalise sur la partie 5' des transcrits. Depuis Yu et *al*, 2015 [36].

Des consortiums internationaux se sont formés afin d'assurer l'annotation fonctionnelle du génome de plusieurs espèces. Nous pouvons citer le projet ENCODE [8], le projet GTEx [10] ou encore le projet FANTOM [35]. Ces consortiums ont permis d'établir d'imposantes bases de données ouvrant la voie à une meilleure compréhension du nombre, du rôle et de la fonction des gènes ainsi que les mécanismes de régulations sous-jacents.

Lancé par l'institut national de recherche sur le génome humain (USA, National Human Genome Research Institute), le projet de recherche publique ENCyclOpedia of DNA Element (ENCODE) [8] vise à identifier et caractériser les éléments fonctionnels du génome humain. Le consortium propose une base de données conséquente composée, entre autres, d'échantillons de séquençage de type RNAseq [34] (Fig. 1.1). Cette technologie consiste à identifier et quantifier l'ARN issu de la transcription du génome permettant notamment de déterminer la structure des gènes et leur composition en exons (portions de gène transcrites en ARN) et introns (portions non traduites en protéines).

L'institut de recherche japonais RIKEN a initié le projet Functional ANnoTation Of the Mammalian genome (FANTOM) [35] proposant l'identification de l'ensemble des sites d'initiation de la transcription le long du génome, par application de la méthode Cap Analysis of Gene Expression (CAGE) [9]. Cette technologie de séquençage de l'ARN permet de capturer la partie 5' (la coiffe) des transcrits suggérant un potentiel régulateur et dont l'activité est directement mesurable par la quantité d'ARN produit. Les régions

identifiées dans le cadre de ce projet permettent d'affiner la prédiction des TSSs et sont donc exploitables dans le but de définir les régions régulatrices d'intérêt en amont des gènes (promoteurs).

Enfin, le programme Genotype-Tissue Expression (GTEx) [10] est dédié à la création d'une banque de données et de tissus dont l'objectif est d'étudier la relation entre les variants génétiques et l'expression génique dans plusieurs tissus humains et entre individus. Les données GTEx sont utilisées pour améliorer l'interprétation fonctionnelle des résultats des études d'association pan-génomiques (Genome Wide Association Studies) (GWAS) et pour l'identification de gènes pertinents dans le cadre d'une maladie. L'ambition du programme GTEx repose sur le développement d'une biobanque d'échantillons biologiques de tissus ainsi que d'ARN, d'ADN, d'échantillons de sang et de lignées cellulaires d'environ 960 donneurs. Les données et échantillons sont conservés au Broad Institute de Harvard et au MIT (Massachusetts Institute of Technology). GTEx a nécessité la création d'une ressource de données en ligne (portail GTEx : https://gtexportal.org/home/) pour le stockage, l'inventaire et le partage des données agrégées. Il est important de souligner que les données génétiques issues des vastes ressources telles que GTEx sont aussi utilisées comme un ensemble de données de référence bien décrites et formalisées, qui jouent un grand rôle dans la conception et la réalisation de nouvelles méthodes statistiques et outils informatiques.

Lien entre mutations régulatrices non-codantes et expression des gènes

Suite aux résultats obtenus par le biais du séquençage du génome humain, les projets d'envergure internationale comme les consortiums ENCODE ou FANTOM ont permis d'explorer les fonctions du génome et notamment du génome non-codant. Ces programmes mettent à disposition de la communauté scientifique des données précieuses. En effet, de nombreuses études mettent en évidence le fait que les mutations survenant dans les régions non-codantes du génome sont susceptibles de jouer un rôle considérable dans les maladies complexes [37, 38]. Ces régions régulatrices présentent des signes de sélection négative suggérant un rôle important au sein des cellules et que des mutations survenant dans ces régions entraînent des effets délétères. De plus, des résultats issus d'études GWAS indiquent que la plupart des loci génomiques statistiquement associés à des phénotypes particuliers n'apparaissent pas au sein d'une séquence codante (ou Coding DNA Sequence) (CDS) [39]. L'interprétation fonctionnelle et la hiérarchisation

des variants non codants représentent un défi persistant, et les variants non codants impactant des traits et des maladies demeurent pour une grande partie non identifiés.

Les facteurs de transcription (Transcription Factors) (TF) jouent un rôle central dans la régulation de l'expression des gènes en se liant à des motifs spécifiques de courtes séquences d'ADN, appelées sites de fixations des facteurs de transcription (Transcription Factor Binding Sites) (TFBS). Ces séquences d'ADN se trouvent dans les promoteurs (et les enhancers) des gènes, des régions impliquées dans la régulation de l'expression et situées généralement en amont de la séquence codante des gènes (Fig. 1.2). Ainsi, une mutation survenant dans ces régions peut avoir de lourdes conséquences. En effet, elle peut suffire à la création d'un motif de fixation de facteur de transcription et ainsi conduire à l'activation, l'expression d'un gène qui ne l'était pas initialement ou réciproquement supprimer un TFBS et abolir l'expression d'un gène important.

L'exemple de mutations dans une région promotrice qui vont influencer l'expression de gène avec un rôle majeur dans les cancers est bien illustré par le gène *TERT*. En effet des mutations récurrentes ou "hotspot" du promoteur de *TERT* ont été identifiées dans divers types de cancers tels que le mélanome cutané ou le cancer de la vessie [40]. Ces mutations entraînent la génération de motifs de liaison consensus *de novo* pour des TFs de la famille des Erythroblast Transformation Specific (ETS) à travers lesquels la transcription de *TERT* et la télomérase sont activées (Fig. 1.3). Les ETS sont une famille de facteurs de transcription spécifiques à la transformation E26 des mammifères qui jouent un rôle essentiel dans le développement, la différenciation cellulaire, la prolifération, l'apoptose et l'oncogenèse. Les mutations du promoteur de *TERT*, identifiées à l'origine dans les mélanomes malins sporadiques et familiers, se produisent principalement à deux points chauds du chromosome 5 à -124 pb et -146 pb du TSS avec une transition dipyrimidine de type cytidine-thymidine (C $>$ T), nommées respectivement C228T et C250T. Les mutations C228T ou C250T créent des motifs de liaison ETS *de novo* nécessaires à l'activation de la transcription de *TERT* et de la télomérase [41]. Cet exemple de mutations régulatrices dans le promoteur de *TERT* nous a permis de tester les développements de notre outil au cours de nos travaux.

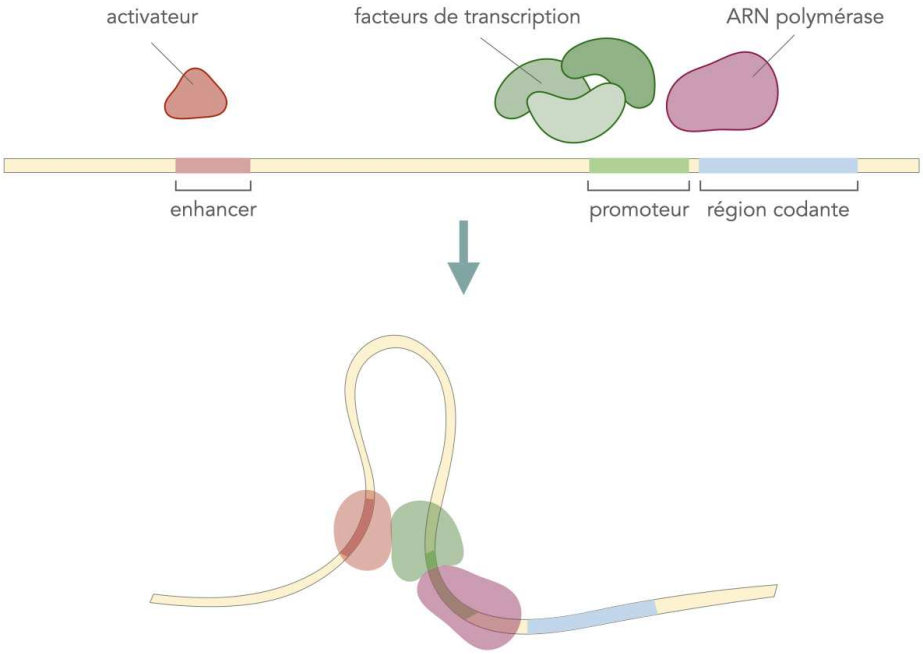

Figure 1.2 – Schéma simplifié du mode d'action des facteurs de transcription. Le facteur de transcription (ici en vert) est une protéine qui vient se lier à l'ADN au niveau d'un TFBS du promoteur, en amont du gène. Plusieurs TFs peuvent se lier aux promoteurs des gènes. Ce processus engendre l'activation de l'ARN polymérase II qui va provoquer la transcription du gène en générant la molécule d'ARN. Des protéines activatrices ou des TFs peuvent aussi se lier à l'ADN au niveau des enhancers qui, *via* la protéine cohésine, obligea l'ADN à se plier (schéma du bas) rapprochant l'enhancer du promoteur et participent ainsi au contrôle de la transcription.

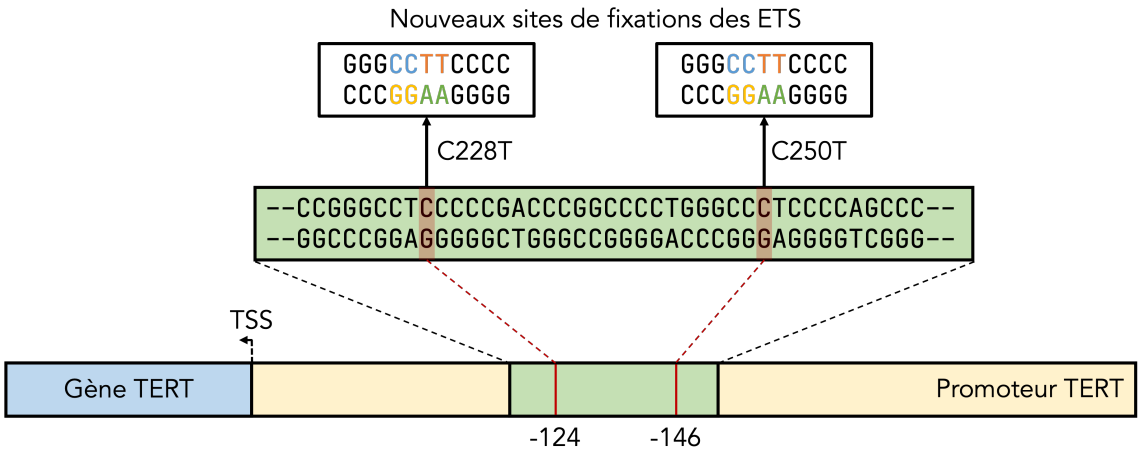

Figure 1.3 – Mutations du promoteur de *TERT*. Illustration des deux mutations C >T identifiées en position -124 et -146 par rapport au TSS du gène. (Cordonnées hg19). Adapté de Colebatch et *al*, 2019 [42].

1.2.2 Les avancées et l'évolution de la bioinformatique

Le rôle essentiel de la bioinformatique dans l'essor de la génomique

La bioinformatique décrit l'utilisation des statistiques et de l'informatique dans leur application aux sciences du vivant. Cette discipline a vu le jour afin d'organiser et analyser les premières données issues d'expériences en biologie. C'est à partir des années 90 que la bioinformatique apparaît comme une discipline de recherche en tant que telle [43]. C'est à cette époque que furent développés des algorithmes efficaces permettant de faire face au volume grandissant de données biologiques. Les applications de la bioinformatique se rapportent essentiellement au traitement automatique et de manière plus générale, à l'intégration, l'analyse, la modélisation ou encore la prédiction de données émanant des sciences du vivant.

Grâce aux avancées techniques permettant de séquencer le génome et le transcriptome d'un organisme [18, 34] les informations génétiques, essentiellement humaines mais aussi d'autres espèces, alimentent des bases de données colossales. Cette quantité d'information a nécessité l'élaboration de méthodes computationnelles abouties afin de l'exploiter de manière optimale. En effet, les technologies NGS [24] produisent des dizaines de millions de fragments de petites tailles qu'il est nécessaire de traiter par des algorithmes bioinformatiques puissants pour étudier ces génomes. Pour séquencer un gé-

nome via DNASeq, des outils d'assemblage des séquences [44] ont ainsi été développés pour reconstruire le puzzle génomique produit par NGS. Puis des algorithmes dédiés aux données RNASeq [45] permettront d'annoter le génome c'est-à-dire de reconstruire les gènes, les localiser sur le génome et les quantifier en fonction d'un tissu, d'un stade de développement donné. Si un génome de référence proche est disponible, ces algorithmes pourront utiliser une étape d'alignement de ces fragments sur ce génome de référence, sinon, des approches d'assemblages *de novo* seront privilégiées.

L'essor des méthodes de séquençage haut-débit associé aux méthodes computationnelles a rendu possible l'identification de l'ensemble des variants d'un individu à l'échelle des régions exoniques (Whole Exome Sequencing) ou bien à l'échelle du génome entier (Whole Genome Sequencing). Les séquences obtenues grâce aux méthodes de séquençage haut-débit sont ensuite comparées avec le génome de référence. Puis, l'utilisation d'un algorithme d'appel de variant, comme GATK (The Genome Analysis ToolKit) [45] va permettre d'en connaître les différences (SNPs ou Indels). Dans ce cadre, l'utilisation de méthodes bioinformatiques permet de détecter des variants rares, inhérents à un individu ou à sa famille.

Certaines anomalies génétiques peuvent en revanche être observées à l'échelle de la population et être responsables de pathologies répandues. Les GWAS (Fig. 1.4) permettent d'établir des associations statistiques fonctionnelles de variations génétiques, généralement les SNP, issues d'un grand nombre d'individus manifestant des traits phénotypiques tels que les maladies humaines majeures. Cette technique consiste dans l'ensemble à comparer les fréquences alléliques des SNP présents dans une population présentant le phénotype d'intérêt avec ceux présents dans une population témoin. Bien que la méthode semble intuitive, elle nécessite des approches bioinformatiques et statistiques spécifiques afin d'intégrer les données et de les analyser pour en déduire une association fonctionnelle. À titre d'exemple, le logiciel libre PLINK [46] permet de réaliser des études d'associations pangénomiques. Un ensemble d'algorithmes et des tests statistiques sont formulés spécifiquement pour proposer des statistiques descriptives, un calcul de déséquilibre de liaison ou encore une stratification de la population étudiée.

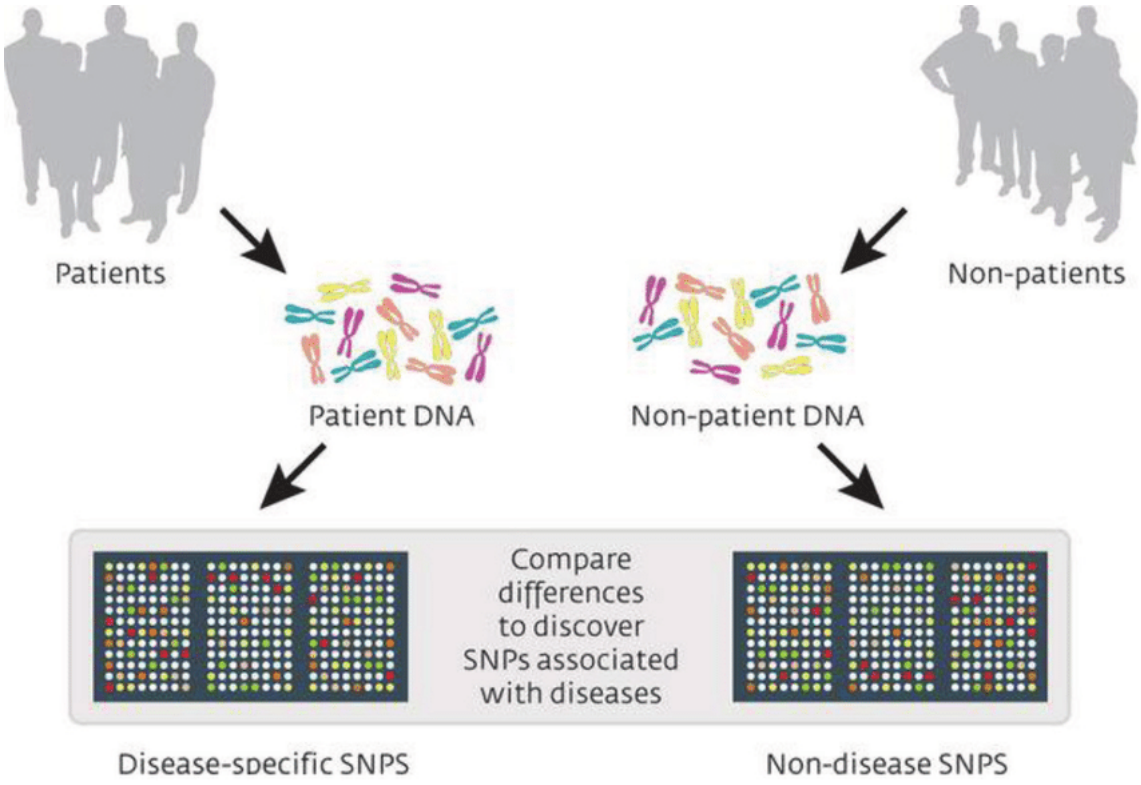

Figure 1.4 – Principe d'une étude d'association pangénomique. L'ADN de personnes présentant un phénotype d'intérêt est comparé à celui d'un groupe de personnes témoin dans le but de détecter le ou les variants génétiques statistiquement associés au phénotype.

Utilisation des algorithmes pour modéliser l'expression des gènes

La nature et la quantité de données produites par les différentes techniques de séquençage occasionnent de nombreux potentiels de valorisation grâce aux méthodes bioinformatiques. Plus particulièrement, les méthodes d'apprentissage automatique (Machine Learning) (ML) présentent un intérêt singulier quant à l'exploitation des données génomiques et transcriptomiques [1]. L'apprentissage automatique représente un ensemble de méthodes et d'approches mathématiques et statistiques dédié à la résolution d'une tâche. Le principe général de la méthode, consiste en la conception d'une fonction permettant de définir le résultat que l'on souhaite faire valoir d'un ensemble de données, appelées observations. L'élaboration de cette fonction nécessite une phase dite d'entraînement ou d'apprentissage avant de pouvoir l'appliquer à des données similaires.

Le domaine de l'apprentissage automatique se divise en deux types de tâches bien distinctes. Selon la disponibilité de l'information que l'on souhaite extraire de nos données, l'apprentissage est dit supervisé ou non supervisé. Lorsque les données dont on dispose ne contiennent pas l'information que l'on souhaite en extraire, nous parlons de données non étiquetées et donc d'apprentissage non supervisé. L'objectif est donc de découvrir, modéliser une structure contenue dans les données. En revanche, lorsque les données d'entraînement sont étiquetées, l'apprentissage est dit supervisé. Il s'agit de classification, lorsque les valeurs à modéliser sont discrètes, ou de régression lorsqu'il s'agit de valeurs continues.

Appliqués à la génomique, les algorithmes d'apprentissage automatique ont démontré leur grande utilité pour l'identification de sites d'épissage ou de promoteurs à partir de séquences nucléotidiques [47, 48]. En effet, bien que de nombreux projets bioinformatiques ont permis de mettre en lumière des relations fortes entre des variations génomiques et certaines maladies ou traits phénotypiques, il demeurait une part d'ombre sur les mécanismes par lesquels ces relations opèrent. Ainsi de nombreux travaux ont été réalisés pour développer des prédicteurs basés sur l'apprentissage automatique pour l'identification *in silico* des signaux génomiques et des régions fonctionnelles telles que des signaux de polyadénylation, des sites d'initiation de la traduction et des sites d'épissage [47, 49, 50].

Pour prédire l'impact fonctionnel de variants non-codants identifiés par séquençage haut-débit, de nombreux outils basés sur des approches d'apprentissage ont été dé-

veloppés ces dernières années. Par exemple, les outils CADD (Combined Annotation Dependant Depletion) [51] et FATHMM-MKL [52] utilisent des approches de classification supervisée via un algorithme de type SVM (Support Vector Machine) entraînées sur plusieurs types de données complémentaires (conservation de séquences, profils bioichimiques issues de consortiums publics ou encore composition des séquences). Plus récemment, l'outil FINSURF [53] propose trois modèles entraînés par forêts aléatoires (RandomForest) sur trois jeux de contrôles négatifs spécifiquement sélectionnés (Adjusted, Local et Random). L'outil constitue une méthode efficace de hiérarchisation automatisée de variants non codants permettant la priorisation de variants candidats et son utilisation est facilement accesible via un serveur en ligne [54].

Les méthodes d'apprentissage profond (Deep Learning) (DL) sont des techniques issues de l'apprentissage automatique capables d'identifier des motifs très complexes dans de grands ensembles de données à partir d'une structure d'algorithme en réseau de neurone [7]. Tout comme les approches traditionnelles d'apprentissage automatique, l'apprentissage profond peut être employé pour réaliser des tâches supervisées ou non. Contrairement aux méthodes d'apprentissage automatique, l'apprentissage profond ne dépend pas du traitement préalable des données à considérer car la méthode permet d'en extraire automatiquement les caractéristiques importantes.

Les exemples d'applications des méthodes de deep learning à la génomique sont déjà nombreux. Des algorithmes permettent par exemple de prédire le statut de méthylation, le contrôle de l'épissage ou encore l'expression des gènes. Parmi les approches existantes, les CNN ont connu un fort succès ces dernières années. Ces méthodes ont permis l'élaboration de l'outil Deep-STORM [55] permettant de transformer des images issues de la microscopie de fluorescence afin de les convertir en haute résolution. Une autre approche [56] en imagerie intègre le CNN à une structure de réseau antagoniste génératif (GAN, Generative Adversarial Network) afin de reconstruire des séquences vidéo de cellules vivantes dynamiques capturées à l'aide d'une technique de microscopie computationnelle. Aussi, à partir d'un réseau de neurones convolutif, ExPecto [57] propose de prédire dans 218 tissus et types cellulaires l'impact de mutations sur la transcription. L'outil AttentiveChrome [58] permet quant à lui de prédire le niveau d'expression des gènes grâce à une architecture de réseau de neurones récurrents modélisant des signaux de modification d'histones.

L'expression des gènes détermine la diversité des types et des états cellulaires d'un

organisme et représente ainsi un phénotype intermédiaire. La modélisation de ce processus biologique offre alors la possibilité de mieux comprendre les mécanismes sous-jacents. Depuis moins de 10 ans, nous avons constaté que ce phénomène pouvait être modélisé comme une fonction de l'ADN et de manière précise grâce aux méthodes de machine learning et plus récemment, celles de deep learning [4, 59]. En effet, un génome représente une base d'apprentissage très large, comme chez l'homme avec 3,2 milliards de nucléotides. Dans ce contexte, les algorithmes de deep learning ouvrent la voie à l'analyse précise des génomes grâce à la puissance de ces méthodes appliquées aux grands ensembles de données.

En dépit de l'intérêt manifeste de ces algorithmes, la recherche en deep learning dans le domaine de la génomique rencontre toutefois des obstacles. En effet, la discipline dépend de la disponibilités des données nécessaires à l'aboutissement des algorithmes et son application à la génomique relève donc des données de séquençage. Le manque de données réelles est évidemment un point épineux pour ce type de recherche, leur production étant encore assez coûteuse et leur utilisation encadrée par les lois de protection des données personnelles.

Face à cette limitation, plusieurs stratégies sont développées dont l'apprentissage par transfert [60]. Cette approche consiste à transférer les connaissances acquises par un algorithme appliqué à une problématique disposant de nombreux exemples à une autre problématique, souvent similaire, aux données moins abondantes. L'approche montre aujourd'hui son intérêt en génomique, notamment pour prédire l'accessibilité de la chromatine à partir d'une séquence [61]. Alors que les volumes de données sont encore faibles, les méthodes d'apprentissage par transfert pourraient bien perdurer pour certaines maladies rares ou encore pour les espèces non-modèles et celles qui bénéficient de moindre financements.

1.3 Le chien comme modèle spontané des cancers de l'homme

1.3.1 L'intérêt du modèle canin en génétique

Le chien, *Canis lupus familiaris*, développe naturellement des cancers cliniquement homologues à ceux de l'homme. En passant de 155 000 cas en 2012 à 382 000 en 2018 [5], l'incidence du cancer est en augmentation et le présente comme une préoccupation capitale en matière de santé publique. La diversité clinique et génétique de cette maladie s'impose comme un défi de taille auprès de la communauté scientifique dans son objectif de mieux en comprendre les causes. Pour contourner cette difficulté, l'étude des cancers survenant spontanément dans un organisme physiologiquement proche de l'homme se présente comme une ressource essentielle. L'apparition spontanée des cancers canins respecte également une aspiration éthique dans le cadre de la recherche en génétique. À titre d'exemple, l'équipe "Génétique du chien" dispose d'un centre de ressources biologiques (CRB) (Cani-DNA) constitué d'environ 30 000 échantillons prélevés sur des chiens, vivant ou ayant vécu au sein d'un foyer.

Depuis 2003 et l'accomplissement de l'assemblage du premier génome humain, les techniques et outils utilisés en biologie moléculaire sont en constant progrès et permettent des avancées médicales considérables. Cette tendance s'observe également chez le chien, avec l'intérêt grandissant pour la génétique de l'espèce [62] et une considération accrue de la santé des animaux de compagnie et donc des soins vétérinaires [63]. En effet, l'année 2003 marque également le premier séquençage du génome d'un chien, un caniche, permettant alors la connaissance du génome canin et la mise en évidence de grandes similitudes entre celui-ci et le génome humain [64].

Composé d'environ 2,4 milliards de nucléotides, le génome du chien compte plus de 20 000 gènes codant pour des protéines répartis sur 38 paires d'autosomes et une paire de chromosomes sexuels [65]. En comptant une divergence nucléotidique de 0,35 substitution par site, le génome canin présente l'avantage d'être plus proche de celui de l'homme que ne l'est celui d'un autre organisme modèle, la souris, qui compte 0,51 substitution par site [66].

Enfin, le chien représente un modèle de choix dans le sens où il peut être le patient dans le cadre d'essais cliniques. L'espèce canine est à considérer comme modèle intermé-

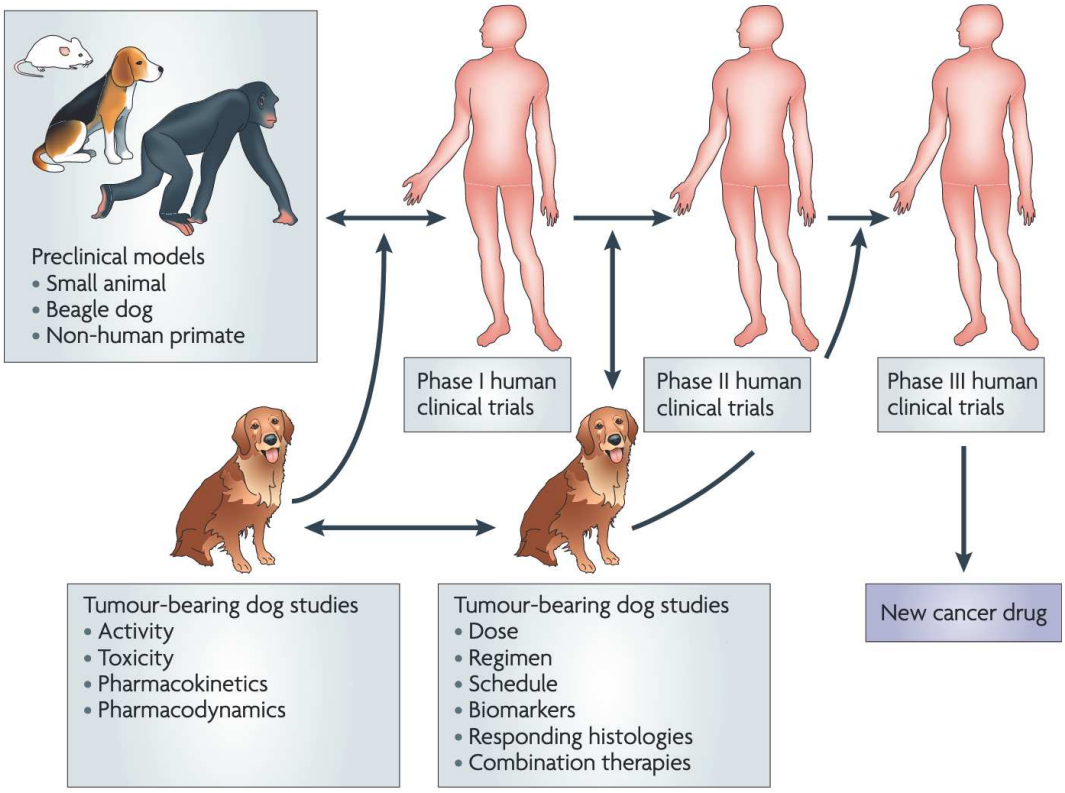

Figure 1.5 – Approche intégrée de l'intérêt du chien domestique atteint de cancer au cours de phases de développement d'un médicament anti-cancéreux. Figure issue de Paoloni et *al*, 2008 [63].

diaire dans le développement des médicaments, entre les modèles précliniques classiques, comme la souris, et les essais sur l'homme [67]. En effet, l'efficacité d'une molécule peut être testée sur une tumeur canine, homologue à l'homme auprès d'un hôte immunocompétent et partageant le même environnement. Aussi, la durée de vie des chiens, plus courte que celle des hommes, permet d'évaluer l'efficacité d'un médicament dans des délais bien plus restreint que chez l'homme (Fig. 1.5).

1.3.2 Les prédispositions du chien à différentes maladies génétiques

Le chien est le premier animal domestiqué. Des études suggèrent que toutes les races de chien connues aujourd'hui sont issues du loup gris d'Eurasie (*Canis lupus lupus*)

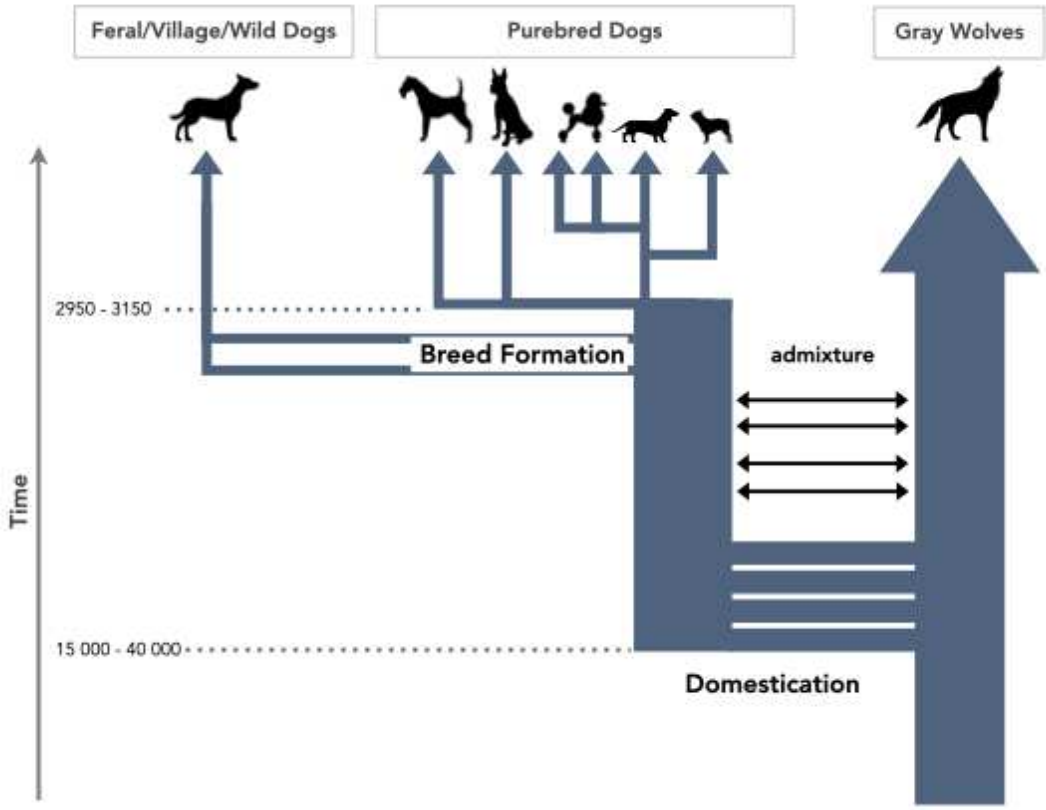

Figure 1.6 – Schéma simplifié de l'histoire évolutive du chien.Le processus de la domestication canine est caractérisé par sa complexité et notamment la présence de deux principaux goulets d'étranglement. Le premier est lié à la domestication du loup par l'homme, le second à la création des races modernes.

ou potentiellement d'Asie du Sud-Est, bien que la localisation géographique exacte et la datation du processus de domestication restent l'objet de nombreuses études [68]. En effet, si l'origine de la domestication du chien remonte à au moins 15 000 ans [68], certaines études [69, 70] proposent une date de divergence entre l'ancêtre du loup et celui du chien dès le paléolithique supérieur ($-30\,000$ ans) (Fig. 1.6), ces individus pourraient toutefois appartenir à une lignée canine éteinte avant l'Holocène 9700 ans av. J.-C [71].

La majorité des 344 races de chien reconnues par la Fédération Cynologique Internationale (FCI) [72] par des caractéristiques physiques et comportementales bien spécifiques, ont moins de 200 ans. Cette variation phénotypique que l'on observe d'une race canine à l'autre est le résultat d'une forte sélection artificielle, menée par des éleveurs

dans le but d'accentuer des caractéristiques distinctives (Fig. 1.7). De nombreuses races sont nées d'un nombre limité d'animaux fondateurs et l'utilisation de géniteurs prisés est une pratique courante. En conséquence, chaque race représente une population d'élevage isolée avec des niveaux élevés d'homogénéité phénotypique, une diversité génétique réduite et un enrichissement en affections spécifiques à la race. Par comparaison, la variation génétique entre les races représente plus de 27% de la variation observée dans toute la population canine, alors que la différenciation génétique reportée pour les populations humaines est de 5 à 10% [73].

Essentiellement, la faible diversité génétique résulte de la consanguinité qui consiste à accoupler deux chiens relativement apparentés pour engendrer une progéniture. Cela peut être deux frères et sœurs, mère et fils, ou père et fille qui s'accouplent. L'objectif principal de cette pratique d'élevage est de développer et d'améliorer la qualité de la lignée de la race. Il est compréhensible que le développement précoce de la race ait nécessité la consanguinité dans une certaine mesure. Il faut mentionner que les éleveurs cherchent à maîtriser la consanguinité dans leurs pratiques d'élevages. Ils sont intéressés à obtenir les meilleurs standards d'élevage pour leurs chiens. Si une race de chien montre une aptitude donnée, un standard atteint, certains éleveurs sont tentés de maintenir ces standards en croisant le chien avec des parents proches qui partagent les mêmes qualités. Les éleveurs vont aussi accoupler des champions de race au sein des lignées afin de garder plus de champions dans la même lignée, ce qui donne un meilleur pedigree pour les portées à venir. Les éleveurs ont longtemps défini les pratiques d'élevage pour maintenir les traits désirables et éliminer les traits non souhaités. Bien que certaines raisons de la consanguinité soient compréhensibles, le processus a de réelles conséquences. Des études récentes ont montré que la consanguinité chez les chiens peut entraîner de nombreux de problèmes de santé, notamment une réduction de 6 % de la taille des adultes, une réduction de la durée de vie de 6 à 10 mois et une réduction globale de la taille de la portée et de la fertilité [75]. Chaque population de chiens de race pure possède une multitude de mutations récessives rares qui étaient soit présentes dans leur population fondatrice, soit apparues dans leur population par la suite. Malheureusement, de nombreuses races sont maintenant hautement consanguines.

Une analyse génétique récente sur 227 races canines [76] a révélé un taux de consanguinité moyen de 25%. C'est plus ou moins la même chose que de partager du matériel génétique similaire avec un frère ou une sœur à part entière. Ces niveaux sont considérés



Figure 1.7 – Diversité de l'espèce canine. Les races canines varient en fonction de nombreux traits phénotypiques comme la taille, la longueur des pattes, le type et la couleur de pelage ou encore la forme du crâne. Sur l'image sont représentés un barzoï (A), un basset hound (B), un chihuahua (C), un schnauzer géant (D), un bichon frisé (E), un colley (F), un bouledogue français (G), un teckel (H), un braque allemand (I), un épagneul nain papillon (J) et un mâtin de Naples (K). D'après Ostrander et *al*, 2012 [74].

comme supérieurs à ce qui est autorisé pour les animaux sauvages et les populations humaines. Par exemple, chez les humains, des niveaux de consanguinité de 3 à 6 % sont souvent associés à des taux accrus de problèmes de santé complexes [77]. L'étude a également mis en évidence les races de chiens les plus consanguines, le lundehund norvégien, le carlin, le bulldog anglais, le basset, et le golden retriever. Les deux causes principales de cette consanguinité sont le désir et le maintien de traits spécifiques et une petite taille de l'effectif des populations originales. À l'inverse, parmi les races canines les moins touchées par la consanguinité, on retrouve des races dites primitives telles que l'akita, le basenji, le samoyed, le malamute, ou des races locales comme chien de ferme dano-suédois et des races issues de croisement récents comme le barbet ou le labraniche (croisement du Labrador et du caniche). Ces exceptions sont probablement liées à une population fondatrice relativement importante et aussi au fait que les individus ont été sélectionnés pour une fonction et non sur des critères esthétiques. Les principales conclusions de cette étude avancent que les races avec des niveaux de consanguinité plus élevés nécessitent plus de soins vétérinaires, qu'une consanguinité élevée est le résultat de livres généalogiques fermés ou d'un petit nombre de fondateurs ou des deux. Il s'agit d'exercer une gestion prudente des populations reproductrices pour éviter une perte supplémentaire de la diversité génétique existante, grâce à l'éducation des sélectionneurs et au suivi des niveaux de consanguinité permis par les technologies de génotypage direct. En particulier, dans les quelques races à faible niveau de consanguinité, tout doit être fait pour maintenir la diversité génétique présente.

D'un point de vue de généticien, la faible variabilité génétique entre chiens d'une même race facilite grandement les études génétiques basées sur l'association entre génotype et phénotype [67]. En effet, le choix d'utiliser peu de reproducteurs pour créer et multiplier une race va favoriser la propagation d'anomalies génétiques identiques responsables de pathologies particulières. Ainsi les études GWAS chez le modèle canin simplifient la découverte de loci et de gènes responsables de maladies. Les races canines, sous pression de sélections artificielles drastiques, connaissent malheureusement de nombreuses prédispositions aux maladies génétiques. Ces pathologies souvent courantes chez le chien peuvent être rares chez l'homme. En effet, auprès de la population humaine, une maladie est considérée comme très fréquente dès lors qu'elle touche 0,2% de la population. Pour l'espèce canine, certaines maladies touchent entre 1% et 10% des individus d'une même race [78], atteignant même 30% dans le cas de l'ichtyose chez le golden retriever [79].

Ces éléments illustrent l'intérêt du chien comme modèle génétique spontané efficace pour l'étude de maladies fréquentes dans certaines races et rares chez l'homme où il est difficile d'en identifier les causes génétiques. Le modèle canin permet alors de pouvoir répondre à un panel agrandi de questions scientifiques et cela de manière plus précise, au bénéfice de la santé humaine mais également canine. Le chien dispose donc d'une histoire génétique unique permettant une identification simplifiée des relations phénotype / génotype par rapport aux études menées chez l'homme. De plus, la stratification en race de l'espèce canine forme des isolats génétiques. En utilisant les récentes techniques de biologie moléculaire adaptées au chien en génomique et transcriptomique, le modèle canin permet non seulement de pointer des altérations génétiques impliquées dans des maladies à transmission mendélienne, mais aussi dans des maladies plus complexes et multifactorielles comme les cancers.

1.3.3 Le projet Dog10K

En 2019, la communauté internationale de la génomique canine a lancé le projet Dog10K, un effort international pour séquencer et analyser les génomes de plusieurs milliers de canidés sur une période de 5 ans [11]. Notre équipe fait partie de ce consortium à la fois par notre contribution aux échantillons canins séquencés (fournis par le CRB - caniDNA), par notre investissement dans la réalisation d'une nouvelle séquence de référence [80] et enfin par l'analyse des variants identifiés et de leurs caractéristiques structurelles et fonctionnelles. Les principaux objectifs de l'analyse des données ont été centrés sur la caractérisation de la variabilité génétique de l'espèce canine, à partir de la plus grande représentativité possible de populations canines. En effet, même avec la publication de nouvelles séquences de lectures longues de haute qualité, y compris le boxer original [80], un basenji [81], un grand danois [82], un berger allemand [83], et d'autres en cours, l'absence de données de séquences de races de chiens moins courantes, de races qui représentent des populations non européennes, de chiens de village et sauvages et de canidés sauvages limite presque toutes les études génétiques canines en cours. Les autres objectifs du consortium Dog10K sont une meilleure caractérisation de la formation et la morphologie des races, la sensibilité et la progression des maladies et la génétique de la domestication des chiens.

L'échantillonnage du projet Dog10K a permis d'inclure trois à cinq membres de chacune de plus de 300 races canines avec une distribution qui comprenait des races rares

et communes, une distribution géographique d'ordre mondial et une diversité en termes d'esthétique et de comportement. Un recrutement important de chiens de village (village dogs) a été également réalisé. Ils représentent une population de chiens qui ne sont pas sauvages mais qui ne relèvent pas de la race pure ou de la race mixte. Nous avons également recherché environ 20 populations de chiens qui résident dans une zone géographique définie qui partagent des traits comportementaux et/ou morphologiques [84] [85], qui sont souvent liés à une occupation hautement spécialisée comme l'élevage ou le gardiennage. La collection d'au moins 30 pedigrees étendus de chiens va permettre d'étudier les taux de recombinaison et de mutation à petite échelle dans le génome de populations distinctes de canidés était particulièrement intéressante. Enfin, l'échantillonnage a été étendu à 20 populations de canidés sauvages, notamment les loups, coyotes, chacals dorés, dohls et lycaons.

Aujourd'hui, le consortium Dog10K a réalisé les données de 1929 séquençages complets de génomes (Whole Genome Sequencing) (WGS). Le consortium Dog10K a utilisé l'assemblage du génome version canFam4 pour l'appel de variants. L'identification des variants de l'ensemble des données Dog10K contient près de 30 millions de variants bialléliques et inférieurs à 50 bp de haute qualité, représentant un catalogue de variants à la fois complet et diversifié pour les canidés. Ainsi, cet échantillonnage comprenant également 281 chiens de village et de 57 loups a révélé une multitude de variations inédites non capturées par les chiens de race, avec près de 30% de toutes les variations observées étant exclusives à ces deux groupes. Enfin, l'analyse Dog10K a produit le catalogue le plus complet de variations structurelles ($>$50 bp) chez les chiens, composé de plus de 144 000 variants de structure, élargissant le répertoire global de variation entre les races de chiens. L'ensemble des variants comprend 27 878 361 des SNV autosomiques et 847 128 du chromosome X a été identifiés après l'application de filtres de profondeur et de qualité de séquences, ce qui représente un variants toutes les 80 pb lorsque les 1 929 individus sont pris en compte. La qualité des variants a été validée par un taux de concordance de 99,8% calculé avec les génotypes de 168 individus communs, également typés sur l'Illumina Canine HD Array. Une comparaison avec la variation entre des chiens et des loups contenus dans les données canines existantes telles que DBVDC [86], du NIH [87] et European Variation Archive (EVA) RS Release 3, a montré que 43 % sont uniques à cet ensemble. En revanche, 98% de ces variants uniques sont rares (AF$<$1%), et non dues à des différences entre la version d'assemblage CanFam3.1 [88] et CanFam4 [83].

1.3.4 L'oncologie comparée entre l'homme et le chien

L'espérance de vie des chiens de compagnie s'est allongée de manière considérable dans les pays les plus riches. Ce fait s'explique entre autres par une médecine vétérinaire plus performante et le développement des assurances réservées aux animaux de compagnie. Cependant, la principale raison se cache dans le fort investissement moral et financier des propriétaires qui, souvent aujourd'hui, considèrent leur animal comme un membre à part entière de leur famille [63]. Ainsi, les chiens voient fréquemment leur durée de vie dépasser les 10 ans. Malheureusement, cet âge avancé s'accompagne d'une survenue de plus en plus fréquente des cancers, notamment chez les chiens de pure race [89]. De plus, tout comme chez l'homme, les cancers de causes génétiques et environnementales touchent également les chiens plus jeunes [90].

La durée de vie moyenne d'un chien étant de 12 ans et celle d'un homme 79 ans, le vieillissement accéléré du chien induit une évolution tumorale plus rapide et plus agressive. Ce processus s'explique également par le fait que les cancers canins sont détectés plus tardivement que chez l'homme, permettant en revanche de recueillir des données cliniques précieuses et dans un temps réduit par rapport aux essais menés chez l'homme. L'étude du cancer s'est fortement développée depuis les années 2000, avec notamment la notion de « cancer hallmarks », c'est-à-dire des caractéristiques spécifiques du cancer ou Hallmark [91]. Ces marques de cancer permettent de comprendre la pathogénicité des tumeurs et sont publiées en 2011 de nouvelles caractéristiques des tumeurs avec les cibles thérapeutiques potentielles associées [6] (Fig. 1.8).

Les facteurs liés au processus de tumorigenèse impliquent directement l'hérédité et par conséquent sous-tendent des prédispositions génétiques. Il existe de nombreux facteurs environnementaux allant des modes de vie, consommation d'alcool, de tabac, alimentation, etc., à l'exposition à la pollution, au soleil par exemple. Les cellules cancéreuses, heureusement rares, représentent finalement une exception où un type cellulaire a acquis un avantage de survie grâce à des mutations particulières et ainsi échappera à la surveillance de l'organisme.

Parmi les options thérapeutiques en oncologie, la thérapie dite ciblée est, à la différence de la chimiothérapie ou de la radiothérapie, un traitement qui va cibler les cellules cancéreuses et sera capable de détruire spécifiquement les cellules tumorales. Cette approche, largement encouragée, est désormais une option possible grâce aux avancées de

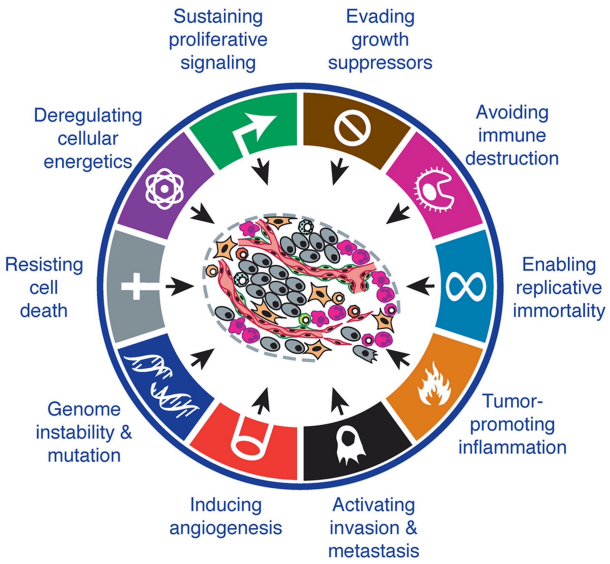

Figure 1.8 – Principales caractéristiques biologiques du cancer. D'après Hanahan et *al*, 2011 [6].

la recherche et de la technologie des NGS, qui permettent de détecter les anomalies génétiques de type somatiques et proposent une caractérisation plus ou moins précise des types et sous-types de cancer. Ces approches ciblées reposent en effet sur l'analyse et la description précise des profils génétiques des tumeurs des patients et donc sur la qualité des concepts, des méthodes et des interprétations issues des analyses bioinformatiques [5]. En effet, en identifiant avec une grande sensibilité et spécificité les mutations somatiques à partir d'une biopsie, à partir de l'ADN circulant, on est en mesure de décider si une thérapie ciblée est possible. C'est la spécificité du profil génétique de la tumeur qui permettra de choisir l'option thérapeutique la plus appropriée, au niveau de la tumeur ou des cellules environnantes par exemple. Ainsi, certaines thérapies ciblées utilisent des anticorps monoclonaux capables de se lier à des protéines présentes uniquement à la surface des cellules cancéreuses. D'autres interviennent dans l'environnement de la tumeur en limitant l'angiogenèse, pour limiter l'apport de nutriments et d'oxygène à la tumeur et la propagation des métastases et d'autres pénètrent directement dans la cellule pour atteindre leur cible [92]. Certaines thérapies ciblées sont capables de bloquer les échanges entre cellules tumorales, notamment les messages responsables de leur prolifération anarchique. Leurs cibles incluent les facteurs de croissance, les récepteurs membranaires. Les thérapies ciblées font partie de l'arsenal de la médecine de précision, aussi appelée médecine personnalisée.

N'importe quelle cellule peut être à l'origine de l'apparition d'un cancer. L'organisme humain compte environ 200 types cellulaires différents et donc autant de formes de cancers. Ces dernières peuvent être classées en cinq grands types :

— Les carcinomes surviennent dans les épithéliums de la peau, du système nerveux, du tube digestif ou encore du système respiratoire.
— Les sarcomes apparaissent dans les tissus conjonctifs, notamment dans les muscles et les os.
— Les leucémies sont des cancers du sang.
— Les lymphomes se développent dans les glandes qui luttent contre les infections, comme les ganglions lymphatiques.
— Les myélomes prennent naissance dans la moelle osseuse.

Les cancers canins apparaissent spontanément et sont cliniquement, histologiquement et dans leur réponse aux traitements, très semblables aux cancers humains. D'autre part, des études récentes ont également montré une similarité génétique et moléculaire entre les cancers humain et canins [93, 94].

Enfin, le séquençage des génomes humains et canins a permis de mettre en lumière la présence de régions génomiques disposant de motifs nucléotidiques conservés [12]. Ce mécanisme de conservation appuie l'hypothèse selon laquelle ces régions sont sous contrainte et que leur rôle fonctionnel est important. Pour le génome codant, les séquences conservées indiquent la préservation de la séquence protéique tandis que pour le génome non-codant, la conservation d'une séquence indique le potentiel régulateur de celle-ci.

Plusieurs métriques permettent d'évaluer le niveau de conservation entre espèces à la résolution du nucléotide. Ainsi le score PhyloP mesure la conservation évolutive sur les sites d'alignement individuels [95]. Les interprétations des scores sont comparées à l'évolution attendue en dérive neutre. Les scores positifs mesurent la conservation, qui évolue plus lentement que prévu, sur les sites qui devraient être conservés. Les scores négatifs mesurent l'accélération, qui est une évolution plus rapide que prévue, sur les sites qui devraient évoluer rapidement. Les scores PhyloP sont extrêmement utiles pour évaluer les signatures de sélection au niveau de nucléotides particuliers ou de classes de nucléotides. Les valeurs absolues des scores phyloP représentent des valeurs de probabilité sous une hypothèse nulle d'évolution neutre. Les scores PhyloP seront prochainement calculés et disponibles pour la dernière version de l'assemblage canin CanFam4 [83], qui

contribuera à valoriser le modèle canin en oncologie comparée et confortera le chien comme un puissant modèle spontané d'étude des cancers humains.

Dans les études des caractéristiques conservées des régions régulatrices des gènes orthologues, il a été observé que les composants fonctionnels essentiels du promoteur tels que le site d'initiation de la transcription, les boîtes TATA et les motifs régulateurs, sont plus conservés que les séquences qui les entourent (70 à 100 % par rapport à 30-50%) [96]. Ainsi l'architecture gène-spécifique des séquences promotrices rend difficile la prédiction des promoteurs en général et des promoteurs de gènes orthologues. Le promoteur contient le site d'initiation de la transcription (Transcription Start Site) (TSS) et s'étend généralement de -60 à +40 par rapport au TSS. Environ 30 à 50 % de tous les promoteurs connus contiennent une boîte TATA située à environ 30 bp en amont du TSS et constitue un signal fonctionnel conservé dans les promoteurs eucaryotes. La région de 200-300 pb immédiatement en amont du promoteur central constitue le promoteur proximal qui contient généralement plusieurs sites de liaison aux facteurs de transcription (TFBS), généralement courts (6 à 10 pb), et qui sont responsables de la régulation de la transcription. La partie distale du promoteur est située plus en amont et peut également inclure des sites de liaison aux facteurs de transcription.

1.4 Objectifs des travaux de thèse

Les progrès réalisés dans les disciplines scientifiques au cours des dernières décennies ont permis d'initier la compréhension des processus biologiques régissant la génération, la croissance et l'extinction des organismes vivants. Désormais les défis sont d'exploiter et de valoriser les différentes informations collectées, notamment dans les disciplines de génomique et transcriptomique. Parmi les principaux enjeux, le premier est de prendre en compte la diversité biologique individuelle et le second doit viser à raccourcir les délais entre la découverte d'une connaissance telle qu'une mutation, une molécule et son utilisation. Afin de répondre aux problématiques soulevées par les enjeux en santé publique, les méthodes bioinformatiques et la véritable révolution conceptuelle et technique que représente aujourd'hui l'apprentissage profond, se présentent comme une proposition prometteuse.

L'application de ces méthodes dans le domaine de la recherche sur le cancer peut grandement contribuer à l'appréhension des mécanismes suscitant l'apparition de cette

maladie. Il est également instructif de combiner ces ressources à la possibilité d'étudier des phénomènes biologiques similaires chez des espèces modèles. En effet, l'espèce canine représente un modèle d'étude performant en aidant l'homme à découvrir de nouveaux gènes impliqués dans le développement des cancers. L'utilisation du chien en tant que modèle génétique permet aussi de déchiffrer les gènes déjà connus chez l'homme et de les cibler par différents essais thérapeutiques.

Dans ce contexte, mon projet de thèse a consisté en l'utilisation de cancers canins naturels comme modèles pour les cancers humains, ce qui confère une démarche originale et des intérêts spécifiques dans mes travaux. En effet, les cancers chez les chiens se développent naturellement, tout comme chez les humains, et il est de plus en plus reconnu que les souris ne sont pas les modèles les plus adaptés au cancer humain. La plupart des efforts d'oncologie comparative se sont concentrés sur les chiens, car ils développent spontanément bon nombre des mêmes types de cancers que les humains, notamment le mélanome, le cancer du sein et le lymphome. Ainsi, les objectifs et enjeux de nos travaux sont d'analyser les cancers canins avec des angles innovants que ce soit avec l'apport des méthodes et des concepts modernes en informatique, et en génétique avec la considération des variants non codant.

Dans un premier temps, mon projet de thèse a consisté en la modélisation de l'expression des gènes canins par le moyen d'un algorithme d'apprentissage profond. Il s'est tout d'abord agi d'évaluer l'intérêt de la constitution de ce modèle de prédiction étant donné les solutions envisageables pour exploiter des ressources pré-existantes. Puis, nous avons utilisé le modèle de prédiction obtenu pour investiguer une question biologique essentielle qui cherche à établir l'interprétation fonctionnelle et la hiérarchisation des variants non codants. Ainsi nous avons réalisé une analyse de prédiction sur l'impact de mutations régulatrices situées dans les régions promotrices sur le niveau d'expression des gènes impliqués en cancérologie.

MATÉRIEL ET MÉTHODES

2.1 Description des données

Nous avons réalisé deux modèles de prédiction du niveau d'expression des gènes canins et utilisé un modèle de prédiction du niveau d'expression des gènes humains, établi par Kelley et *al*. Ces modèles ont été réalisés à partir de génomes de référence et de profils d'expression séquencés via la technologie CAGE [9].

Le premier modèle canin est réalisé à partir de la version d'assemblage du génome canin CanFam3 [22], le second modèle canin à partir de la récente version CanFam4 [83] et enfin le modèle de prédiction de l'expression des gènes humains se base sur la version d'assemblage du génome humain GRCh38 [97].

2.1.1 Assemblages de référence

L'utilisation de l'assemblage du génome de l'homme et celui du chien est nécessaire comme support direct à l'établissement des modèles de prédiction du niveau d'expression des gènes. Ces données sont également exploitées pour l'alignement des profils d'expression CAGESeq.

canFam3

Le génome canin est composé d'environ 2,5 milliards de paires de bases. Le génome de référence correspondant à la version d'assemblage canFam3 est celui d'une femelle boxer (Tasha), séquencé en 2005 par une approche de WGS utilisant la méthode de séquençage traditionnelle de Sanger [17] et offrant une profondeur de séquençage de 7,4x. Le choix de cet individu et de cette race reposait sur le faible taux de d'hétérozygotie du génome du boxer par rapport à d'autres races. La version canFam3 compte 23 876 régions gaps (de taille moyenne 764bp) dont 19,6% se situant dans le corps de gènes et 9,8% dans

une région de 5kb en amont des gènes. Ces gaps de séquences, en partie due à la perte du gène PRDM9 chez le chien [98] et donc à la présence de régions riches en GC par le phénomène de conversion génique biaisé (gBGC), ont induit une absence d'information concernant certaines régions du génome canFam3 dont les promoteurs.

canFam4

Depuis 2020, une nouvelle version d'assemblage du génome canin est disponible. Celle-çi correspond au séquençage du génome d'une femelle berger allemand (Mischka), dont la génétique est considérée comme représentative de la race. En effet, cet individu a été génotypé avec une puce SNP haute-densité (170K - CanineHD BeadChip Illumina) et comparée à une population de 260 bergers allemands d'une étude précédente. Mischka a été évaluée comme étant représentative de la population via la valeur de consanguinité attendue (F=0,037) et des mesures de distance génétique à échelle multidimensionnelle (MDS - PLINK v1.9) [46] et donc sélectionnée pour l'assemblage de son génome. Le génome de cette chienne a été séquencé par une combinaison de technologies de nouvelle génération permettant d'atteindre une profondeur de lecture de 100x. Ainsi l'assemblage a utilisé plusieurs technologies de séquençage. Des librairies de séquences de lectures longues ont été préparées (SMRTbell Template Prep Kit 1.0) et 70 cellules SMRT ont été séquencées sur le système PacBio Sequel avec une chimie v2.1 (Pacific Biosciences) qui ont généré 276,86 Gb de données. Les linked-read ont été produites à partir de librairies de la technologie Chromium (10x Genomics) et séquencées sur un Illumina HiSeq X (2 x 150 bp) générant 269,75 Gb de données. La technologie de Dovetail Genomics a été utilisée pour préparer trois librairies HiC. Les librairies résultantes partagent de nombreuses caractéristiques des données Hi-C qui sont utiles pour l'assemblage et l'orientation des fragments génomiques à de longues distances. Ces librairies ont été séquencées sur un Illumina HiSeq X (lectures appariées de 2 x 150 bp) et ont produit 121,47 Gb de données). Cette version, plus récente, du génome de référence canin permet d'obtenir un assemblage plus complet, ne comptant désormais que 367 gaps (dont la taille moyenne est de 100 bp).

GRCh38

La version d'assemblage du génome humain GRCh38 a été publiée en 2013 par le consortium de référence sur le génome humain (GRC, Genome Reference Consortium). Cette version d'assemblage du génome est composite, c'est-à-dire que 70% de son contenu provient du génome d'un donneur homme anonyme et 23% provient de 10 librairies d'individus distincts. Enfin, 7% du contenu provient de plus de 50 librairies représentant des donneurs masculins et féminins anonymes et des chromosomes de différentes lignées cellulaires [97]. La séquence du génome de référence GRCh38, ou hg38, compte 349 gaps d'une longueur totale de ~160 Mb.

2.1.2 Profils d'expression CAGE-seq

Les promoteurs de gènes, généralement situés en amont de TSS, sont les régions régulatrices de l'expression des gènes. Les régions promotrices des gènes permettent la fixation et la libération de facteurs de transcriptions spécifiques qui vont moduler l'expression des gènes de manière coordonnée et tissu-spécifique. La technologie CAGE [9] permet de capturer la coiffe, c'est-à-dire l'extrémité en 5' des ARNs de l'échantillon biologique. Ces fragments sont extraits, rétrotranscrits en ADNc, amplifiés par PCR et séquencés. Ces courtes séquences produisent un instantané de l'extrémité 5' du transcriptome et permettent d'identifier les TSS et quantifier les transcrits correspondants dans un échantillon biologique donné. La technique CAGE a été publiée pour la première fois par Hayashizaki, Carninci et ses collègues en 2003. La méthodologie CAGE est largement utilisée dans les projets de recherche du consortium FANTOM, un consortium de recherche international créé en 2000 dans le cadre de l'institut de recherche RIKEN au Japon, pour aider à l'annotation fonctionnelle des génomes.

Le projet d'annotation du génome canin : le consortium DoGA

DoGA (Dog Genome Annotation project) [99] est un projet collaboratif pour générer une annotation fonctionnelle du génome du chien afin d'améliorer les modèles canins pour la santé humaine. Le projet est dirigé par les Pr. Hannes Lohi, (University of Helsinki, Helsinki, Finland) ; Associate Professor Carsten Daub (Karolinska Institutet, Stockholm, Sweden) et le Pr. Juha Kere, (University of Helsinki and Karolinska Institutet). DoGA a

pour but de générer la source d'informations fonctionnelles la plus complète du génome du chien afin d'en faciliter l'analyses des traits, de la morphologie, du comportement, et des maladies génétiques et pour définir des scénarios de traitement plus efficaces et plus précis, ce qui n'est pas possible avec les données de référence actuelles.

De manière générale, la compréhension d'un génome englobe l'identification des gènes codant pour les protéines et des ARN régulateurs de l'expression génique. Il est tout aussi important de comprendre comment la production de protéines est orchestrée dans chaque type de cellule ainsi que sa régulation. Quelques soient les génomes étudiés, la clé de cette compréhension est l'identification et la description, également appelées annotations, des régions régulatrices des gènes dans le génome ainsi que des gènes qui ne produisent finalement pas de protéines mais des molécules d'ARN régulatrices.

Chez le chien, la variabilité des maladies, de la morphologie et du comportement chez les races canines se produit plus souvent dans les régions régulatrices que dans les gènes codant pour les protéines. Les régions promotrices de gènes présentent un intérêt particulier. En effet, ces régions régulatrices, qui ont suscité beaucoup d'intérêt au cours des deux dernières années, ne peuvent être détectées qu'avec les technologies récentes et sont considérées comme essentielles pour donner aux cellules leur identité spécifique. La complexité des organismes supérieurs est codée dans les régions régulatrices des génomes et dans la complexité de l'ARN régulateur, qui évolue avec la complexité de l'organisme.

Chez l'Homme, les projets d'annotation fonctionnelle incluent les projets ENCODE, FANTOM et GTEx [8, 10, 35]. Les génomes des organismes modèles nécessitent une annotation améliorée pour les mêmes raisons. La séquence du génome de référence du chien est disponible depuis 2005 [22] et a grandement facilité la découverte de gènes associés à des centaines de traits normaux ou pathologiques [100]. Cependant, l'annotation du génome canin de nombreux gènes et des ARNs régulateurs non codant pour les protéines fait encore défaut et reste encore incomplète malgré les efforts continus d'annotation fonctionnelle auxquels notre équipe contribue largement [50, 101].

Les objectifs spécifiques de DoGA impliquent de générer l'atlas d'annotation génomique le plus complet pour le chien y compris le loup. Un second objectif est de décrire les points communs et les différences entre les gènes humains et canins, y compris les régions régulatrices des gènes qui sont essentielles pour utiliser le chien comme modèle pour les maladies humaines. Un autre objectif est de fournir cet atlas d'annotation du génome canin à la communauté scientifique, y compris un système de base de données

élégant, similaire à la base de données FANTOM. Un dernier grand objectif est d'utiliser ces nouvelles ressources génomiques pour identifier de nouveaux gènes à risque et leurs variants dans les troubles cérébraux canins tels que l'épilepsie, l'anxiété et la neurodégénérescence en tant que modèles pour les maladies à composantes génétiques humaines correspondantes.

DoGA a généré une biobanque de tissus frais à partir de chiens et de loups (~120 échantillons) chez 10 chiens (de 9 races) et 2 loups. Le projet a généré des approches transcriptomiques telles que STRT [102] et CAGE pour générer des données massives pour l'annotation. La technologie de séquençage STRT est similaire à la technologie CAGE dans le sens où elle cible aussi les parties 5' des transcrits mais elle nécessite moins de quantité d'ARN en entrée (~40-50ng) par rapport au CAGE et elle permet aussi de séquencer plus en aval du TSS du transcrit engendrant des profils de distribution de lectures autour de TSS légèrement décalés vers le 3' du transcrit [99]. Nous avons établi une collaboration, formalisée par un accord de Memorandum Of Understanding, pour avoir accès aux échantillons de séquençage, nécessaires comme données d'apprentissage pour nos travaux.

Données alignées sur canFam3

Pour le modèle de prédiction du niveau d'expression des gènes canins basé sur la version d'assemblage CanFam3, nous avons utilisé 9 profils d'expression CAGE provenant de la base de données publique proposée par le consortium FANTOM [35]. Ces 9 profils d'expression reflètent 3 tissus canins, prélevés sur un beagle, chacun représenté par 3 réplicats biologiques. Aussi, dans le cadre d'une collaboration avec le consortium international DoGA, nous avons pu intégrer au modèle 125 profils d'expression canins supplémentaires. Ceux-ci décrivent le niveau d'expression des gènes dans 42 tissus, représentés par 1 à 5 réplicats biologiques (Tableau 2.1). Ces séquençages indiquent un nombre moyen de lectures de 6 552 473 (Fig. 2.1). Après séquençage, les lectures de ces échantillons canins ont été alignées sur la version d'assemblage du génome canin canFam3 (Sept 2011, Broad canFam3.1/canFam3).

Tableau 2.1 – Tissus séquencés et alignés sur canFam3

Nom de tissu	Nombre de réplicat
Adénohypophyse	3
Glande surrénale	3
Artère + veine cave	3
Moelle osseuse	6
Muscle cardiaque	1
Noyau caudé + amygdale	1
Hémisphère cérébral	3
Hémisphère cérébral + vermis du cervelet	2
Colon	3
Corps calleux	3
Embryon	6
Cortex frontal + cortex pariétal + cortex temporal	1
Oreillette du coeur	3
Endocarde + valve cardiaque	3
Formation de l'hippocampe	3
Rein	3
Foie	6
Poumon	3
Noeud lymphatique	3
Mésencéphale + tronc cérébral	3
Neurohypophyse	3
Bulbe olfactif	3
Bulbe olfactif + organe vomeronasal + cavité nasale	1
Ovaire	3
Pancreas	1
Glande parathyroïde	3
Lobe piriforme	4
Glande prostatique	3
Rétine	1
Rétine + nerf optique	2
Muscle squelettique	3
Peau	6
Moelle épinière	3
Ganglion rachidien	3
Rate	3
Estomac	3
Testicule	3
Thalamus	3
Glande thyroïdienne	3
Trachée	3
Trachée + bronches	3
Utérus	3

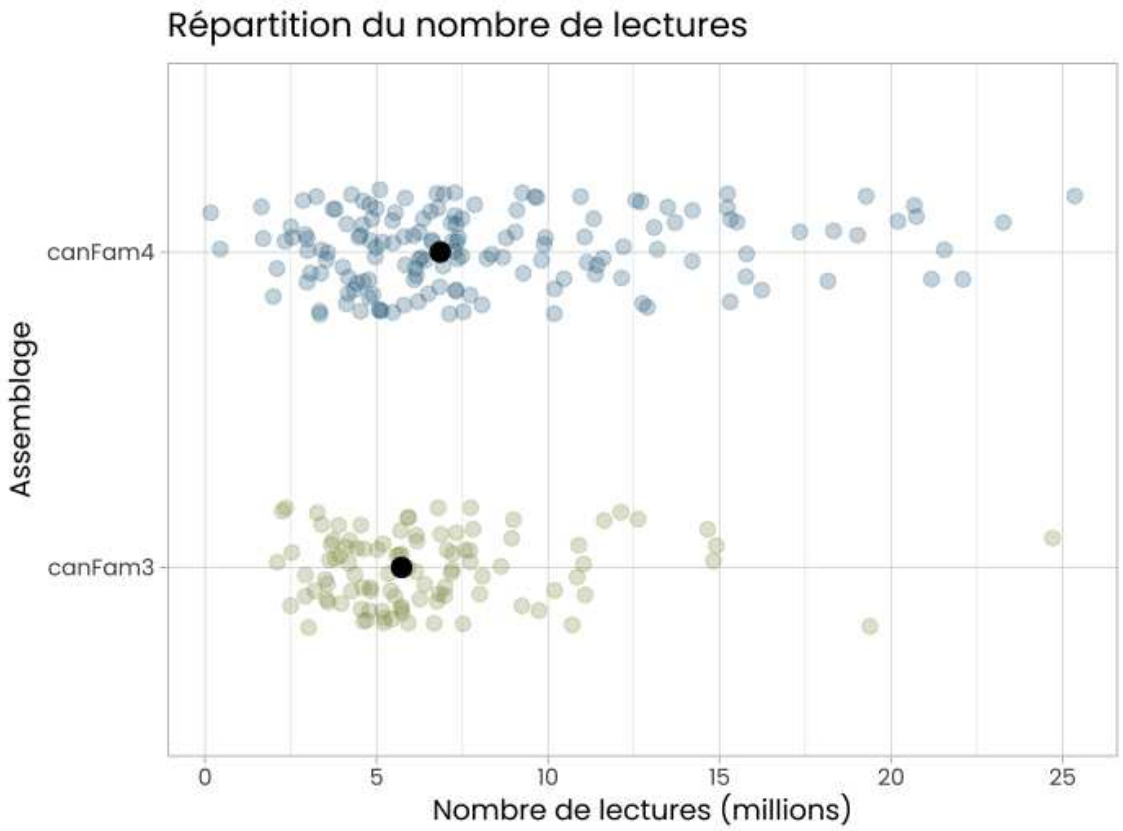

Figure 2.1 – Répartition du nombre de lectures selon la version d'assemblage. Les points bleus représentent les fichiers CAGE des lectures alignées sur l'assemblage canFam4, les points verts représentent ceux alignés sur la version canFam3. Les points noirs plus épais représentent les médianes respectives.

Données alignées sur canFam4

Le deuxième modèle prédictif du niveau d'expression des gènes canins est construit avec la version d'assemblage du génome canin canFam4 (Mar 2020, Uppsala University UU_Cfam_GSD_1.0). Pour cette approche, nous avons pu utiliser 116 séquençages d'échantillons canins alignés sur la version canFam4 du génome canin par nos collaborateurs du DoGA. Ces 116 échantillons représentent 36 tissus comptant 3 à 6 réplicats biologiques chacuns (Tableau 2.2). Nous comptons un nombre de lectures moyen de 8 322 759 (Fig. 2.1) pour ces échantillons et un nombre de pics s'élevant à 54 295. Le consortium FANTOM ne propose pas de données CAGE canines alignées sur la version canFam4 du génome.

Tableau 2.2 – Tissus séquencés et alignés sur canFam4

Nom de tissu	Nombre de réplicat
Adénohypophyse	3
Glande surrénale	3
Artère	3
Moelle osseuse	3
Muscle cardiaque	3
Noyau caudé	3
Hémisphère cérébral	3
Colon	3
Corps calleux	3
Embryon	6
Cortex frontal	3
Endocarde	3
Formation de l'hippocampe	3
Rein	3
Foie	3
Poumon	3
Mésencéphale	3
Neurohypophyse	3
Bulbe olfactif	4
Ovaire	3
Pancreas	3
Glande parathyroïde	3
Lobe piriforme	5
Glande prostatique	3
Rétine	2
Muscle squelettique	3
Peau	6
Moelle épinière	3
Ganglion rachidien	3
Rate	3
Estomac	3
Testicule	3
Thalamus	3
Glande thyroïdienne	3
Trachée	3
Utérus	3

Données alignées sur GRCh38

Le modèle de prédiction du niveau d'expression des gènes humains établi par Kelley et *al.* [103] utilise un total de 5313 échantillons humains séquencés par plusieurs technologies. Nous comptons 3991 échantillons séquencés par la technologie ChIP-seq [104] permettant de cartographier les sites de liaison sur l'ADN d'une protéine d'intérêt, 10 échantillons ATAC-seq [105], caractérisant les régions accessibles de la chromatine, 674 échantillons DNase-seq [106] identifiant et localisant les régions régulatrices. Enfin, le modèle de prédiction chez l'homme compte 638 échantillons séquencés par la technologie CAGE [107], provenant également de la base de données publiques produite par le consortium FANTOM.

Les différents usages que nous faisons du modèle humain portent sur les profils d'expression CAGE le composant. Ces derniers représentent 595 tissus ou cellules uniques.

2.1.3 Panel de gènes impliqués en cancérologie

Avec pour ambition d'évaluer la performance et l'exploitabilité des modèles de prédiction dans la perspective d'une étude d'oncologie comparée entre l'homme et le chien, nous avons constitué un panel de gènes décrits comme impliqués dans les processus de tumorigenèse et orthologues entre l'homme et le chien.

Définition du répertoire de gènes de cancer

Dans un premier temps, nous avons répertorié un ensemble de 1039 gènes humains proposés dans les bases de données OncoKB (actualisation de mai 2019) [108]. De cet ensemble, nous avons conservé les gènes identifiés comme orthologues à des gènes canins avec un haut niveau de confiance, selon la base de données Ensembl Compara (v98) [109], comme décrit dans Herrero et *al.* [110]. Nous avons également inclus à notre panel 672 gènes orthologues entre l'homme et le chien, décrits par la société Nanostring comme impliqués dans la réponse immunitaires aux thérapies du cancer [111]. En associant ces deux ressources complémentaires, nous avons établi un jeu de données final composé de 1317 gènes connus comme impliqués dans la biologie des cancers et orthologues entre l'homme et le chien. Ce panel de gènes de cancers est librement disponible à https://

github.com/ckergal/BLIMP/blob/main/manuscript/input_data/coordinates/. Cet ensemble nous assure un double objectif, avec à la fois l'évaluation de la performance des modèles en termes de prédiction, mais aussi la prédiction de l'impact des mutations régulatrices au sein des séquences promotrices de ces gènes. Nous définissons ces séquences promotrices comme étant la région de 1024 nucléotides en amont du site de début de transcription des gènes (TSS).

Caractéristiques génétiques

Pour toutes les séquences promotrices de ces 1317 gènes, nous avons calculé leur composition en certaines caractéristiques génétiques comme le pourcentage en GC, leur teneur en éléments transposables ou encore leur niveau d'orthologie entre l'homme et le chien.

Les éléments transposables, ou transposons, représentent des séquences d'ADN capables de se déplacer de manière autonome au sein d'un génome. Parmi eux, la position génomique des SINEs (Short Interspred Nuclear Elements), définit par RepeatMasker [112], a été extraite depuis UCSC [113]. La présence d'éléments transposables spécifiques à une espèce particulière constitue une difficulté lors d'études comparatives interespèces [114]. La famille de SINE unique à l'ordre Carnivora est appelée Can-SINE ou SINEC. Le génome canin est composé à 10,7% de SINECs, représentant 1,6 millions d'éléments d'une taille moyenne de 200 nucléotides [101]. Parmi ceux-ci, la sous-famille des SINEC_Cfs, éléments spécifiques du génome canin, est prépondérante. De plus, la sélection artificielle chez le chien a permis la co-sélection et/ou la fixation de SINEs ayant des effets neutres ou légèrement délétères, alors que les insertions de TEs sont généralement éliminées par la sélection naturelle et/ou inhibées de façon épigénétique. Par conséquent, les SINECs sont des contributeurs majeurs aux traits domestiques canins, tels que les motifs de couleur de la robe, la taille des pattes, la morphologie du crâne et plus généralement sont associés au vieillissement et aux maladies canines [101]. Nous avons donc pu définir l'intersection des SINECs avec les promoteurs de notre panel de gènes grâce à la commande intersect de l'outil BEDTools, version 2.25 [115]. Les séquences promotrices des gènes canins de notre panel détiennent entre 0 et 6 SINECs. Nous comptons 499 séquences détenant au moins un SINEC avec une taille moyenne de 205,9 nucléotides.

La composition en GC du génome canin est supérieure à celle retrouvée dans le génome d'autres mammifères, dont l'homme [82]. Aussi, les séquences génomiques canines particulièrement riches en GC chevauchent les TSS des gènes et sont donc très représentées au sein de notre panel des séquences promotrices des gènes impliqués dans la tumorigenèse (Fig. 2.2). Le taux médian de GC dans les séquences génomiques canines d'intérêt est de 57,5% et de 51,9% chez l'homme.

Le taux de conservation (Fig. 2.3) des séquences promotrices d'intérêt entre l'homme et le chien a été mesuré grâce à l'outil BLAT (BLAST-Like Alignment Tool) v35 [116]. BLAT est régulièrement utilisé en génomique comparative car optimisé pour l'alignement pairé de séquences d'ADN. Les scores calculés à partir de la longueur de l'alignement et la similarité des séquences à l'issue de l'utilisation de cet outil ont été utilisés pour évaluer l'alignement des séquences de notre panel.

2.2 Algorithme de prédiction

L'outil Basenji [103] utilise l'architecture d'apprentissage profond des CNN, Convolutional Neural Network) [117] pour modéliser le niveau d'expression des gènes comme une fonction de l'ADN. Proposé par Kelley et *al.*, Basenji est disponible via la plateforme collaborative GitHub (https://github.com/calico/basenji) et l'ensemble des codes qui composent l'outil sont écrits en langage Python.

2.2.1 Intégration des données

L'utilisation de l'outil Basenji pour la création d'un modèle de prédiction du niveau d'expression des gènes nécessite en premier lieu une étape d'intégration des données. Ce processus a pour but de prendre en compte et de corriger l'influence de certains biais provenant des expériences de séquençage génomique fonctionnel. Le script bam_cov.py de Basenji permet d'estimer la couverture génomique des lectures alignées à l'échelle du nucléotide et de normaliser le biais dû à la teneur en GC et en séquences répétées des régions génomiques. Cette étape a pour effet de convertir les fichiers BAM d'alignements en fichiers bigWig, un format binaire indexé utilisé pour stocker des ensembles de données denses et continues.

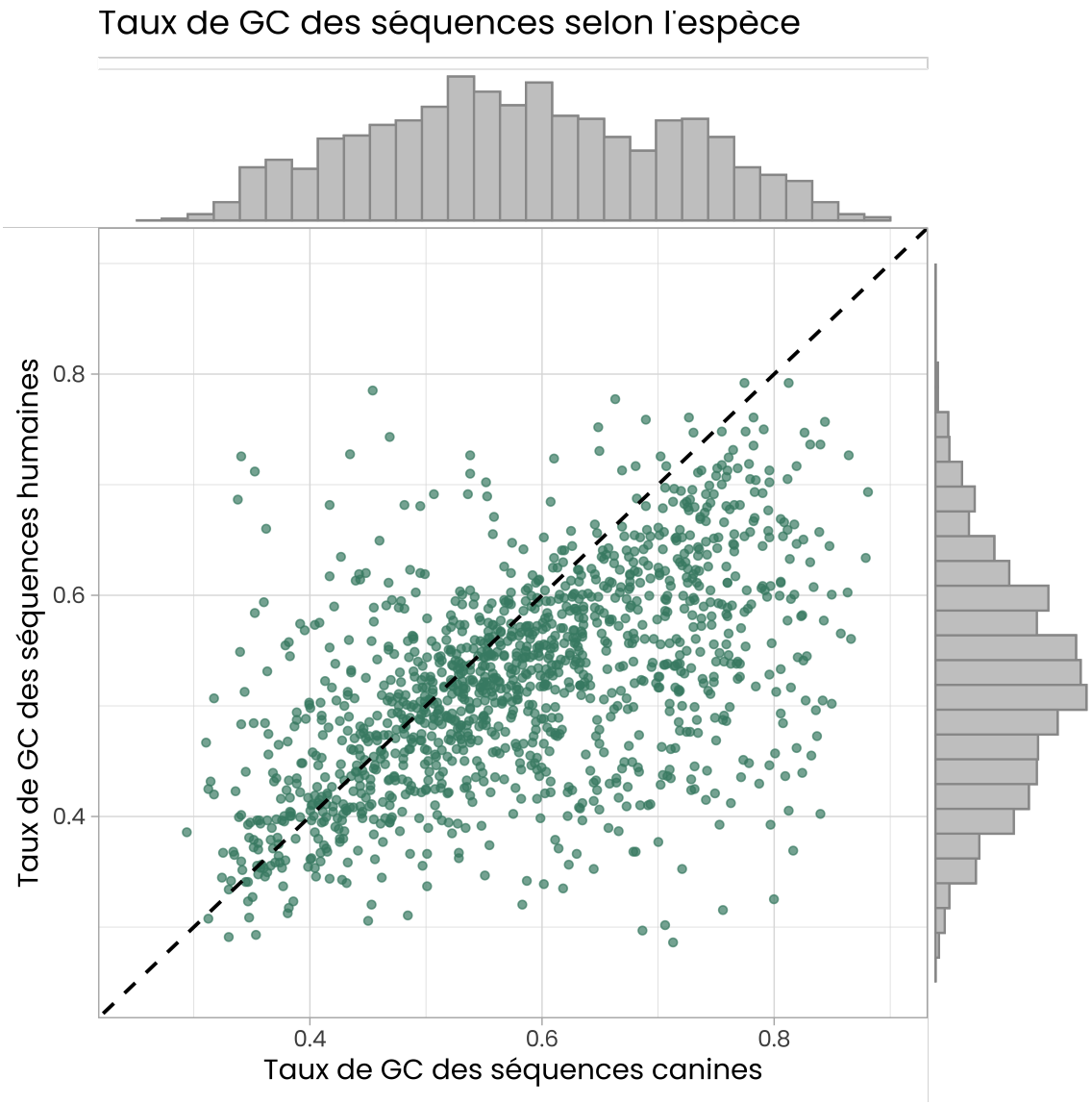

Figure 2.2 – Taux de GC des séquences promotrices des gènes de cancers selon l'espèce. Un point représente la séquence promotrice d'un gène inclu dans notre panel avec en abscisse le taux de GC pour les séquences canines et en ordonnées pour les séquences humaines. En haut et à droite du graphique sont représentées les répartitions respectives.

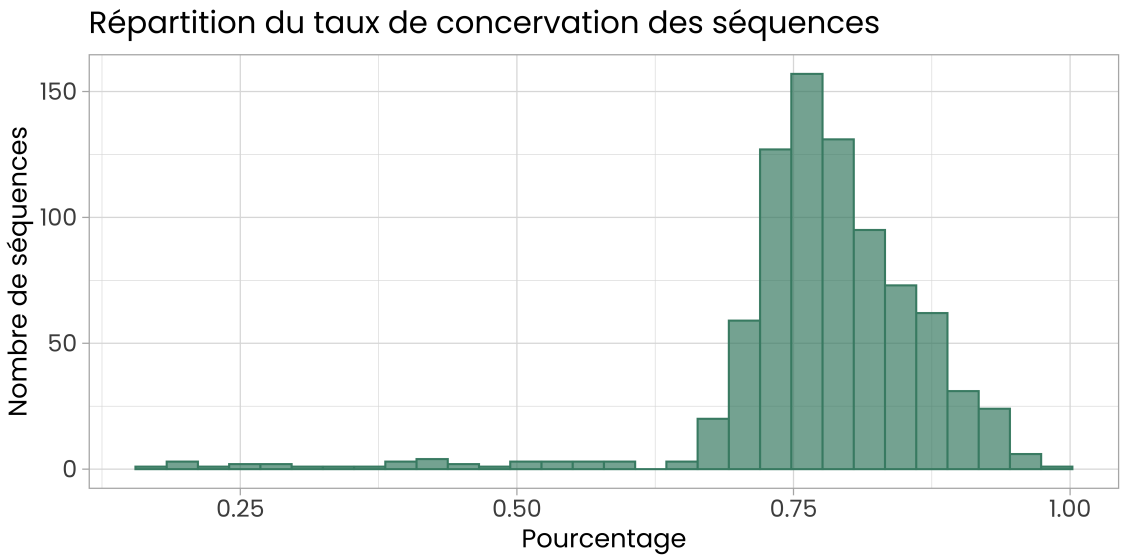

Figure 2.3 – Répartition du taux de conservation des séquences. Le graphique représente le pourcentage de conservation avec l'homme des séquences promotrices canines des gènes de notre panel.

Ensuite, en excluant les régions unmappables c'est-à-dire difficilement alignables et les gaps d'assemblage de plus de 1000 nucléotides, des séquences non chevauchantes d'une longueur de 131 072 (2^{17}) nucléotides sont extraites des chromosomes des génomes de références. Pour le modèle canin basé sur la version canFam3, nous obtenons 16 986 séquences et pour la version canFam4, contenant moins de régions avec gaps, nous obtenons 17 400 séquences. Ces séquences sont alors séparées en trois sous-ensemble correspondant au jeu d'entraînement, de validation et de test de l'algorithme d'apprentissage automatique (proportion de 80%, 10% et 10%, respectivement) [117]. Pour pouvoir être intégrées par l'algorithme, les séquences génomiques subissent une transformation que l'on appelle encodage one-hot. C'est-à-dire que chaque nucléotide est représenté par une matrice formée d'une combinaison de 0 et de 1, comme présenté dans la figure 2.4.

Ces séquences sont finalement associées aux profils d'expression contenus dans les fichiers d'alignement CAGE [9] en additionnant les valeurs de couverture de 128 nucléotides pour servir d'étiquette de prédiction aux modèles (Fig. 2.5).

Afin de réaliser cette première étape d'intégration des données, nous avons utilisé le script basenji_data.py de l'outil Basenji en précisant la longueur des séquences de 131 072 nucléotides, la taille de fenêtre de 128 nucléotides, le pourcentage de séquence souhaité dans chaque ensemble d'entraînement, validation et test du modèle, le génome

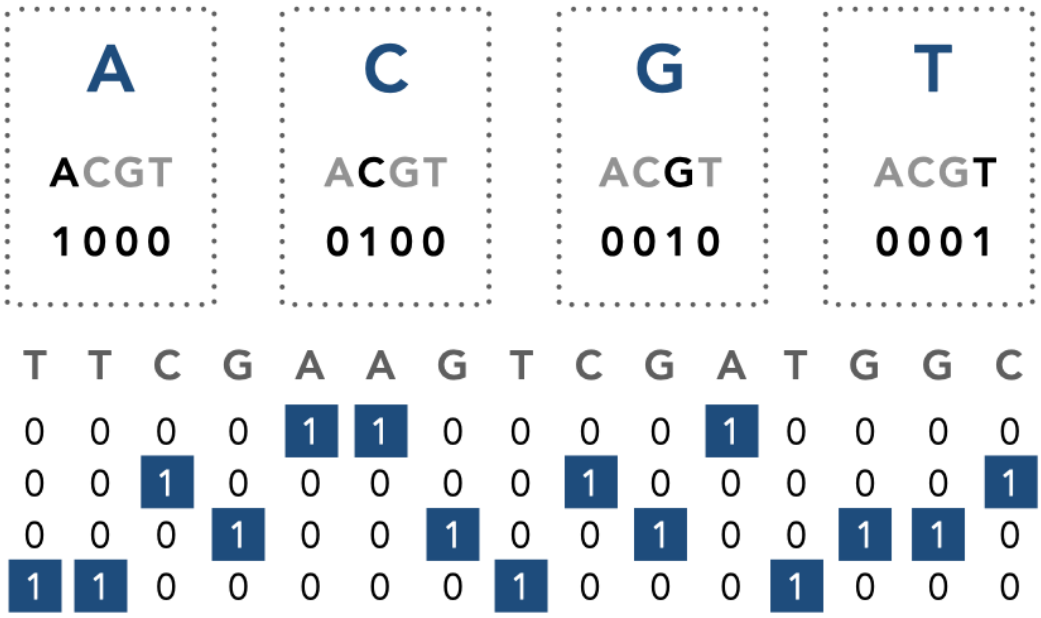

Figure 2.4 – Principe de l'encodage one-hot. Chaque nucléotide des séquences génomiques est codé par une combinaison de 0 et de 1.

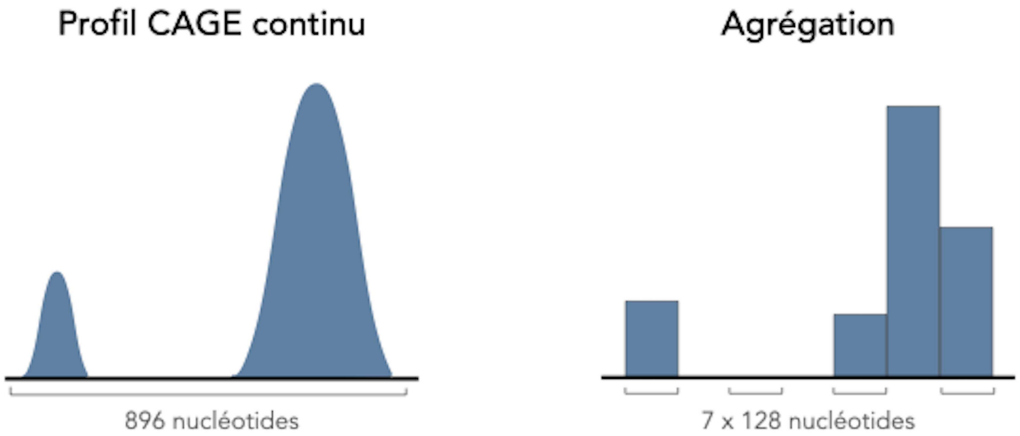

Figure 2.5 – Schéma des profils CAGE. Le schéma de gauche représente le profil d'expression continu de pics CAGE dans un tissu, associés à une séquence génomique de 896 nucléotides. Le schéma de droite symbolise la synthétisation du profil de gauche suite à l'étape d'intégration des données de Basenji.

de référence (canFam3 ou canFam4) au format FASTA ainsi l'emplacement des fichiers d'alignement CAGE d'intérêt. L'ensemble des lignes de commandes ayant permis cette tâche est disponible sur le dépôt GitHub dédié (https://github.com/ckergal/BLIMP).

2.2.2 Entraînement du modèle de prédiction

L'algorithme proposé avec l'outil Basenji déploie un réseau neuronal convolutif profond pour prédire les valeurs de couverture expérimentales (*i.e* données CAGE) en fonction de la séquence d'ADN sous-jacente. La structure du réseau est constituée en premier lieu de six couches de convolution, suivies de onze couches de convolution dilatées et d'une couche de convolution finale (Fig. 2.6). Toutes les couches appliquent une normalisation par lots (ou batch normalization), une fonction d'activation softplus et une fonction de dropout qui va permettre d'optimiser l'apprentissage et d'éviter le surajustement (ou overfitting). Le sur-ajustement, ou sur-apprentissage, décrit le phénomène selon lequel un modèle décrit trop spécifiquement les données d'entraînement au point d'en devenir inadapté pour la prédiction de nouvelles valeurs. Les couches de convolution standards appliquent une agrégation du maximum dans des fenêtres de deux pour chacune des six couches pour atteindre la taille de fenêtre de 128 nucléotides (26). Cette taille de fenêtre a été choisie car elle correspond à la puissance de deux la plus proche de la distance de 146 nucléotides entre les composants du nucléosome. Les valeurs prédites et mesurées sont ensuite comparées via une fonction de log-vraisemblance de régression de Poisson. Les implémentations du package TensorFlow [118] ont été utilisées pour réaliser l'architecture de ce réseau de neurones.

La particularité de Basenji est de proposer des convolutions dilatées qui sont définies comme des filtres de convolution avec des nœuds vides et dont la taille augmente d'un facteur deux à chaque couche. La connexion dense de ces couches signifie que chaque couche prend en entrée toutes les couches précédentes, au lieu de ne prendre que la couche précédente (Fig. 2.6). Sept couches de convolution dilatées sont appliquées afin d'atteindre une largeur de champ réceptif de ~32 kb, permettant de capturer une partie des interactions régulatrices distales.

Le principe de l'entraînement consiste dans un premier temps à attribuer des poids, initialisés à l'aide de la méthode Glorot, à chacun des liens constituant notre réseau de neurones. Ces poids constituent les paramètres du modèle, que l'on peut donc voir

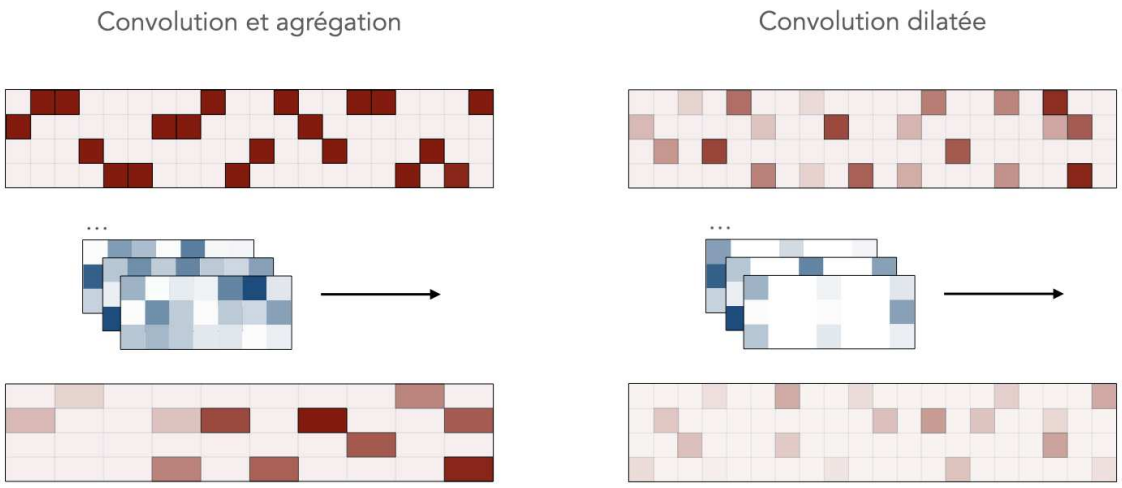

Figure 2.6 – Convolutions de l'algorithme. Le schéma de gauche représente la convolution classique implémentée dans Basenji. Les matrices dans les tons bleus sont les différents filtres de convolution (paramètres) entraînés. Le schéma de droite représente une convolution dilatée.

comme une fonction de prédiction. À partir de ce premier ensemble de poids, la dernière couche du réseau (couche de prédiction) propose un niveau d'expression pour chaque séquence génomique de l'ensemble d'entraînement dans chacun des échantillons composant notre jeu de données initial. À cette étape, la version du modèle résultant est appliquée pour prédire le niveau d'expression des séquences génomiques du jeu de validation. Pareillement, nous obtenons une prédiction du niveau d'expression dans chacun des échantillons du jeu de données. Ces prédictions sont comparées au niveau d'expression réel et un coefficient de corrélation de Pearson est calculé entre les deux grandeurs pour évaluer la qualité du modèle. Cet ensemble d'instructions désigne une époque (ou epoch) de l'algorithme.

Une deuxième époque est ensuite réalisée en ajustant les paramètres (valeurs des poids attribuées aux liens du réseau) et en comparant le coefficient de corrélation nouvellement calculé, toujours à partir des séquences du jeu de validation, à celui obtenu à l'issue de la première époque. Le principe de l'algorithme consiste à réaliser un certain nombre d'époques et de sélectionner le modèle déduit de l'époque aboutissant au coefficient de Pearson le plus élevé. Cet enchaînement d'étape et d'ajustement des valeurs attribuées aux paramètres désigne l'entraînement du modèle. Au cours de l'entraînement, la valeur du coefficient de corrélation converge vers un maximum comme la fonction de

coût (écart entre les valeurs prédites et celles mesurées expérimentalement) converge vers un minimum. Une fois cet optimum atteint, les époques consécutives ne permettent pas d'améliorer la performance du modèle. Il appartient à l'utilisateur de l'algorithme de déterminer combien d'époques sans amélioration il est nécessaire de réaliser pour définir la dernière meilleure époque comme convergeant vers l'optimum et donc arrêter l'algorithme.

La mise en œuvre de l'algorithme d'entraînement du modèle de prédiction repose sur l'utilisation du script basenji_train.py et nécessite une infrastructure de calcul intensif. En effet, l'utilisation de GPU (Graphics Processing Units, processeur graphique) est incontournable pour le déploiement d'un algorithme de cette envergure. Dans ce contexte, nous avons pu nous appuyer sur les ressources de la plateforme bioinformatique Genouest [119], disposant de deux nœuds de calcul GPU (avec au total 7 cartes graphiques GPU).

2.2.3 Evaluation du modèle de prédiction

Une fois l'algorithme d'entraînement achevé, le modèle appris à partir des séquences génomiques provenant de l'ensemble de test (10% du génome) est évalué. De même qu'avec le jeu de validation, le modèle final prédit le niveau d'expression du jeu de test dans chacun des échantillons admis lors de l'entraînement. L'évaluation globale du modèle est portée par le coefficient de corrélation entre les prédictions et le niveau d'expression mesuré expérimentalement sur ces régions génomiques (Fig. 2.7) qui n'interviennent pas dans l'entraînement du modèle. Cette étape se réalise grâce à l'utilisation du script basenji_test.py.

2.2.4 Méthode d'optimisation des modèles de prédictions

Nous avons vu que l'algorithme d'entraînement du modèle de prédiction avait pour objectif de converger vers un optimum et qu'une fois cet optimum atteint, les paramètres du réseau sont optimisés et le modèle est alors considéré comme le plus performant possible. En revanche, il est possible de modifier d'autres variables liées à l'entraînement du modèle avec l'intention d'en améliorer les performances. Les caractéristiques liées à l'architecture du modèle comme le nombre de couches de convolution, la valeur du dropout ou encore le taux d'apprentissage constituent ce que l'on appelle les hyperparamètres du

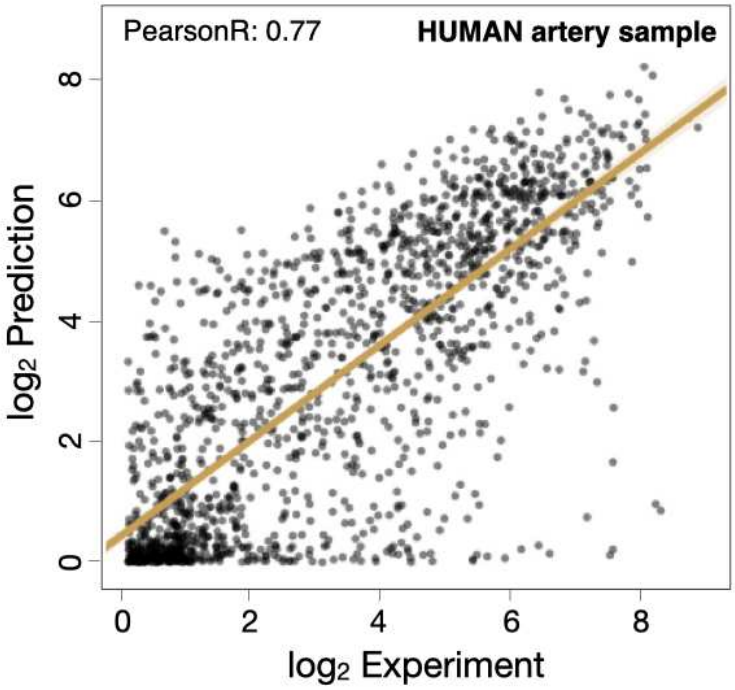

Figure 2.7 – Évaluation du modèle de prédiction. Un point représente une fenêtre de 128 nucléotides provenant des séquences du jeu de test de l'algorithme. En abscisse est indiqué le niveau d'expression mesuré expérimentalement dans un tissu (ici l'artère de l'homme) et celui prédit par le modèle est indiqué en ordonnée. La corrélation entre les prédictions et les valeurs réelles est calculée par un coefficient de Pearson, ici 0,77.

réseau de neurones. Les hyperparamètres sont fixés en amont de l'étape d'entraînement, sont donc externes au processus d'entraînement, restent statiques durant l'entraînement et ne sont pas ajustés lors de la réalisation de l'algorithme, ce qui les différencient des paramètres classiques.

L'optimisation des hyperparamètres consiste alors à ajuster certains éléments de la structure de l'algorithme et à évaluer l'impact de ces changements sur les performances de prédictions. Pour la création des deux modèles de prédiction du niveau d'expression des gènes canins, nous avons fait le choix d'optimiser conjointement le taux d'apprentissage et un hyperparamètre de normalisation, appelé L2_Scale dans l'outil Basenji qui va permettre de trouver l'optimum lié à l'apprentissage plus rapidement. Cette décision a été motivée suite à des discussions avec l'auteur de Basenji (David Kelley) et aussi par le temps de calcul nécessaire induit par l'ajout d'hyperparamètres supplémentaires à optimiser.

La méthode choisie, appelée "recherche par quadrillage" ou plus communément "grid search" [120], consiste à entraîner les modèles formés depuis plusieurs combinaisons possibles d'hyperparamètres. À partir de la littérature [121], nous avons fixé les bornes de recherche des deux hyperparamètres entre 10^{-4} et 0,1 pour le taux d'apprentissage et 10^{-8} et 10^{-3} pour le L2 scale, avec un pas logarithmique entre chaque valeur. Ainsi, de ce design expérimental résulte la création de 42 architectures de modèles dont les performances sont finalement comparées pour sélectionner la combinaison menant aux prédictions les plus proches des valeurs expérimentales.

RÉSULTATS

3.1 Prédiction du niveau d'expression des gènes canins

Le premier objectif de ce projet de thèse consistait en la création d'un modèle de prédiction du niveau d'expression des gènes canin à partir d'un algorithme performant basé sur une architecture de réseaux de neurones convolutifs.

Dans cette optique, nous nous sommes appuyés sur l'utilisation de l'outil Basenji [103] et des génomes de référence canins associés à des profils CAGE [9] représentant le niveau d'expression des gènes dans divers échantillons biologiques afin d'établir deux modèles. Dans un premier temps, nous avons réalisé l'entraînement d'un modèle avec la version d'assemblage du génome canin canFam3.1 [22] et les profils d'expression CAGE de 134 échantillons alignés sur canFam3.1. Par la suite, grâce à la disponibilité des données de séquençage alignées sur la version canFam4 [83] du génome canin par nos collaborateurs du consortium DoGA, nous avons entraîné un second modèle de prédiction comprenant 116 profils d'expression.

3.1.1 Création des modèles

Version basée sur l'assemblage canFam3

Nous avons créé un modèle de prédiction du niveau d'expression des gènes canins à partir de la version d'assemblage canFam3.1 et des données CAGE représentant le profil d'expression d'échantillons biologiques canins. À partir d'une architecture d'hyperparamètres disponible *via* https://github.com/ckergal/BLIMP (cf. section Méthode du manuscrit), l'étape d'entraînement du modèle a nécessité la réalisation de 50 époques pour une durée moyenne de 16,5 minutes par époques et une durée totale de 13,8 heures. La méthode que nous utilisons pour stopper l'entraînement est de limiter le nombre d'époques lorsqu'un critère de performance a cessé de s'améliorer. Ici l'algorithme était

programmé pour s'arrêter lorsque 30 époques sans amélioration du coefficient de Pearson étaient effectuées. Utilisée par l'algorithme initial de Basenji [103], cette approche de l'arrêt précoce permet de stopper l'entraînement une fois que les performances du modèle cessent de s'améliorer sur l'ensemble de données de validation correspondant à 10% du génome.

Ainsi, le modèle défini par cette étape d'entraînement a été réalisé à l'occasion de la 19ème époque. Le jeu d'entraînement, composé de 13 860 séquences génomiques de 131 kb associées au niveau d'expression des gènes dans 136 échantillons et représentant 80% du génome, est utilisé à chaque époque pour ajuster les paramètres du réseau de neurones convolutif. Nous disposons ainsi d'une valeur de corrélation de Pearson et de fonction de coût calculées entre le niveau d'expression prédit et celui mesuré expérimentalement pour les séquences génomiques du jeu d'entraînement et cela à chaque époque de l'algorithme d'entraînement (Fig. 3.1).

Aussi, les 1551 séquences génomiques du jeu de validation sont utilisées pour contrôler la capacité du modèle appris à être généralisable, c'est-à-dire performant lorsqu'il est appliqué à d'autres données que celles du jeu d'entraînement. À chaque époque de l'algorithme, la corrélation de Pearson et la fonction de coût sont calculées entre les prédictions du niveau d'expression des séquences du jeu de validation et les niveaux d'expression attendus, connus par les données CAGE. Lorsque ces valeurs convergent, cela signifie que le modèle est le plus performant possible dans le sens où il a suffisamment appris des données d'entraînement pour modéliser de manière fiable la relation entre les séquences d'ADN et le niveau de transcription, sans pour autant engendrer de sur-ajustement.

Version basée sur canFam4

Au cours de cette thèse et suite à la réalisation du modèle de prédiction canin basé sur la version canFam3 du génome du chien, le consortium DoGA [99] a réalisé une mise à jour de sa banque de données en alignant les données CAGE sur la version canFam4 du génome de référence canin. Nous avons ainsi pu disposer de 116 fichiers CAGE afin de réaliser un nouveau modèle de prédiction du niveau d'expression des gènes canins, basé sur cette nouvelle version d'assemblage.

L'entraînement de ce nouveau modèle a nécessité la réalisation de 62 époques pour un temps de total de 17,4 heures (16,8 minutes en moyenne par époque). Tout comme

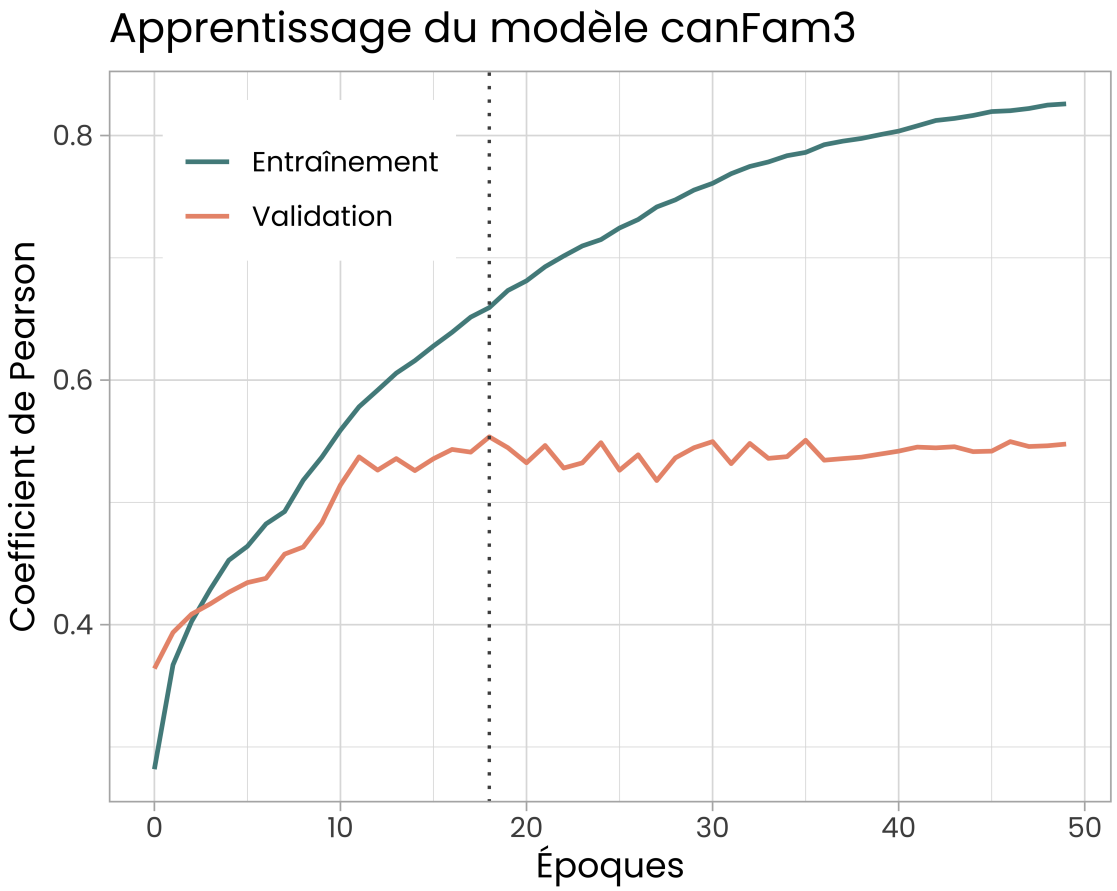

Figure 3.1 – Entraînement du modèle canFam3. Évolution du coefficient de Pearson (axe des ordonnées), pour le jeu d'entraînement et le jeu de validation au cours des époques (axes des abscisses) de l'algorithme d'apprentissage du modèle basé sur l'assemblage canFam3.1. La ligne en pointillée marque le modèle retenu, ici réalisé lors de la 19ème époque.

avec le modèle de la version d'assemblage canFam3, l'algorithme était programmé pour s'arrêter après 30 époques sans amélioration du coefficient de Pearson. L'époque menant au modèle sélectionné à l'issue de l'algorithme apparaît à la 30ème position.

La séquence du génome canFam4 étant plus complète que celle de la version canFam3, les ensembles d'entraînement, de validation et de test contiennent davantage de séquences génomiques de 131 kb. Ainsi, nous disposons d'un jeu d'entraînement composé de 14 214 séquences associées, cette fois, à 116 profils d'expression CAGE pour entraîner le modèle de prédiction du niveau d'expression des gènes. L'évolution de la valeur de corrélation de Pearson suit la même tendance que celle observée pour l'entraînement du modèle basé sur canFam3 (Fig. 3.2). Le jeu de validation contient 1591 séquences génomiques dont on mesure également la corrélation et la fonction de coût entre les niveaux d'expression prédits et ceux mesurés expérimentalement.

3.1.2 Mesure des performances

Une fois les étapes d'entraînement des modèles achevées, nous avons évalué leurs performances grâce aux jeux de données de test (10% restant du génome). En effet, cette dernière étape consistait à prédire le niveau d'expression des séquences génomiques non exploitées lors de la phase l'entraînement grâce au modèle issu de la précédente étape. Le niveau d'expression est prédit par fenêtres de 128 nucléotides pour chacune des séquences du jeu de test. Nous avons obtenu, par conséquent, 1024 prédictions par séquence de 131 kb dans chaque échantillon admis dans l'entraînement des modèles.

Évaluation du modèle entraîné avec canFam3

Pour le modèle entraîné sur la version d'assemblage canFam3, nous avons utilisé 1575 séquences génomiques représentant donc 10% du génome n'intervenant pas dans la phase d'entraînement du modèle. Afin d'évaluer la performance du modèle, nous avons considéré la corrélation entre les (1024 x 1575) ~1,61 millions valeurs correspondant aux niveaux d'expression prédits et ceux mesurés expérimentalement pour les 134 échantillons. Pour associer la performance du modèle à une valeur, nous avons calculé la médiane des 134 valeurs de corrélation, s'élevant ici à 0,64 et un intervalle des valeurs allant de 0,43 pour le tissu foie à 0,75 pour l'aorte (Fig. 3.3).

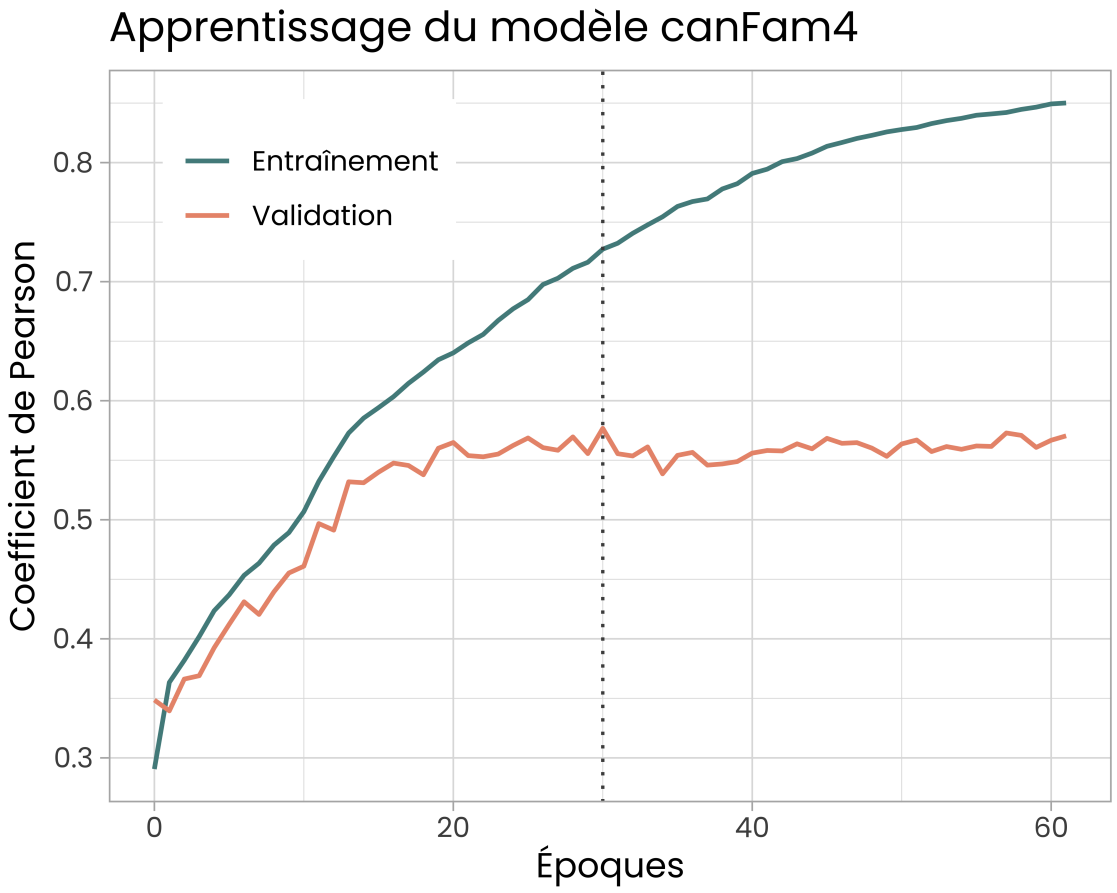

Figure 3.2 – Entraînement du modèle canFam4. Évolution du coefficient de Pearson pour le jeu d'entraînement et le jeu de validation au cours de l'algorithme d'apprentissage du modèle basé sur l'assemblage canFam4. La ligne en pointillée marque le modèle retenu, ici réalisé lors de la 30ème époque.

Évaluation du modèle entraîné avec canFam4

Le modèle entraîné à partir de l'assemblage canFam4 du génome canin est évalué selon le même procédé. L'utilisation de cette version du génome de référence du chien aboutissait à un jeu de test composé de 1595 séquences génomiques. De ce fait, (1024 x 1595) ~1,63 millions niveaux d'expression prédits sont comparés aux niveaux d'expression réels correspondant en calculant le coefficient de corrélation de Pearson entre les deux grandeurs, dans les 116 échantillons utilisés pour la création de ce modèle de prédiction. La médiane des 116 corrélations s'élevait pour ce modèle à 0,67 avec un intervalle des valeurs allant de 0,34 pour le tissu pancréas à 0,76 pour la neurohypophyse (Fig. 3.3).

Comparaison avec le modèle établi chez l'homme

Le modèle de prédiction du niveau d'expression des gènes humains a été établi par l'auteur de l'outil Basenji, David Kelley [103]. Son accès et son utilisation sont disponibles grâce au répertoire GitHub (https://github.com/calico/basenji/) associé à l'outil et son application. Ainsi, nous avons pu utiliser le jeu de données réservé au test du modèle, composé de 1937 séquences de 131 kb extraites de la version d'assemblage GRCh38 [97] du génome humain et associées à 5313 profils transcriptomiques dont 638 profils CAGE [107]. Du fait de la composition en profils CAGE uniquement des jeux de données canins, nous avons analysé les prédictions du niveau d'expression des séquences de test du modèle humain seulement dans les échantillons séquencés par la technologie CAGE. Nous avons donc calculé la corrélation entre (1024*1937) ~1,98 millions niveaux d'expression et ceux mesurés expérimentalement dans les 638 jeux de données CAGE humain. La médiane de ces valeurs s'élève à 0,66 avec un intervalle des valeurs allant de 0,18 pour la substantia nigra à 0,76 pour les cellules du muscle squelettique (Fig. 3.3).

3.1.3 Optimisation des hyperparamètres

Afin de proposer des modèles de prédiction du niveau d'expression des gènes canins plus performants, nous avons procédé à l'optimisation des hyperparamètres liés aux architectures de réseaux de neurones utilisés (cf. la section méthode de ce manuscrit). Suite à des échanges avec le concepteur de Basenji (David Kelley) quant aux stratégies à appliquer pour employer au mieux son outil sur un nouvel organisme, nous avons

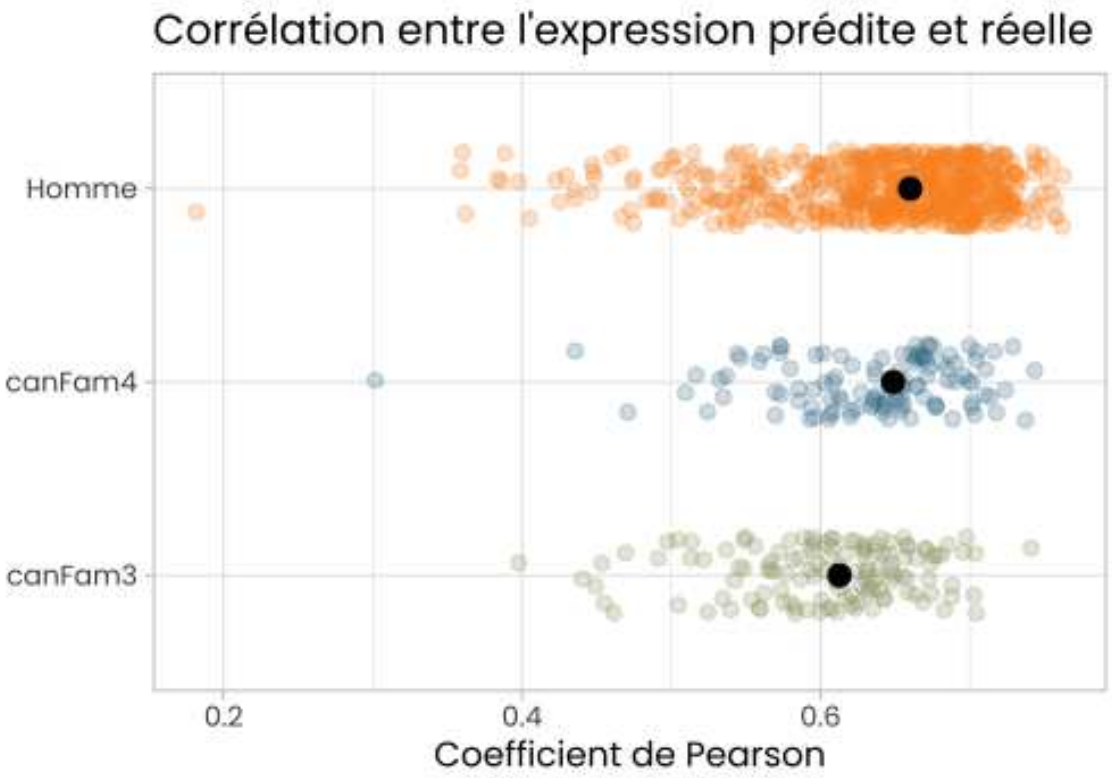

Figure 3.3 – Évaluation des modèles de prédiction. Un point correspond au coefficient de Pearson associé à un tissu composant le modèle précisé en ordonnée. Les points noirs représentent les médianes des corrélations de chaque modèle, 0,66 pour le modèle humain (Homme en vert), 0,65 pour le modèle basé sur la version canFam4 du génome canin (en bleu) et 0,61 pour le modèle basé sur la version canFam3 (en jaune).

entrepris l'optimisation de deux hyperparamètres (learning rate ou taux d'apprentissage et L2 scale ou régularisation L2) au moyen d'une méthode de recherche par quadrillage ou "grid search" (cf section Matériel et Méthodes). Cette approche consiste à fixer un ensemble de valeurs pour chaque hyperparamètre que l'on souhaite optimiser. Puis chaque combinaison d'hyperparamètres va être utilisée pour entraîner un modèle dont on garde en mémoire la performance. Finalement, il s'agit de sélectionner l'ensemble des hyperparamètres qui aboutit au modèle le plus performant (cf. la section méthode de ce manuscrit).

Pour chaque version d'assemblage, nous avons donc créé 42 modèles correspondant aux architectures construites à partir des combinaisons de 6 valeurs pour le taux d'apprentissage et de 7 valeurs pour la régularisation L2 (Fig. 3.4). L'entraînement de ces 84 modèles au total s'est déroulé sur une période de temps considérable. En effet, un modèle doit être entraîné sur une des sept cartes graphiques GPU du serveur de calcul genouest. Ainsi, selon les disponibilités et les priorités accordées aux différents utilisateurs de cette ressource de calcul, sept modèles au maximum pouvaient être entraînés simultanément. Avec un temps moyen d'entraînement d'environ 30 heures, il a fallu plus d'une semaine pour réaliser l'optimisation de chaque modèle.

Pour le modèle entraîné avec la version d'assemblage du génome canin canFam3 [22], l'optimisation indique que le meilleur ensemble d'hyperparamètres est constitué d'un taux d'apprentissage de 10^{-2} et d'un L2 Scale de 10^{-6} (Fig. 3.4). L'entraînement de ce modèle a nécessité 34 heures, 123 époques avec un temps moyen de 16,6 minutes par époque. L'époque aboutissant à la meilleure valeur de corrélation de Pearson a été réalisée en 92ème position. L'évaluation du modèle retenu avec les données provenant du jeu de test indique une corrélation médiane de 0,64, avec un intervalle des valeurs allant de 0,43 pour le tissu foie à 0,75 pour l'aorte (Fig. 3.5).

En ce qui concerne le modèle de prédiction entraîné à partir de la version canFam4 [83] du génome du chien, le processus d'optimisation à conduit au choix d'un modèle construit à partir d'un taux d'apprentissage valant 10^{-2} et un l2 scale de 10^{-6} (Fig. 3.4). Avec une durée moyenne de 16,8 minutes, 200 époques ont étées nécessaires pour l'entraînement de ce modèle d'une durée totale de 56 heures. La 169ème époque de l'apprentissage du modèle proposait le meilleur coefficient de corrélation calculé à partir du jeu de validation. L'évaluation de ce modèle avec les données du jeu de test indique un coefficient de corrélation de Pearson de 0,67, avec un intervalle des valeurs allant de

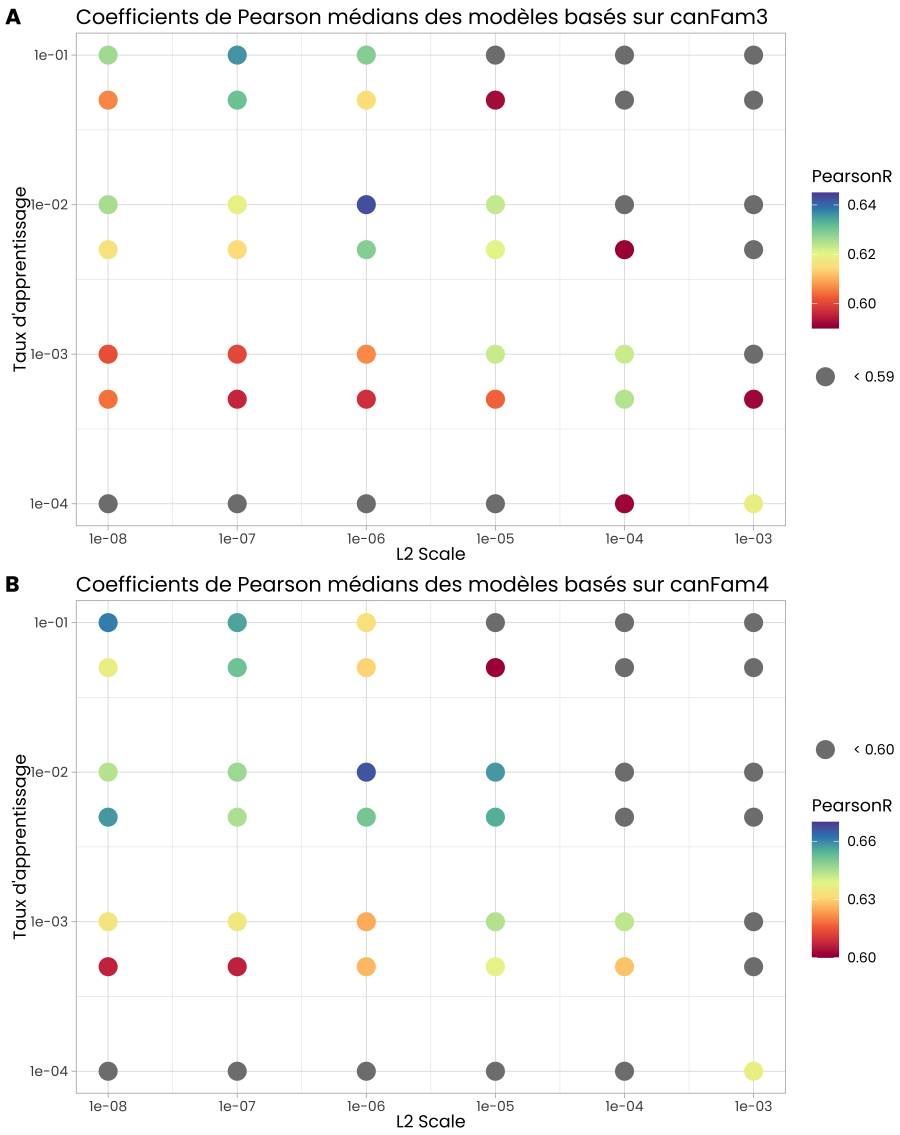

Figure 3.4 – Optimisation des modèles de prédiction canins. Un point représente un modèle de prédiction entraîné à partir de la combinaison d'un L2 scale en abscisse et d'un taux d'apprentissage en ordonnée. La couleur du point correspond à la valeur médiane du coefficient de Pearson évalué à partir du jeu de test. **A.** Modèles basés sur la version d'assemblage canFam3.1. **B.** Modèles entraînés à partir de la version canFam4 du génome canin.

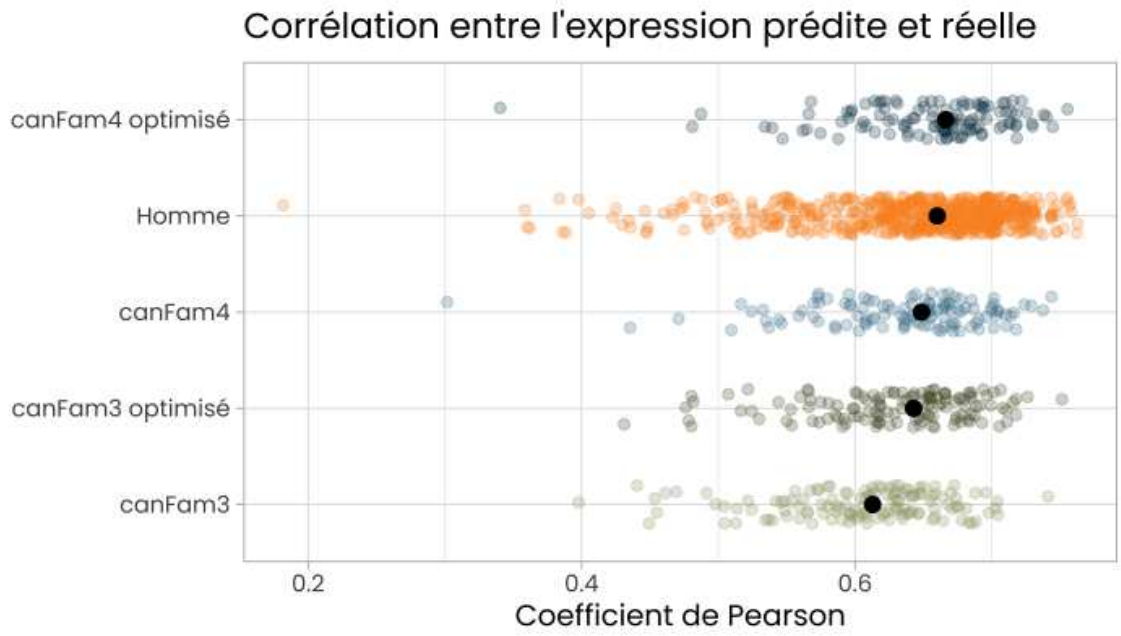

Figure 3.5 – Évaluation des modèles de prédiction après optimisation. Un point correspond au coefficient de Pearson associé à un tissu composant le modèle précisé en ordonnée. Les points noirs représentent les médianes des corrélations de chaque modèle. Le modèle optimisé canFam3 atteint une médiane de 0,64 (0,61 sans optimisation), le modèle canFam4 atteint une médiane de 0,67 (0,65 sans optimisation).

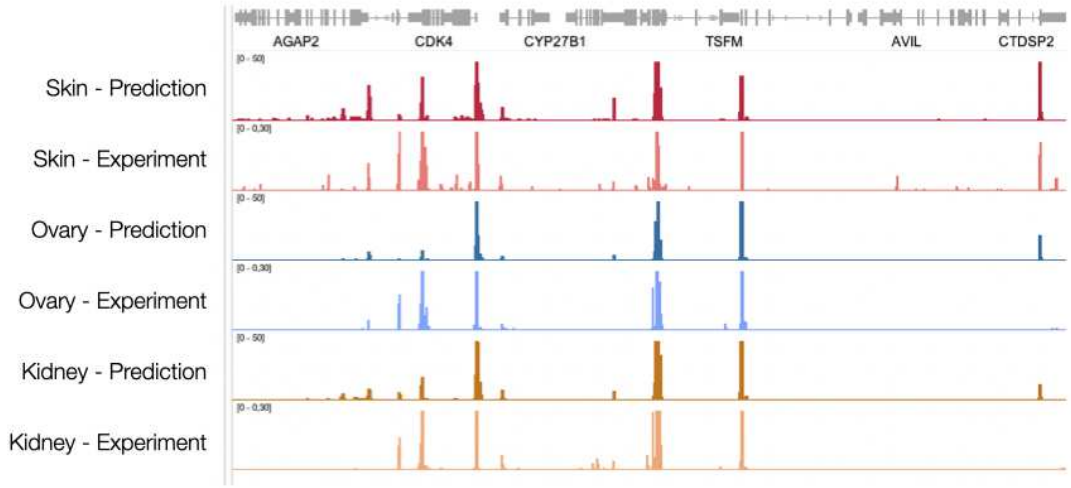

Figure 3.6 – Visualisation des prédictions d'expression par le logiciel IGV (Integrative Genomics Viewer) [122]. Cette illustration permet de comparer le niveau d'expression mesuré expérimentalement dans trois tissus avec celui prédit par le modèle basé sur le génome canFam4 (de haut en bas sont représentés les profils pour les tissus peau en rouge, ovaire en bleu et reins en orange). Une teinte représente un tissu, les versions claires représentent les valeurs expérimentales et les versions foncées sont les niveaux d'expression prédits. La piste du haut représente l'annotation des gènes sur la région génomique considérée.

0,34 à 0,76 (Fig. 3.5).

Une autre fonctionnalité de Basenji permettant d'apprécier les prédictions établies par le modèle d'apprentissage consiste à utiliser le script *basenji_predict_bed.py* en précisant un génome d'intérêt dont on souhaite obtenir la prédiction du niveau d'expression. Ainsi, l'outil constitue des fichiers exploitables par les logiciels de visualisation de génomes (fichiers .bigwig) qui permettent d'évaluer de manière visuelle la fiabilité du modèle (Fig. 3.6) en comparant le niveau d'expression prédit par le(s) modèle(s) au niveau d'expression réel issus des données expérimentales CAGE.

3.2 Comparaison des approches inter-espèce et intra-espèce

La création des modèles de prédictions du niveau d'expression des gènes canins repose sur l'hypothèse selon laquelle ces derniers permettent de prédire de manière plus performante que d'autres modèles déjà établis. Ce questionnement est justifié dans le sens où il est possible d'utiliser un modèle de prédiction entraîné à partir du génome et de profils d'expression d'une espèce précise en prédisant le niveau d'expression de séquences génomiques provenant d'une autre espèce. L'approche classique, utilisée précédemment pour évaluer les modèles canins et le modèle humain correspond à une méthode dite intra-espèce (ou within-species) tandis que la prédiction de séquences génomiques provenant d'une autre espèce que celle correspondant au modèle décrit une approche inter-espèce, ou cross-species.

3.2.1 Approches inter-espèces

Plusieurs études [123, 124] ont montré l'intérêt de s'appuyer sur des prédictions réalisées par des approches inter-espèces. Différents critères poussent à considérer ces stratégies à l'instar du manque de données disponibles chez une espèce, d'un défaut de maîtrise technique ou même de temps pour réaliser un modèle de prédiction dédié à une espèce d'intérêt. À titre d'exemple, Chen et *al*. [123] ont montré que la prédiction de régions régulatrices de type enhancers chez le rat, à partir d'un modèle de réseau de neurones convolutif entraîné chez l'homme, est possible et mène à des performances assimilables à celles que l'on peut observer pour des prédictions intra-espèces.

En revanche, des résultats plus récents ont montré l'utilité, lorsque les données sont disponibles, d'entraîner un modèle prédictif relatif à l'espèce d'intérêt [114]. En effet, les approches intra-espèces atteignent de meilleures performances globalement, notamment en raison des difficultés des approches inter-espèces à réaliser des prédictions à partir de séquences faiblement conservées au cours de l'évolution. De plus, les approches intra-espèces présentent l'avantage de la meilleure représentativité des échantillons étudiés.

Afin de conduire l'étude inter-espèces de prédiction du niveau d'expression de séquences génomiques canines à partir du modèle de prédiction entraîné chez l'homme, nous avons dû établir une liste de tissus communs entre les deux espèces. Parmi les

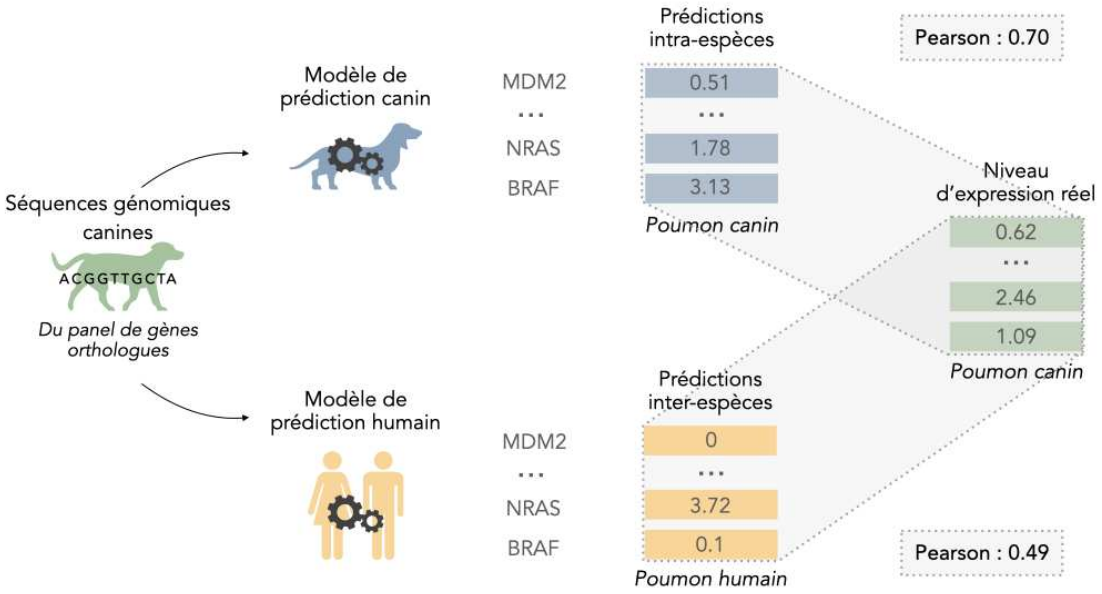

Figure 3.7 – Principe des approches inter-espèce et intra-espèce. Les séquences génomiques canines issues du panel de gènes orthologues sont prises en compte par le modèle de prédiction canin (approche intra-espèce) et le modèle de prédiction humain (approche inter-espèce) pour en prédire le niveau d'expression. Ces deux grandeurs sont ensuite comparées avec le niveau d'expression réel pour un déduire une corrélation, mesurant la performance des deux approches.

638 échantillons séquencés via la technologie CAGE composant le modèle de prédiction humain et les 116 échantillons canins utilisés lors de l'entraînement du modèle basé sur canFam4, nous avons retenu 19 tissus communs selon la description détaillée des métadonnées c'est à dire les informations qui complètent les données expérimentales disponibles. En outre, afin d'éviter un biais dans les prédictions du modèle entraîné sur les données humaines, nous avons fait le choix d'utiliser un panel de gènes orthologues entre l'homme et le chien (cf Méthodes) pour évaluer l'approche inter-espèces. Une représentation schématique des approches intra- et inter-espèces est disponible Fig. 3.7.

Ainsi, les prédictions inter-espèces du niveau d'expression des gènes canins sont réalisées à partir de séquences génomiques canines représentant 1024 nucléotides en amont de gènes orthologues, connus pour être impliqués dans les processus de tumorigenèse. À partir du script basenji_predict_bed.py disponible avec l'outil Basenji et permettant de prédire le niveau d'expression d'une séquence génomique fournie en entrée, nous

avons créé les scripts basenji_predict_par.py et basenji_predict_h5_out.py accessibles *via* notre site dédié https://github.com/ckergal/BLIMP afin de réaliser nos études inter-espèces. Leur utilisation permet notamment de comparer les prédictions réalisées avec les niveaux d'expression des gènes relatifs à une espèce différente de celle utilisée pour la création du modèle de prédiction mais aussi d'amoindrir le temps de calcul nécessaire à la réalisation de ces prédictions. Les lignes de commandes spécifiques à notre utilisation des scripts sont également disponibles sur le dépôt GitHub de BLIMP.

Les résultats de notre étude comparative entre les approches inter-espèces et intra-espèces (modèle réalisé avec la version canFam4 du génome canin) suggèrent l'intérêt de l'utilisation du modèle de prédiction du niveau d'expression des gènes spécifique à l'espèce canine. En effet, les performances de ce dernier sont supérieures à celles du modèle de prédiction humain, appliqué aux séquences canines. Quel que soit le tissu commun analysé, le coefficient de corrélation de Pearson entre les niveaux d'expression prédits et ceux mesurés expérimentalement sont supérieurs avec le modèle canin qu'avec le modèle humain. Le coefficient médian relatif au modèle canin s'élève à 0,67, avec un coefficient minimal de 0,47 atteint avec le tissu moelle osseuse et maximal de 0,73 atteint avec le tissu trachée. Tandis que pour le modèle humain, la médiane du coefficient de corrélation de Pearson s'élève à 0,45, un minimum de 0,2 par le tissu pancréas et un maximum de 0,56 par le tissu utérus (Fig. 3.8).

3.2.2 Impact des caractéristiques génétiques sur les performances du modèle

Nous nous sommes intéressés à différentes caractéristiques génétiques composant les séquences canines d'intérêt afin de comprendre les différences de performance du modèle de prédiction humain et du modèle de prédiction canin pour estimer le niveau d'expression des gènes,

La structure du génome du chien se différencie de celle de l'homme notamment de par sa teneur en GC. Comme évoqué précédemment, la perte du gène PRDM9 chez le chien [98] explique en partie la présence de régions riches en GC liée au processus de conversion génique biaisé (gBGC). De ce fait, nous avons fait le choix de comparer les performances associées aux prédictions réalisées par l'approche inter-espèces et intra-espèces en fonction de la teneur en GC des séquences génomiques canines dont on prédit

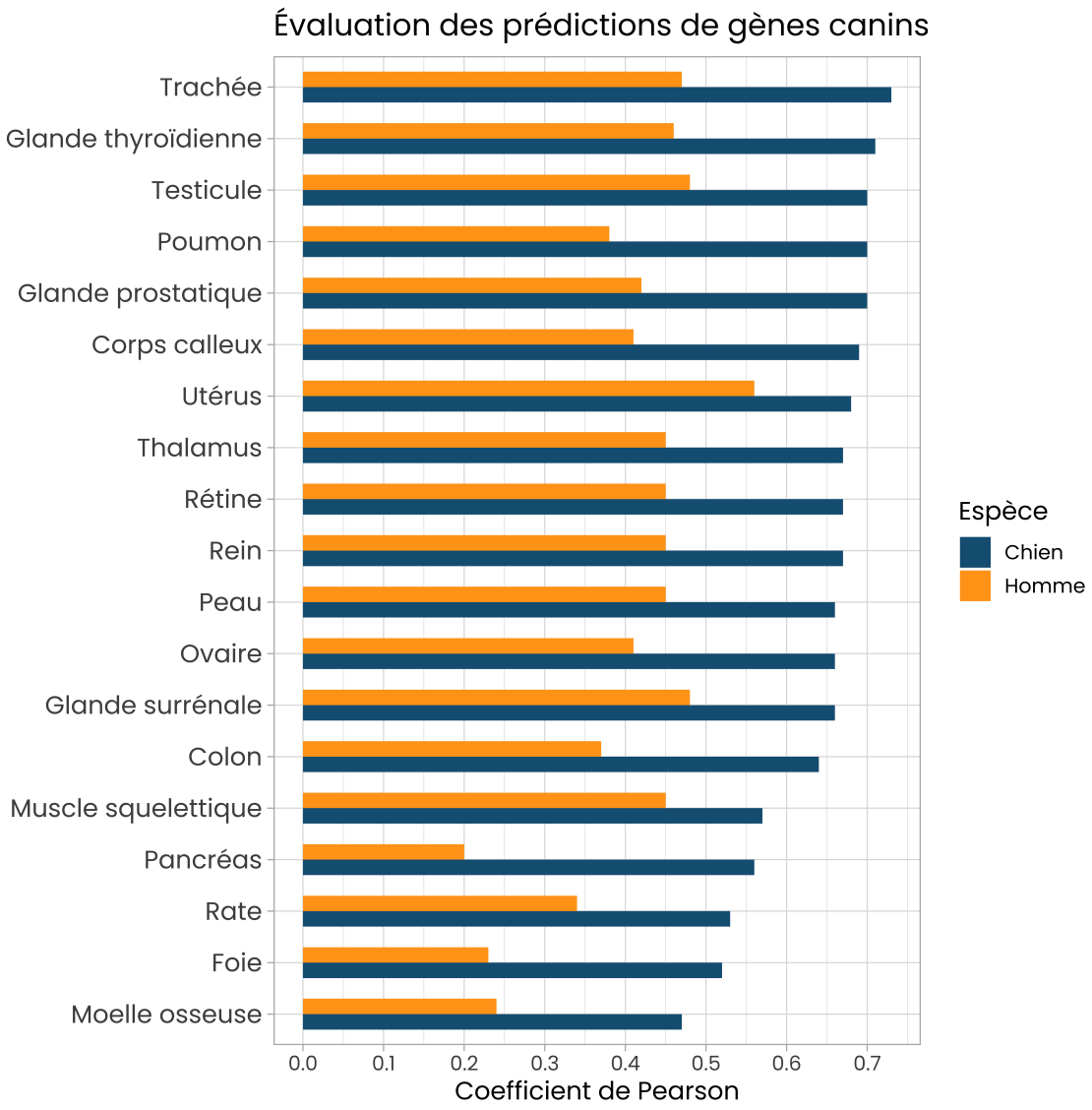

Figure 3.8 – Évaluation des approches inter-espèce et intra-espèce. Chaque barre atteint le coefficient de corrélation de Pearson (en abscisse) calculé dans 19 tissus entre la prédiction et le niveau d'expression réel des séquences génomiques canines du panel de gènes orthologues dans les tissus communs entre l'homme et le chien (ordonnée). Les barres jaunes représentent les prédictions réalisées via le modèle humain, les barres bleues représentent les prédictions réalisées par le modèle de prédiction canin (canFam4).

le niveau d'expression. Dans cette perspective nous avons constitué deux catégories de gènes canins, en fonction de la médiane du taux de GC (57,5%) de leurs séquences promotrices. En prédisant le niveau d'expression des séquences canines à faible taux de GC à l'aide du modèle de prédiction humain, nous observons une corrélation moyenne de 0,39 lorsque la prédiction des séquences canines à fort taux de GC conduit à une corrélation moyenne de 0,43. Cette différence de performance observée avec l'approche inter-espèces selon la composition en GC des séquences canines n'a pas été retrouvée lorsque l'on a prédit le niveau d'expression de ces mêmes séquences via l'approche intra-espèces, c'est-à-dire avec le modèle de prédiction canin (Fig. 3.9).

Nous avons également conduit cette analyse en différenciant les séquences promotrices canines selon leur composition en éléments répétés de type SINEC, une catégorie d'éléments transposables spécifiques des carnivores (Fig. 3.9). Nous avions préalablement montré que l'expression de certains types de gènes canins était associée à la présence d'éléments transposables spécifiques dans leurs promoteurs [101]. De la même manière que pour l'analyse des performances selon la constitution en GC, nous avons donc réparti les séquences canines selon leur composition en SINECs, inférieure ou supérieure à la médiane de l'ensemble. Avec le modèle spécifique à l'homme, la prédiction du niveau d'expression des séquences avec le taux d'éléments transposables le plus faible indique un coefficient de corrélation de Pearson médian de 0,48 tandis que pour les séquences comprenant un taux d'éléments transposables plus élevé, le coefficient médian est de 0,43 suggérant ainsi que la présence de SINEC diminue les performances de l'approche inter-espèces. Cette tendance n'est pas retrouvée en évaluant les prédictions du modèle spécifique du chien selon la teneur en éléments répétés des séquences génomiques étudiées. En effet, le coefficient moyen est de 0,65 pour une faible teneur en éléments répétés et de 0,64 pour les séquences détenant le plus fort taux d'éléments répétés. Ceci suggère que l'approche intra-espèce est moins impactée par la présence de séquence d'éléments transposables dans les promoteurs de gènes car ces derniers ont été intégrés lors de l'entraînement du modèle canin.

Enfin, nous avons analysé le niveau de conservation entre les séquences génomiques canines utilisées dans notre étude et leurs orthologues humains (Fig. 3.9). Cette analyse a confirmé le résultat attendu selon lequel les performances de prédiction sont meilleures lorsque le modèle humain est employé à prédire le niveau d'expression de promoteurs canins fortement conservés avec l'homme, en atteignant un coefficient de corrélation de

Pearson moyen de 0,42. Appliqué à la prédiction de l'expression des séquences canines avec le plus faible taux de conservation avec l'homme, le modèle humain présente une performance plus faible, représentée par une corrélation moyenne de 0,38.

Ces différentes analyses suggèrent une certaine sensibilité des modèles de prédiction quant aux caractéristiques génétiques spécifiques des espèces étudiées. Le génome du chien dispose d'une forte composition en GC et en éléments répétés de type SINEC spécifiques, occasionnant des motifs génomiques particuliers absents du génome humain. La présence de ces motifs inhérents à l'espèce canine explique en partie les performances plus faibles du modèle de prédiction humain. En revanche, ces résultats suggèrent également que sous certaines conditions relatives à la composition des séquences canines (forte conservation évolutive et faible taux de SINECs), il peut être intéressant d'utiliser le modèle de prédiction humain.

L'ensemble des travaux présentés jusqu'ici ont fait l'objet d'une communication orale sélectionnée et d'une publication (cf. Annexe 1) dans le cadre des actes de la conférence JOBIM (Journées Ouvertes en Biologie, Informatique et Mathématiques).

3.3 Impact des mutations régulatrices sur le niveau d'expression des gènes

Les modèles de prédiction canins que nous avons décrits précédemment permettent de prédire le niveau d'expression des gènes d'une séquence nucléotidique dans les échantillons biologiques (134 pour la version canFam3, 116 pour la version canFam4) admis lors de leurs entraînements respectifs. Ainsi, en mutant la séquence du génome de référence utilisé lors de la conception du modèle, l'outil Basenji permet de mesurer la différence entre le niveau d'expression prédit, à partir d'une nouvelle séquence, et celui mesurée expérimentalement, affecté au génome de référence. De cette manière, il est possible de calculer un score d'impact sur le niveau d'expression de mutations génomiques (allèle alternative) par rapport à la séquence de référence (allèle de référence).

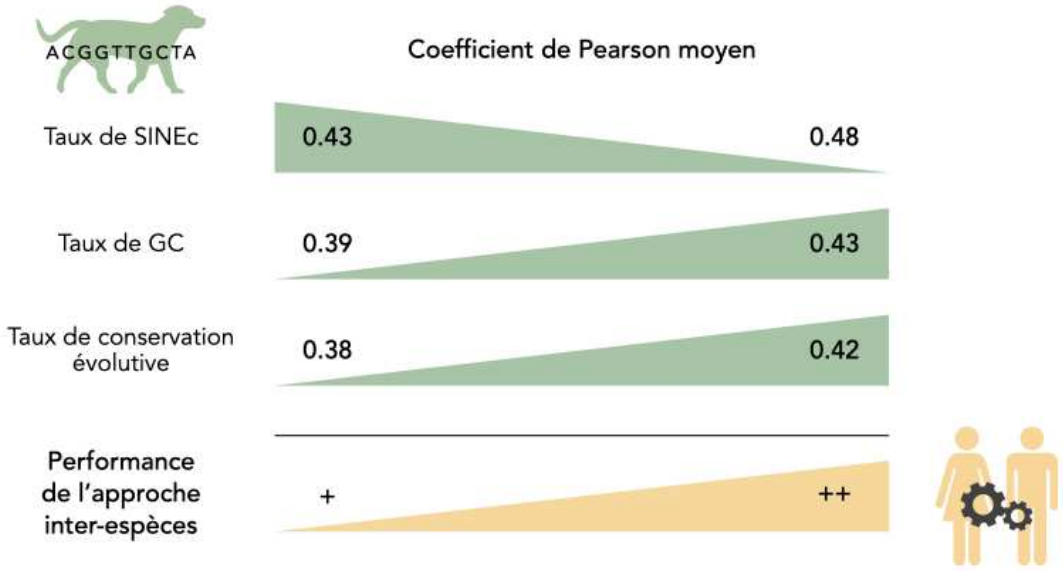

Figure 3.9 – Impact des caractéristiques génomiques sur la performance de l'approche inter-espèces. Lorsque les séquences canines contiennent un taux de SINECs élevé, le modèle de prédiction humain indique de plus faibles performances qu'avec des séquences canines à plus faible taux de SINECs (coefficient de Pearson de 0,43 et 0,48 respectivement). Les séquences canines avec un fort taux de GC ou un taux de conservation évolutive, avec les séquences orthologues chez l'homme, permettent de meilleures performances du modèle de prédiction humain (0,43 et 0,42 respectivement) que les séquences canines avec de faibles taux de ces deux caractéristiques génétiques (0,39 et 0,38 respectivement).

3.3.1 Mutagenèse saturée *in silico*

Parmi les différentes fonctionnalités proposées par l'outil Basenji, la mutagenèse saturée *in silico* (ISM) représente une fonctionnalité essentielle. Le principe consiste à simuler toutes les mutations possibles sur une région génomique choisie afin de prédire l'impact sur l'expression du gène voisin. À chaque position, l'outil prédit l'impact des 3 mutations possibles par rapport au nucléotide de la séquence de référence sur le niveau d'expression grâce au calcul de la différence entre le niveau d'expression induit par les mutations et celui mesuré expérimentalement.

Application avec le modèle de prédiction humain

Cette fonctionnalité étant primordiale dans le cadre de notre projet, nous avons souhaité l'éprouver en utilisant le modèle de prédiction établi chez l'homme dans un premier temps. Le gène *TERT* est très connu dans la littérature scientifique comme étant associé à la cancérologie humaine [40, 41] car des mutations situées dans sa région promotrice induisent la création de nouveaux motifs de liaisons de facteurs de transcription et donc l'activation de son expression. En réalisant l'ISM de cette région du génome humain, nous avons obtenu une prédiction d'impact induit par toutes les mutations possibles au sein de cette région.

Illustrée par la figure 3.10, la fonctionnalité d'ISM de Basenji appliquée à partir du modèle de prédiction humain permet d'identifier des mutations provoquant une dérégulation de l'expression génique dans le sens où celles prédites dans le promoteur de *TERT* ont un impact fonctionnel prouvé expérimentalement dans la littérature.

Utilisation pour l'approche d'oncologie comparée

Dans le cadre de ce travail, nous nous sommes intéressés aux mutations pouvant survenir dans les promoteurs des gènes impliqués en cancérologie. Nous avons donc conduit cette étude d'ISM à partir du panel de gènes orthologues précédemment constitué, en ciblant les prédictions des 1024 nucléotides en amont des TSS de ces gènes. Nous avons réalisé un balayage (scan) mutationnel complet qui mesure l'impact de toutes les mutations possibles à chaque site de la séquence des promoteurs des 1317 gènes de cancer

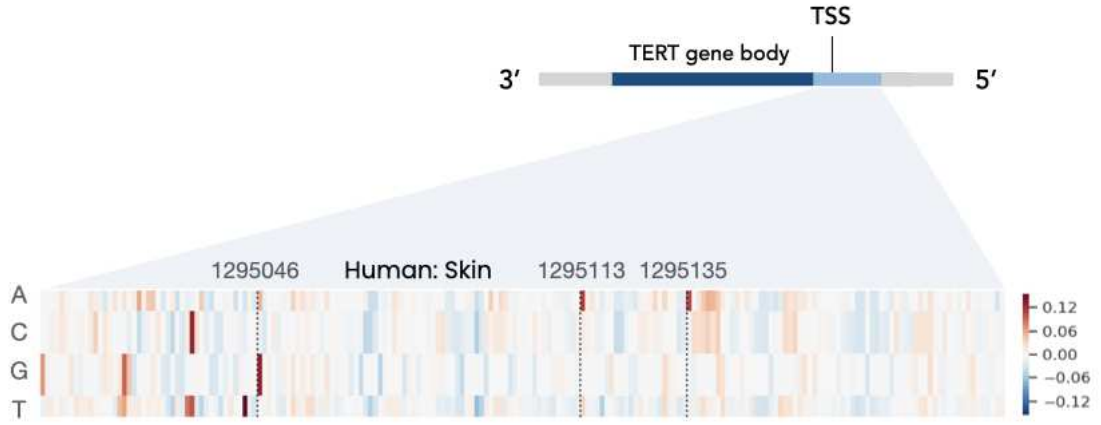

Figure 3.10 – Mutagenèse saturée de *TERT*. Représentation sous forme de heatmap de l'ISM appliquée à la région promotrice (200 nucléotides) du gène *TERT* (chr5 :1253167-1295068), dans un des tissus utilisés pour entraîner le modèle de prédiction d'expression des gènes humains (skin). Chaque colonne représente une des 200 positions et les couleurs rouge et bleu correspondent respectivement à une augmentation et une diminution de l'expression du gène *TERT* lors du changement de nucléotide (en ligne) par rapport à l'allèle de référence. Les 3 mutations aux positions 1 295 046, 1 295 113 et 1 295 135 ont été validées expérimentalement dans la littérature

orthologues correspondant à un total de plus de 1 300 000 sites. L'analyse de ces positions génomiques est réalisée dans l'ensemble des 19 tissus identifiés comme communs entre l'homme et le chien.

Détection des prédictions impactantes

Suite à l'étape de ISM appliquée aux promoteurs de notre panel de gènes orthologues entre l'homme et le chien dans les tissus communs identifiés, nous obtenons les prédictions d'impact sur le niveau d'expression associées. Il s'est ensuite agi d'exploiter ces résultats. Dans cette optique, nous avons conçu le script outliers.py (https://github.com/ckergal/BLIMP/blob/main/bin/outliers.py) permettant d'extraire les mutations prédites commes les plus impactantes sur le niveau d'expression des gènes mais aussi de proposer une mise en forme simplifiant l'exploitation et la lecture de celles-ci.

Afin de détecter les mutations prédites comme étant les plus impactantes sur le niveau d'expression des gènes, nous avons fait le choix de les considérer comme des valeurs extrêmes (ou *outliers*) au sein de leur ensemble d'appartenance. Ainsi, nous considérons

des ensembles de 4096 valeurs représentant les prédictions relatives au nucléotide de référence et les 3 mutations possibles des 1024 positions nucléotidiques du promoteur d'un gène donné et dans un tissu donné. Dans le cas du nucléotide de référence, la prédiction d'impact sur le niveau d'expression est systématiquement de 0. Nous considérons une valeur comme extrême lorsque celle-ci, pour le cas de la distribution normale, se situe à une distance supérieure ou égale à 8 écarts-types de la moyenne.

3.3.2 Analyses des mutations prédites comme impactantes

À partir des étapes de mutagenèses saturées du panel de gènes orthologues, réalisées depuis les prédictions du modèle humain et du modèle canin, et de l'application de notre script outliers.py aux résultats recueillis, nous avons obtenu deux grands ensembles de données représentant les mutations prédites comme impactantes chez l'homme et chez le chien. Nous disposons en effet de 103 217 (7,6%) et 97 946 (7,2%) mutations prédites comme impliquées dans la dérégulation des gènes chez le chien et chez l'homme respectivement.

Description des mutations

Le total de 201 163 mutations prédites comme impactantes par les modèles de prédictions sont assimilées à plusieurs informations. Parmi celles-ci, nous connaissons le nucléotide de référence de la position impactée et la mutation engendrant une dérégulation des gènes. Nous savons aussi à quel écart-type de la moyenne de leur ensemble se situe chaque mutation, permettant de mesurer l'amplitude de la différence d'expression. Nous connaissons enfin leur distance au TSS du gène impacté, mais aussi spécifiquement dans quel tissu elles apparaissent, parmi les 19 communs entre l'homme et le chien.

La distribution de la distance par rapport au TSS des mutations promotrices impactantes est similaire entre le chien et l'homme. En effet, chez les deux espèces elles apparaissent principalement dans les 250 premiers nucléotides du TSS comme illustré sur la figure 3.11.

Nous avons également souhaité analyser de quelle manière les mutations impactent le niveau d'expression des gènes. La figure 3.12 permet de constater leurs répartitions selon le différentiel d'expression qu'elles engendrent. Chez l'homme comme chez le chien,

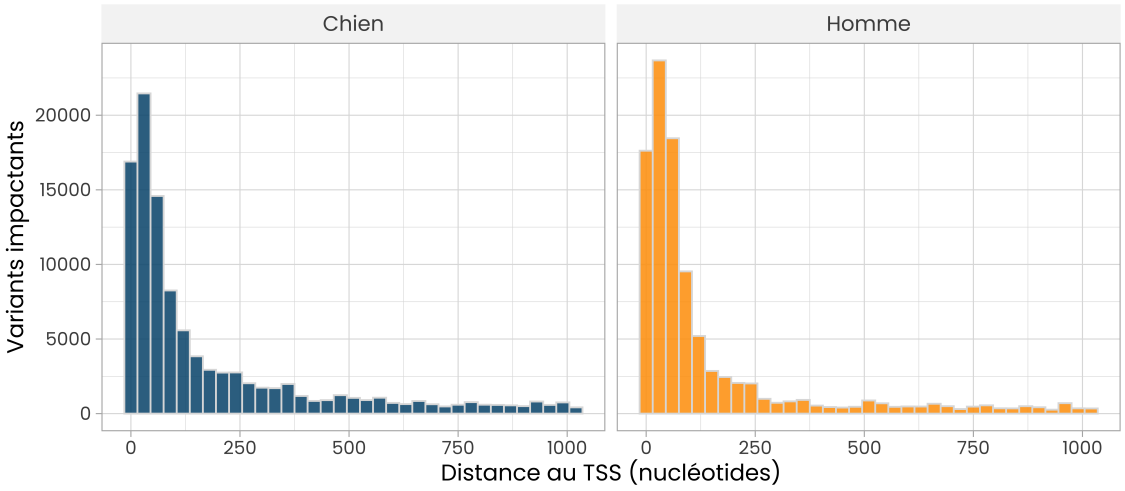

Figure 3.11 – Distance des mutations par rapport au TSS. Répartition des mutations détectées comme impactantes en fonction de leur distance au TSS des gènes, chez le chien (en bleu) et chez l'homme (en jaune).

nous observons davantage de mutations prédites comme engendrant la sur-expression de gènes. En effet, 44% des mutations prédites comme impactantes chez le chien occasionnent une diminution de l'expression. Cette valeur s'élève à 40% chez l'homme.

Nous avons mené une autre analyse pour estimer de quelle manière les mutations prédites comme ayant un impact sur le niveau d'expression des gènes étaient réparties entre les tissus communs entre l'homme et le chien (Fig. 3.13). Nous observons que pour le chien et pour l'homme, tous les tissus possèdent au moins une mutation impactante.

Chez le chien, parmi les 1317 gènes du panel d'intérêt, 1287 possèdent des mutations prédites comme impactantes. C'est dans le pancréas que l'on retrouve le plus de mutations globalement (6036) (Fig. 3.13.A) mais aussi le plus de gènes (1233) ayant au moins une mutation détectée (Fig. 3.13.B). L'utérus est le tissu disposant du moins de mutations (1841) et du moins de gènes (464) ayant au moins une mutation impactante. Chez l'homme, des mutations impactantes ont été détectées dans 1307 gènes parmi 1317 au total. Le tissu dans lequel nous observons le plus de mutations est le colon (5448) et c'est dans la glande surrénale que nous retrouvons le plus de gènes avec au moins une mutation (1206). Le corps calleux est le tissu dans lequel nous trouvons le moins de mutations (2662), mais aussi le moins de gènes disposant d'au moins une mutation (641).

Chez les deux espèces, le nombre moyen de mutations détectées par gènes (Fig.

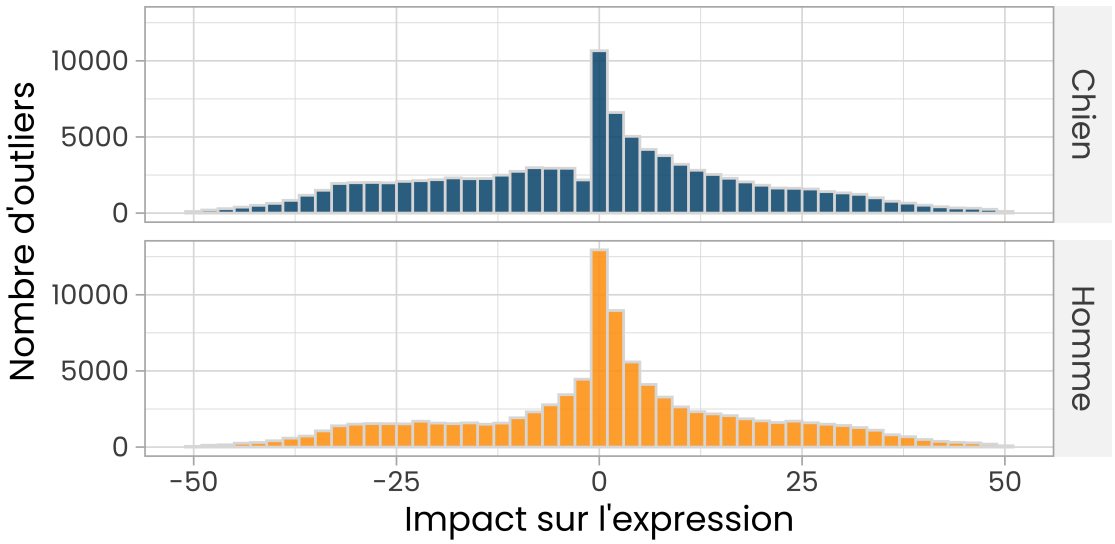

Figure 3.12 – Impacts sur le niveau d'expression. Répartition des différentiels d'expression impliqués par les mutations prédites comme impactantes, en bleu chez le chien et en jaune chez l'homme. Les valeurs d'impact ont été filtrées pour être comprises entre -50 et 50 pour une question de lisibilité du graphique. Chez l'homme, 298 mutations impliquent une sous-expression entre -85 et -50, 998 une sur-expression entre 50 et 150. Chez le chien, 578 mutations impliquent une sous-expression entre -127 et -50 et 1278 une sur-expression entre 50 et 179.

3.13.C) est comparable d'un tissu à l'autre (médiane entre 4 et 6 pour l'ensemble des tissus) et la tendance du nombre de mutations impactantes par tissus (Fig. 3.13.A) est semblable à la tendance du nombre de gènes ayant au moins une mutation, par tissu (Fig. 3.13.B). Ces résultats suggèrent un effet du tissu analysé dans l'étude des mutations ayant un effet sur le niveau d'expression des gènes et proposent des prédictions de régulation distinctes qui contrôlent l'expression des gènes de manière tissu-spécifique. Ces prédictions identifient et hiérarchisent des mutations dans les promoteurs qui impactent l'expression tissulaire spécifique. Par exemple, le gène *MYOD*, une protéine de liaison à l'ADN qui appartient à une famille de facteur de régulation myogéniques (MRF), est synthétisé uniquement dans les cellules musculaires squelettiques et est nécessaire pour la différenciation de leurs précurseurs de fibroblastes indifférenciés [125]. Nous observons que le tissu muscle squelettique possède le plus de mutations (n=14) impactant *MYOD*, présent dans notre panel, comparé aux autres tissus. Ces mutations engendrent toutes une diminution significative de l'expression du transcrit.

Mutations communes au chien et à l'homme

Parmi les simulations de mutations dans les régions promotrices, nous avons recherché les variations communes entre l'homme et le chien prédites avec un impact élevé sur l'expression des gènes. La prédiction précise des promoteurs est fondamentale pour comprendre les modèles d'expression des gènes, la spécificité et le développement des cellules.

Nous émettons l'hypothèse que la cartographie précise entre espèces des variants génétiques associés aux fonctions des promoteurs peut fournir de nouvelles annotations fonctionnelles sur l'expression des gènes. Nous avons ainsi défini les variants communs entre les promoteurs humain et canin selon leurs positions par rapport au TSS. Nous avons utilisé comme proxy les positions identiques qui ont été considérées comme communes entre l'homme et le chien. Cette approche simplifiée repose sur plusieurs éléments. Si les alignements de séquences entre les régions en amont des gènes apparentés sont relativement faibles, généralement $<$60 %, les alignements révèlent en général de nombreux blocs avec un très haut niveau de conservation, qui aident à une prédiction plus précise des promoteurs et des régions renfermant les sites responsables de la régulation de la transcription. De plus, il a été démontré que, contrairement à l'évolution rapide

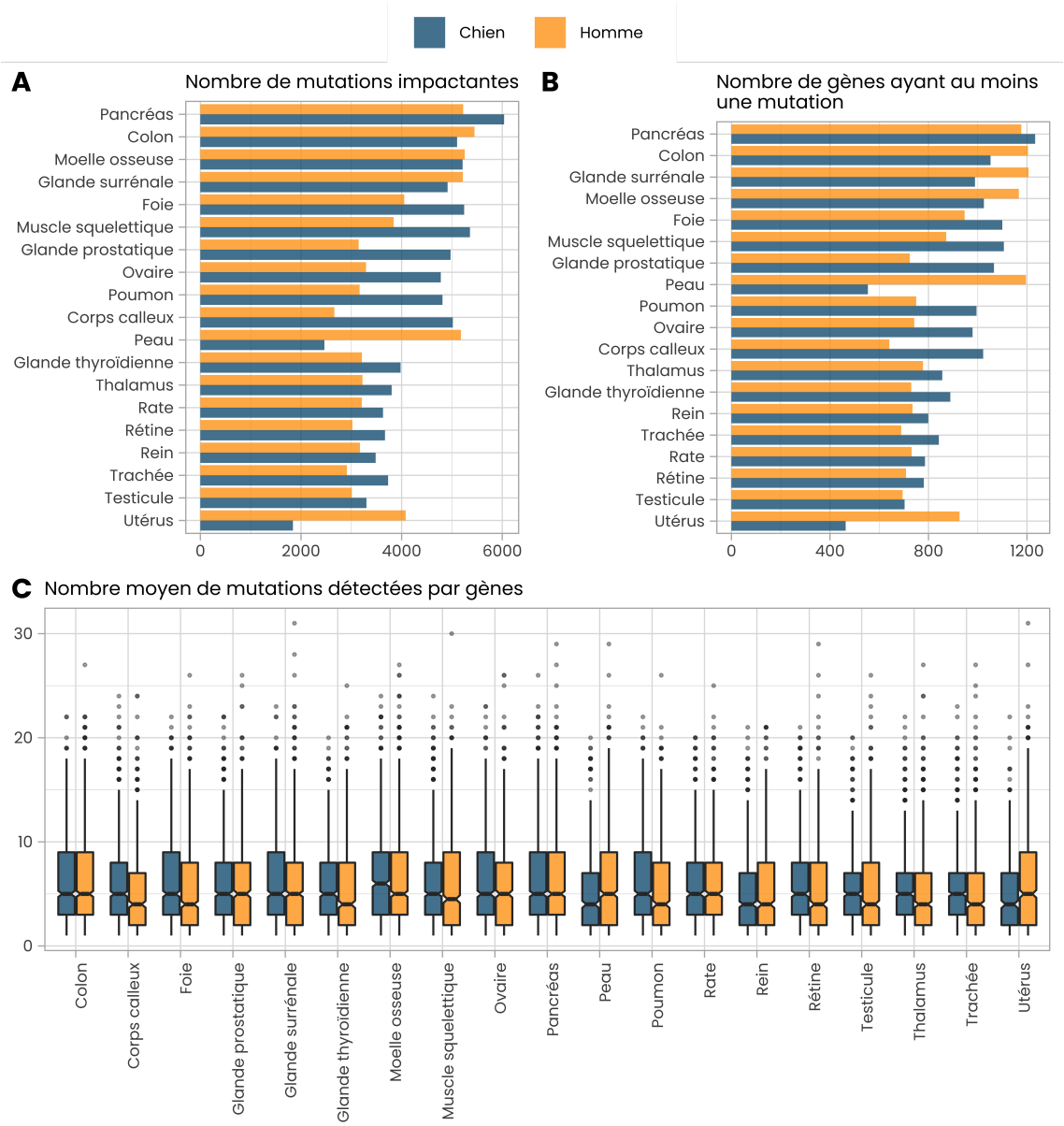

Figure 3.13 – Analyse des mutations en fonction des tissus. Les barres et les boîtes à moustache jaunes concernent les résultats issus de l'utilisation du modèle de prédiction entraîné chez l'homme. Les barres et les boîtes à moustache bleues concernent les résultats établis par les prédictions du modèle canin. **A.** Nombre de mutations impactantes prédites et détectées en fonction du tissu. **B.** Nombre de gènes disposant d'au moins une mutation prédite comme impactante, en fonction du tissu. **C.** Nombre moyen de mutations détectées par gènes, en fonction du tissu.

des enhancers, les promoteurs présentent des profils plus stables au cours de l'évolution [126].

Ainsi, nous considérons comme mutations communes celles qui apparaissent dans le même tissu, le même gène et à la même distance du TSS. Par ces différents filtres, nous avons pu mettre en évidence un ensemble de 1301 mutations communes. Ces 1301 mutations sont présentes dans 19 tissus d'intérêt et affectent 313 promoteurs de gènes, communs entre l'homme et le chien.

La répartition des mutations communes parmi les 19 tissus reflète en partie celle de l'ensemble des mutations prédites comme impactantes (Fig. 3.14). En effet, c'est dans le pancréas et le colon que l'on retrouve le plus de mutations communes et dans l'utérus et le testicule que l'on en retrouve le moins. Cette première description des mutations communes aux deux espèces ne suggère pas l'importance d'un tissu particulier, dans le contexte de notre approche comparative.

Nous avons ensuite exploité la base de données JASPAR CORE [127] qui contient un ensemble de profils organisés et non redondants, dérivés de collections publiées de sites de liaison de facteurs de transcription définis expérimentalement pour les eucaryotes. Nous avons défini un nouvel ensemble de mutations communes à l'homme et le chien en mettant en évidence 476 positions génomiques se référant à un gène et une distance au TSS uniques et ne tenant plus compte de l'information du tissu. L'intersection de 34% des 476 sites prédits dans notre étude avec les coordonnées de TFBS chez l'homme (Hg38) [97] extrait de la base Jaspar est significativement supérieur au hasard ($\chi^2 = 22,18$, p-value $= 2,47^{-6}$). Ces résultats préliminaires suggèrent des sites essentiels dans les sites de TFBS et de manière générale soulignent l'apport des approches de génomique comparative pour prioriser l'implication des variants dans la régulation moléculaire de l'utilisation du promoteur.

Comparaison avec les données Dog10K

La prédisposition des races canines à développer des cancers suggère la présence de variations germinales dans les séquences codantes et régulatrices des gènes. Nous avons utilisé les séquences de génomes entiers des 1929 canidés générées par le consortium Dog10K [11] pour l'analyse du panel de gènes du cancer. Nous avons tout d'abord recherché et extrait les variants génétiques localisés dans les régions promotrices des 1317

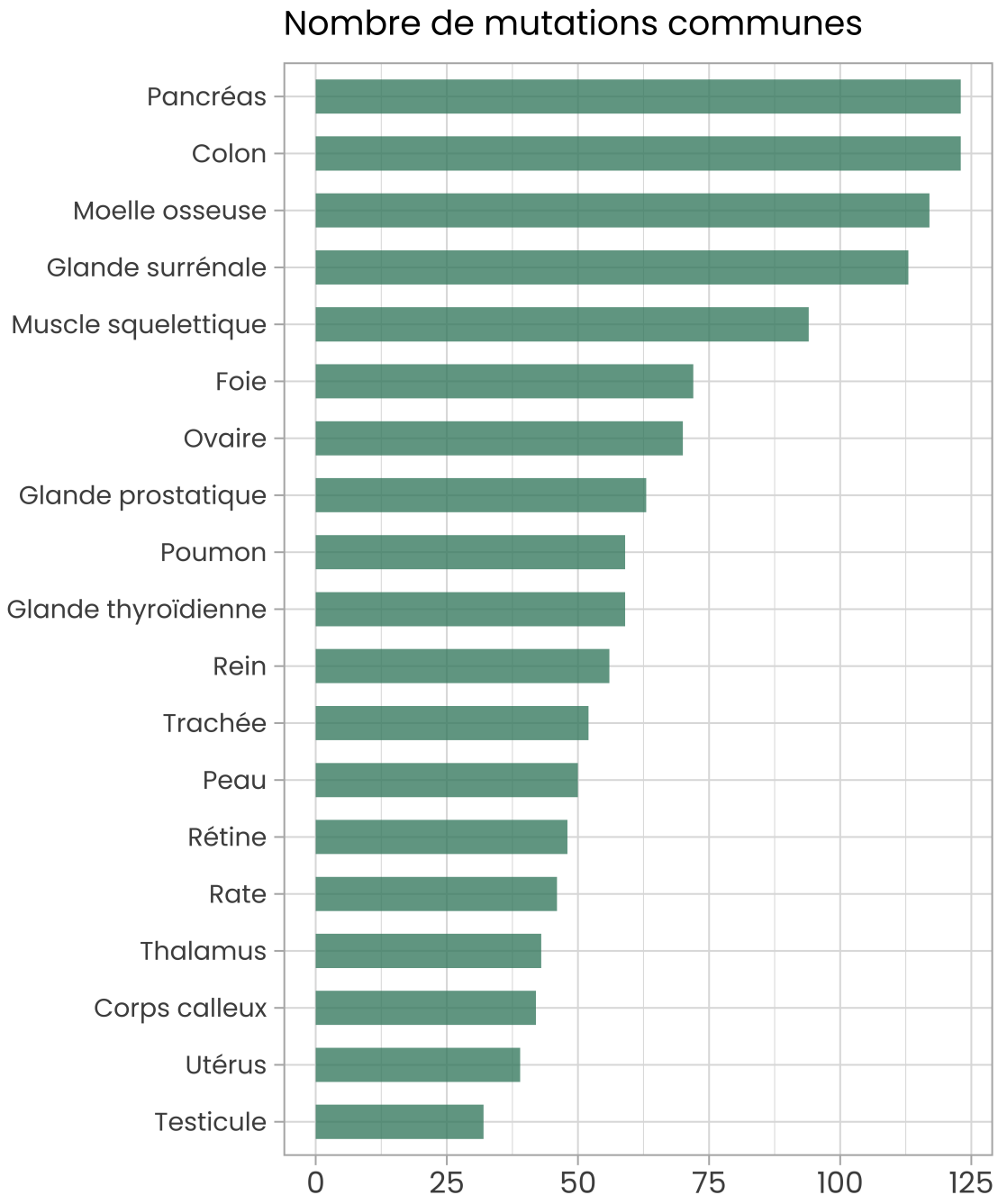

Figure 3.14 – Répartition des mutations communes par tissus. Comptage pour chaque tissu (en ordonnée) du nombre de mutations (en abscisse) prédites comme engendrant une dérégulation de l'expression des gènes.

gènes analysés. Nous avons identifié plus de 15 500 variants localisés dans ces régions promotrices soit un variant tous les 87 nucléotides, un chiffre comparable aux données de variants sur l'ensemble du génome. Le nombre de variants Dog10K, intersectant les régions génomiques de notre étude, spécifiques aux chiens, aux loups et aux chiens de village sont respectivement de 2500, 2200 et 1960.

Par la suite, nous avons extrait les prédictions d'impact de ces variants déterminés avec le modèle de prédiction du niveau d'expression des gènes canins pour l'ensemble des sites des régions promotrices des 1317 gènes. Nous avons défini 3 catégories de variants selon les valeurs d'écart-type à la moyenne, telle que (1) une absence d'impact significatif sur l'expression des gènes pour les valeurs d'écart-type inférieures à 4, (2) un impact défini comme faible à modéré pour les valeurs d'écart-type comprises entre 4 et 8 et (3) un impact défini comme élevé pour les valeurs d'écart-type supérieures à 8.

Ainsi pour l'ensemble des 15 500 variants Dog10K issus des régions promotrices, près de 93% sont prédits avec une absence d'impact significatif, 6,9% sont prédits avec un impact faible à modéré et 1,2 % des variants sont prédits avec un impact élevé. Nous avons également généré des séries (n=1000) de 15 500 variants de manière aléatoire dans les régions promotrices afin d'évaluer la significativité des résultats issus des données réelles. Ainsi, le nombre de variants avec un impact élevé n'est pas significativement différent qu'attendu par hasard (pval = 0,2) pour l'ensemble des populations de canidés de la base Dog10K. Cependant, en différenciant les populations canines, loups et chien de village, le nombre de variants prédits avec un impact élevé est significativement bas chez les populations de loups et de chien village (pval = 0,04 et 0,06). Ces résultats suggèrent que ces variants impactant l'expression des gènes sont contre-sélectionnés chez les loups et les chiens village.

Parmi les régions promotrices ayant des variants prédits comme fortement impactants, nous identifions les gènes *NOTCH1*, *PTPN6* ou encore *GAB2*, dont les fonctions biologiques sont référencées dans les études en oncologie humaine. Cette étude de l'impact des variants issus de données réelles telles que nous les avons identifiées dans les données provenant du consortium Dog10K, nous permet d'établir une évaluation et une classification des variants des régions régulatrices des gènes qui contribuent aux cancers chez le chien.

À la suite des travaux de JOBIM, les diffrentes analyses menées à partir du chapitre 3.3 "Impact des mutations régulatrices sur le niveau d'expression des gènes" feront

prochainement l'objet d'une publication dédiée.

DISCUSSION

4.1 Résumé des travaux

Au cours de ma thèse, nous nous sommes intéressés à la modélisation du niveau d'expression des gènes canins comme fonction de l'ADN grâce à une approche d'apprentissage profond. Cette démarche s'inscrit dans l'ambition plus globale de la simulation *in silico* de l'impact fonctionnel des mutations survenant dans les régions régulatrices des gènes impliqués en cancérologie.

Dans cette optique, nous avons pu disposer du plus vaste jeu de données de séquençage de type CAGE dans le cadre d'une collaboration avec le consortium DoGA dont l'objectif est d'améliorer l'annotation fonctionnelle du génome du chien. À partir de ces données, nous avons employé l'outil Basenji [103], dont l'algorithme d'entraînement repose sur un réseau de neurones convolutif, pour réaliser deux modèles de prédiction du niveau d'expression des gènes canins. Basés sur les versions d'assemblage canFam3 et canFam4 du génome du chien, ils atteignent des performances, mesurées par une moyenne de coefficients de corrélation de Pearson, de 0,61 et 0,65 respectivement.

Les approches d'apprentissage profond permettent d'utiliser un modèle établi chez une espèce pour l'appliquer à la prédiction à partir de données provenant d'une autre espèce. Étant donnée cette possibilité, nous avons souhaité évaluer le modèle de prédiction spécifique à l'espèce canine en comparant les performances des approches inter-espèces et intra-espèces. Bien que le modèle de prédiction établi chez l'homme permette de connaître le niveau d'expression des gènes dans un grand nombre d'échantillons, nous avons montré que son application aux séquences génomiques canines n'était pertinente que sous certaines conditions. En effet, les séquences génomiques canines dont le taux de GC est faible ou bien celles se trouvant être riches en éléments transposables spécifiques aux génomes carnivores (*e.g.* ~SINEC) par rapport aux régions orthologues chez l'homme voient leurs prédictions dégradées. De manière plus évidente, il en est de même pour les séquences génomiques disposant d'un faible taux de conservation évolutive avec

l'homme. Il est ainsi profitable de réaliser un modèle de prédiction spécifique à notre espèce d'intérêt, sous réserve de la disponibilité des données issues de l'annotation fonctionnelle du génome de celle-ci.

Étant donné l'intérêt du chien en tant que modèle biomédical pour les études portant sur le cancer, nous avons constitué un panel de 1317 gènes orthologues entre l'homme et le chien, connus pour être impliqués dans les processus de tumorigenèse. En focalisant notre étude sur les régions de 1024 nucléotides en amont du TSS de ces gènes, nous avons recherché les mutations impliquant des dérégulations de l'expression. Grâce à une approche bioinformatique de mutagenèse saturée utilisant les modèles de prédictions humain et canin, nous avons identifié 97 946 mutations prédites comme impactantes chez l'homme et 103 217 chez le chien. En croisant ces résultats, nous avons pu déterminer un ensemble de 1301 mutations communes entre les deux espèces, ouvrant la voie à une priorisation effective des approches expérimentales.

Nous avons utilisé la version du projet Dog10K [11] comprenant plus de 30 millions de variants nucléotidiques et structurels provenant de 1929 individus constituant la plus grande population représentative de chiens et de loups. En utilisant le modèle spécifique au chien, nous avons attribué à 6% des variants un impact faible à modéré et à 1% un impact élevé sur les niveaux d'expression des gènes du cancer. Nos résultats montrent également une occurrence plus élevée de ces mutations chez les chiens domestiques que chez les loups. Au total, dans cette étude, nous avons développé un outil pour prédire l'impact des mutations non codantes sur l'expression des gènes canins et l'avons utilisé pour prédire de manière exhaustive les mutations régulatrices de l'expression des gènes du cancer et définir celles qui sont partagées entre les humains et les chiens. Notre outil et nos modèles sont disponibles publiquement sur GitHub (https://github.com/ckergal/BLIMP).

4.2 Apport de nos travaux à la littérature

4.2.1 Modèles de prédictions du niveau d'expression des gènes canins

Les modèles de prédiction du niveau d'expression des gènes canins que nous avons créés sont uniques. Pour les réaliser, nous avons utilisé l'ensemble le plus complet de données issues de séquençage canin de type CAGE [9] existant. Ce jeu de données lui-même est employé pour la première fois à cette intention. En proposant deux versions de modèles de prédictions, nous mettons à disposition de la communauté scientifique une version basée sur la version d'assemblage canfam3 [22] et une autre basée sur la version canfam4 [83], les deux assemblages étant largement utilisés par la communauté étudiant la génétique canine.

Un des principaux enjeux de cette thèse était de démontrer l'intérêt de la création d'un modèle d'apprentissage profond spécifique à la prédiction du niveau d'expression des gènes canins. En effet, notre approche fait l'objet d'un débat récurrent dans la littérature [114, 124]. L'utilisation de modèles de prédictions par une approche inter-espèces présente l'avantage que les modèles en question sont déjà réalisés et utilisables pour une espèce dont l'annotation fonctionnelle du génome est faible. En revanche, nous avons pu bénéficier d'une ressource en données non négligeable [99], permettant d'obtenir des prédictions plus performantes. Les expériences que nous avons conduites en prédisant le niveau d'expression de gènes canins avec les modèles de prédiction humain et canin ont montré de meilleures corrélations et donc de meilleures performances à l'utilisation d'un modèle spécifique à une espèce d'intérêt.

4.2.2 Développement de l'outil BLIMP

À partir de la fonctionnalité de mutagenèse saturée *in silico* proposée avec l'outil Basenji et un script python de notre réalisation, nous avons pu recenser des mutations prédites comme engendrant un changement fonctionnel significatif du niveau d'expression des gènes impliqués dans le processus de tumorigenèse. Nous proposons ainsi une liste détaillée de positions génomiques jouant un rôle majeur dans la (de-)régulation de l'expression des gènes.

Dans une démarche FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables), nous avons également créé le dépôt GitHub BLIMP, mettant ainsi à disposition de la communauté scientifique les modèles de prédiction du niveau d'expression des gènes canins. Nous proposons aussi des indications sur la manière d'utiliser l'outil Basenji pour exploiter ces modèles. Un utilisateur peut alors utiliser les nombreuses fonctionnalités de Basenji appliquées aux modèles de prédiction canins pour étudier les gènes et les tissus inhérents à sa problématique de recherche. Nous mettons également à disposition les différents scripts python que nous avons développés afin de faciliter la mise en forme et l'exploitation des différentes prédictions provenant de l'outil Basenji. Ainsi, nous proposons une nouvelle tabulation des sorties initialement proposés par l'outil en y intégrant des informations supplémentaires comme le poids d'impact de chaque mutation, le tissu ou encore le nom du gène dans lesquels elles interviennent. Le dépôt GitHub BLIMP permet donc de réaliser la détection de mutations impactantes du niveau d'expression sur des gènes et des tissus que nous n'avons pas inclus dans notre étude comparative des gènes cancer orthologues entre l'homme et le chien. Notre approche constitue ainsi une ressource précieuse pour investiguer et caractériser toutes nouvelles mutations identifiées dans des projets de séquençage de génomes entiers.

Notre approche de création de modèles d'apprentissage profond du niveau d'expression des gènes chez le chien peut être envisagée pour d'autres espèces modèles en génétique ou d'intérêt en recherche. Nous proposons alors, via le dépôt GitHub BLIMP, des précisions concernant l'ensemble du processus de création des modèles de prédiction canins. Nous expliquons notamment les différentes étapes nécessaires à la bonne intégration des données CAGE utilisées lors de l'entraînement d'un modèle. Ainsi, nous encourageons et promouvons l'utilisation des approches d'apprentissage profond pour la recherche en génétique.

4.3 Limites de nos travaux

4.3.1 Choix de l'algorithme d'entraînement

Au début de cette thèse, le choix d'utiliser l'outil Basenji était justifié par plusieurs aspects tels que la disponibilité des différents scripts bien documentés qui le composent ou encore le support à son utilisation via le répertoire GitHub associé. Aussi, Basenji

promettait de meilleures performances de prédiction du niveau d'expression des gènes comparativement aux outils de l'époque, ceci grâce à son algorithme reposant sur une architecture de réseau de neurones convolutif incluant des couches dilatées pour (théoriquement) prédire l'impact de variants sur des plus grandes distances (~30kb). Cependant, la recherche en apprentissage profond a évolué très rapidement ces trois dernières années et des nouvelles architectures prometteuses issues du traitement automatique du langage naturel (NLP Natural Language Processing) ont été développées plus récemment. Ainsi, les méthodes appelées transformers, basées sur une architecture impliquant des couches d'attention, permettent de modéliser les interactions encore plus distantes et d'atteindre des performances de prédiction plus élevées que les réseaux convolutifs [128]. En octobre 2021, l'outil Enformer [129] a ainsi été développé en se basant sur la méthode transformers et permettant de mesurer l'impact de mutations pouvant survenir à une distance plus élevée du TSS (~100kb) par rapport à l'outil Basenji. En plus de l'étude des promoteurs proximaux, il permet donc d'analyser des régions régulatrices plus distantes telles que des promoteurs distaux ou des enhancers. Ce type d'approche plus complète pourrait donc être favorisé pour de futurs projets d'apprentissage profond appliqués aux problématiques en génomique et notamment pour cartographier plus précisément les variants associés à des maladies génétiques ou des traits phénotypiques.

4.3.2 Validations expérimentales

Un des objectifs principaux de notre travail consistait à prédire l'impact fonctionnel des mutations survenant dans les régions promotrices des gènes afin de cibler des motifs d'intérêt et prioriser les expériences *in vitro* et/ou *in vivo*. Nous avons construit notre approche autour de l'hypothèse selon laquelle les prédictions réalisées à partir du modèle d'apprentissage profond entraîné via l'algorithme de Basenji permettaient de détecter les mutations entraînant un impact significatif sur le niveau d'expression des gènes. Ce postulat a pu être établi par plusieurs faits. Tout d'abord, nous avons vu qu'un modèle de prédiction élaboré grâce à l'outil Basenji permet de prédire efficacement le niveau d'expression de séquences génomiques non incluses dans l'entraînement de celui-ci. Ainsi, la présence d'un variant dans une séquence peut être considérée de la même manière car l'outil mesure l'impact de ces variations comme étant la différence d'expression prédites entre l'allèle de référence et les allèles variants considérés. De plus, avec l'exemple des mutations prédites comme impactantes dans le promoteur du gène *TERT* [41] chez

l'homme, nous avons pu confirmer des positions génomiques établies par approches expérimentales comme impliquant une activation de l'expression du gène. Au-delà du gène *TERT*, un nombre croissant de variations situées dans des régions régulatrices ont été validées fonctionnellement chez l'homme dont certaines ont servi lors d'un challenge CAGI (Critical Assessment of Genome Interpretation, http://www.genomeinterpretation.org/cagi5-regulation-saturation.html) dédié spécifiquement à la prédiction des effets de mutations de type SNV associées à des maladies et localisés dans des promoteurs et enhancers. D'autres ressources comme la base de données ClinVar [130] propose aussi des annotations fonctionnelles de variants non codants selon une classification clinique précise : (probablement) bénins, (probablement) pathogène, non-connu (ou VUS Variant of Unknown Significance) définie selon plusieurs analyses complémentaires (WGS, études in vitro, études familiales, ...). Des résultats préliminaires (Fig. 4.1) montrent que les prédictions d'impact des mutations promotrices dans notre panel de gènes de cancer selon le modèle de prédiction humain sont associées à des classifications ClinVar plus impactantes renforçant ainsi l'intérêt de développer ces analyses de mutagenèse *in silico*.

En revanche, nous ne disposons pas d'exemples bien établis de mutations survenant dans les promoteurs de gènes canins avec un impact fonctionnel mesuré expérimentalement. Par exemple, la base de données OMIA (Online Mendelian Inheritance in Animals) [100] recense 474 variants probablement responsables de maladies génétiques ou de traits phénotypiques chez le chien mais la quasi totalité de ces mutations sont localisées à l'intérieur de la phase codante de gènes. Le développement de ressources similaires pour des mutations non codantes permettraient d'affiner la stratégie d'identification de variations régulatrices orthologues entre l'homme et le chien, de renforcer la portée de notre approche et ainsi d'encourager le recours aux modèles de prédiction que nous avons établis pour l'étude et la compréhension du modèle canin.

4.3.3 Approches et limites de l'oncologie comparée

L'utilisation de cancers canins naturels comme modèles de cancers humains est une approche originale de nos travaux. Comme nous l'avons décrit, les cancers chez le chien se développent naturellement, tout comme chez l'homme, et les chiens souffrent des mêmes types de cancers que les humains, tels que les mélanomes muqueux, le cancer du sein et les lymphomes par exemple [62]. Bien que les chiens partagent de nombreux types de cancer avec les humains, les cancers sont souvent uniques et les corrélations

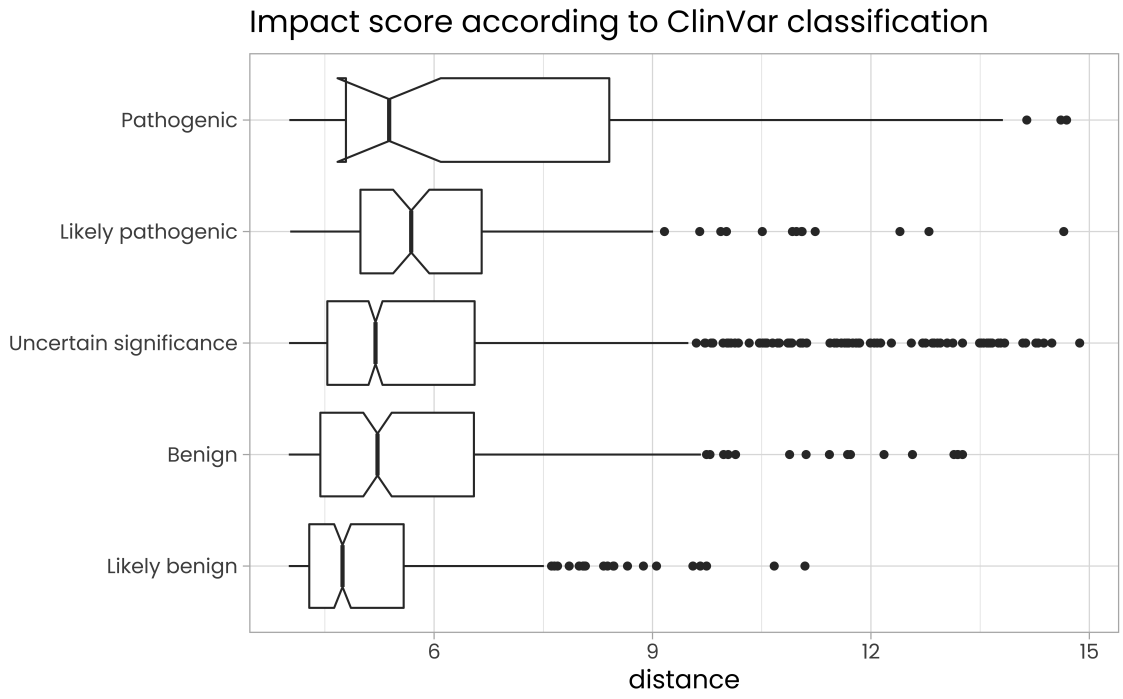

Figure 4.1 – Score d'impact Basenji humain des variants chevauchant des mutations ClinVar localisées dans les promoteurs des gènes humains du panel de cancers. Les variants sont regroupées selon leur classification ClinVar (pour l'illustration, seules les principales classes sont représentées excluant les classes à faible effectif *e.g.* "Conflicting interpretations of pathogenicity").

inter-espèces ne peuvent pas être établies de manières systématiques. Si il est tentant de dresser des comparaisons strictes entre les chiens et les humains lorsqu'il s'agit de cancers, il est important d'éviter les généralisations, notamment en matière de capacité de détection précoce de ces pathologies. En effet, chez l'homme, la connaissance de risques de cancer spécifiques conduit à des approches de dépistage précoce du cancer avec l'objectif de réduire la morbidité et la mortalité. Chez le chien, il existe des races fortement prédisposées à certains cancers, tels que le mélanome muqueux oral chez le caniche, les gliomes chez les races brachiocéphalique (face courte et aplaties), incluant les boxers, bulldogs et certains terriers ou encore l'ostéosarcome des races de très grande taille comme le lévrier irlandais ou le grand danois. Chez ces races de chien, il est documenté qu'elles sont porteuses de mutations génétiques germinales prédisposantes. Il reste cependant à développer et étendre le dépistage spécifique des races chez les chiens asymptomatiques, ce qui n'est pas encore devenu la norme de soins. Si la valeur du modèle canin en oncologie est de plus en plus reconnue [13], il reste à mener de nombreuses études notamment pour identifier les gènes associés au cancer, rechercher et démontrer les facteurs de risque environnementaux, analyser finement la biologie et la progression des tumeurs, et peut-être surtout l'évaluation des nouvelles thérapies contre le cancer.

4.4 Perspectives

La liste des mutations que nous avons identifiées comme engendrant un impact sur le niveau d'expression des gènes canins suggère une valorisation importante. Nous avons pu, dans un premier temps, obtenir des résultats intéressants pour lesquels la richesse des informations peut être exploitée de manière plus approfondie. L'avantage de notre méthode est que l'impact des mutations peut être décrit selon le tissu considéré. L'analyse de l'influence des tissus sur la détection des mutations prédites comme impactantes pourrait aussi être confirmée par le biais de tests statistiques. Nous pouvons également réaliser une étude plus fine des gènes en fonction de la rareté ou, au contraire, de l'abondance des mutations prédites comme impactantes composant leurs promoteurs. Pour reprendre l'exemple du promoteur du gène *TERT* chez l'Homme, les mutations validées expérimentalement sont prédites comme étant impactantes dans 100% des tissus utilisés dans le modèle pour la mutation C250T et dans 73% pour la mutation C228T,

suggérant l'importance de ce critère pour hiérarchiser les mutations prioritaires. Des recherches bibliographiques permettraient d'identifier plus précisément le rôle biologique des gènes singuliers. Enfin, nous pouvons solliciter la littérature pour caractériser les mutations que nous avons identifiées comme communes entre l'homme et les chiens afin d'en connaître les processus fonctionnels. Ces nouvelles informations viendraient ainsi enrichir les indications menant à la priorisation des applications expérimentales.

Les méthodes d'apprentissage profond suscitent un intérêt exceptionnel auprès de la communauté scientifique, notamment de par la diversité de leurs champs d'application et du déluge de données de séquençages disponibles, à la fois génomique, transcriptomique et épigénétique. La discipline connaît donc des améliorations rapides et voit apparaître des approches visant à créer des modèles de prédictions toujours plus performants. Dans ce contexte, les travaux de cette thèse peuvent être poursuivis afin de modéliser de manière plus fidèle le lien entre l'ADN et l'expression des gènes. En plus de la modélisation du taux de transcription comme présenté dans notre travail, la quantité d'ARN dans un tissu ou une cellule dépend aussi de sa vitesse de dégradation, il pourrait être intéressant de prendre en compte afin d'affiner la quantification des ARNs et donc la prédiction d'impact des mutations. Une autre possibilité résiderait dans l'emploi d'un algorithme différent du réseau de neurones convolutif proposé par Basenji ou encore d'envisager des approches complémentaires, comme l'apprentissage par transfert. Cette méthode permettrait par exemple d'exploiter le modèle de prédiction établi chez l'homme comprenant la modélisation du niveau d'expression dans un grand nombre d'échantillons biologiques. Ce modèle servirait alors de support à l'apprentissage des profils d'expression canins, permettant théoriquement d'obtenir de meilleures performances de prédiction [61].

Il est également envisageable de prolonger les travaux de cette thèse en intégrant les données CAGE issues de nouveaux séquençages d'échantillons biologiques canins. L'annotation fonctionnelle du génome des mammifères est un pan considérable de la recherche en génomique et de nouvelles données pourraient venir compléter la collection de séquençages de type CAGE déjà établie par le consortium DoGA. Ce dernier dispose également d'un ensemble conséquent de données séquencées par la technologie STRT [102]. Au cours de cette thèse, nous avons pu tester l'intégration de ces données pour réaliser un modèle de prédiction du niveau d'expression des gènes canins, mais nous n'avons pu constater que de faibles performances de prédiction. En revanche, en réalisant une architecture d'algorithme spécifique, le potentiel représenté par ces données pourrait

être exploité. Cette opportunité permettrait d'étoffer considérablement les applications du modèle de prédiction du niveau d'expression des gènes canins.

L'ensemble de ces travaux permet d'appuyer la valeur des approches d'apprentissage automatique pour la recherche en génomique et renforce l'intérêt du chien comme modèle spontané de nombreuses maladies humaines.

BIBLIOGRAPHIE

1. LIBBRECHT, M. W. & NOBLE, W. S., Machine learning applications in genetics and genomics, en, *Nature Reviews Genetics* **16,** 321-332, ISSN : 1471-0056, 1471-0064, http://www.nature.com/articles/nrg3920 (2022) (juin 2015).
2. ARONSON, S. J. & REHM, H. L., Building the foundation for genomics in precision medicine, en, *Nature* **526,** 336-342, ISSN : 0028-0836, 1476-4687, http://www.nature.com/articles/nature15816 (2022) (oct. 2015).
3. YANDELL, M. D. & MAJOROS, W. H., Genomics and natural language processing, en, *Nature Reviews Genetics* **3,** 601-610, ISSN : 1471-0056, 1471-0064, http://www.nature.com/articles/nrg861 (2022) (août 2002).
4. ZHOU, J. & TROYANSKAYA, O. G., Predicting effects of noncoding variants with deep learning–based sequence model, en, *Nature Methods* **12,** 931-934, ISSN : 1548-7091, 1548-7105, http://www.nature.com/articles/nmeth.3547 (2022) (oct. 2015).
5. *Fondation ARC pour la recherche sur le cancer | Fondation ARC* https://www.fondation-arc.org/ (2022).
6. HANAHAN, D. & WEINBERG, R. A., Hallmarks of Cancer : The Next Generation, en, *Cell* **144,** 646-674, ISSN : 0092-8674, https://www.sciencedirect.com/science/article/pii/S0092867411001279 (2022) (mars 2011).
7. ZOU, J., HUSS, M., ABID, A., MOHAMMADI, P., TORKAMANI, A. & TELENTI, A., A primer on deep learning in genomics, en, *Nature Genetics* **51,** 12-18, ISSN : 1061-4036, 1546-1718, http://www.nature.com/articles/s41588-018-0295-5 (2021) (jan. 2019).
8. The ENCODE (ENCyclopedia Of DNA Elements) Project, en, *Science* **306,** 636-640, ISSN : 0036-8075, 1095-9203, https://www.science.org/doi/10.1126/science.1105136 (2022) (oct. 2004).
9. THE FANTOM CONSORTIUM AND THE RIKEN PMI AND CLST (DGT), A promoter-level mammalian expression atlas, en, *Nature* **507,** 462-470, ISSN : 0028-0836, 1476-4687, http://www.nature.com/articles/nature13182 (2022) (mars 2014).
10. LONSDALE, J. *et al.,* The Genotype-Tissue Expression (GTEx) project, en, *Nature Genetics* **45,** 580-585, ISSN : 1061-4036, 1546-1718, http://www.nature.com/articles/ng.2653 (2022) (juin 2013).
11. WANG, G.-D., LARSON, G., KIDD, J. M., VONHOLDT, B. M., OSTRANDER, E. A. & ZHANG, Y.-P., Dog10K : the International Consortium of Canine Genome Sequencing, en, *National Science Review* **6,** 611-613, ISSN : 2095-5138, 2053-714X, https://academic.oup.com/nsr/article/6/4/611/5505853 (2022) (juill. 2019).
12. HARDISON, R. C., Comparative Genomics, en, *PLoS Biology* **1,** e58, ISSN : 1545-7885, https://dx.plos.org/10.1371/journal.pbio.0000058 (2022) (nov. 2003).

13. Paoloni, M. C. & Khanna, C., Comparative oncology today, eng, *The Veterinary Clinics of North America. Small Animal Practice* **37,** 1023-1032, v, issn : 0195-5616 (nov. 2007).

14. Sanger, F. *et al.,* Nucleotide sequence of bacteriophage phi X174 DNA, eng, *Nature* **265,** 687-695, issn : 0028-0836 (fév. 1977).

15. International Human Genome Sequencing Consortium *et al.,* Initial sequencing and analysis of the human genome, en, *Nature* **409,** 860-921, issn : 0028-0836, 1476-4687, https://www.nature.com/articles/35057062 (2022) (fév. 2001).

16. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, en, *Nature* **431,** 931-945, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature03001 (2022) (oct. 2004).

17. Sanger, F., Determination of Nucleotide Sequences in DNA, en, *Science* **214,** 1205-1210, issn : 0036-8075, 1095-9203, https://www.science.org/doi/10.1126/science.7302589 (2022) (déc. 1981).

18. Nurk, S. *et al.,* The complete sequence of a human genome, en, *Science* **376,** 44-53, issn : 0036-8075, 1095-9203, https://www.science.org/doi/10.1126/science.abj6987 (2022) (avr. 2022).

19. Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A. & Jorde, L. B., Human Population Genetic Structure and Inference of Group Membership, en, *The American Journal of Human Genetics* **72,** 578-589, issn : 00029297, https://linkinghub.elsevier.com/retrieve/pii/S0002929707605746 (2022) (mars 2003).

20. Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, en, *Nature* **420,** 520-562, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature01262 (2022) (déc. 2002).

21. Rat Genome Sequencing Project Consortium *et al.,* Genome sequence of the Brown Norway rat yields insights into mammalian evolution, en, *Nature* **428,** 493-521, issn : 0028-0836, 1476-4687, https://www.nature.com/articles/nature02426 (2022) (avr. 2004).

22. Broad Sequencing Platform members *et al.,* Genome sequence, comparative analysis and haplotype structure of the domestic dog, en, *Nature* **438,** 803-819, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature04338 (2022) (déc. 2005).

23. The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome, en, *Nature* **437,** 69-87, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature04072 (2022) (sept. 2005).

24. Schuster, S. C., Next-generation sequencing transforms today's biology, en, *Nature Methods* **5,** 16-18, issn : 1548-7091, 1548-7105, http://www.nature.com/articles/nmeth1156 (2022) (jan. 2008).

25. Schadt, E. E., Turner, S. & Kasarskis, A., A window into third-generation sequencing, en, *Human Molecular Genetics* **19,** R227-R240, issn : 0964-6906, 1460-2083, https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq416 (2022) (oct. 2010).

26. The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, en, *Nature* **467,** 1061-1073, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature09534 (2022) (oct. 2010).

27. The 1000 Genomes Project Consortium *et al.,* A global reference for human genetic variation, en, *Nature* **526,** 68-74, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/nature15393 (2022) (oct. 2015).

28. *All of Us Research Program | National Institutes of Health (NIH)* en, jan. 2020, https://allofus.nih.gov/future-health-begins-all-us (2022).

29. Sudlow, C. *et al.,* UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age, en, *PLOS Medicine* **12,** e1001779, issn : 1549-1676, https://dx.plos.org/10.1371/journal.pmed.1001779 (2022) (mars 2015).

30. Kurki, M. I. *et al., FinnGen : Unique genetic insights from combining isolated population and national health register data* en, preprint (Genetic et Genomic Medicine, mars 2022), http://medrxiv.org/lookup/doi/10.1101/2022.03.03.22271360 (2022).

31. GenomeAsia100K Consortium *et al.,* The GenomeAsia 100K Project enables genetic discoveries across Asia, en, *Nature* **576,** 106-111, issn : 0028-0836, 1476-4687, https://www.nature.com/articles/s41586-019-1793-z (2022) (déc. 2019).

32. Stark, Z. *et al.,* Australian Genomics : A Federated Model for Integrating Genomics into Healthcare, en, *The American Journal of Human Genetics* **105,** 7-14, issn : 00029297, https://linkinghub.elsevier.com/retrieve/pii/S0002929719302289 (2022) (juill. 2019).

33. *Plan France Médecine Génomique 2025* fr-FR, https://pfmg2025.aviesan.fr/ (2022).

34. Wang, Z., Gerstein, M. & Snyder, M., RNA-Seq : a revolutionary tool for transcriptomics, en, *Nature Reviews Genetics* **10,** 57-63, issn : 1471-0056, 1471-0064, http://www.nature.com/articles/nrg2484 (2022) (jan. 2009).

35. Carninci, P. *et al.,* The Transcriptional Landscape of the Mammalian Genome, en, *Science* **309,** 1559-1563, issn : 0036-8075, 1095-9203, https://www.science.org/doi/10.1126/science.1112014 (2022) (sept. 2005).

36. Yu, N. Y.-L. *et al.,* Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium, en, *Nucleic Acids Research* **43,** 6787-6798, issn : 0305-1048, 1362-4962, https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv608 (2022) (août 2015).

37. Esteller, M., Non-coding RNAs in human disease, en, *Nature Reviews Genetics* **12,** 861-874, issn : 1471-0056, 1471-0064, http://www.nature.com/articles/nrg3074 (2022) (déc. 2011).

38. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M. & Mattick, J. S., Non-coding RNAs : regulators of disease : Non-coding RNAs : regulators of disease, en, *The Journal of Pathology* **220,** 126-139, issn : 00223417, https://onlinelibrary.wiley.com/doi/10.1002/path.2638 (2022) (jan. 2010).

39. Claussnitzer, M. *et al., FTO* Obesity Variant Circuitry and Adipocyte Browning in Humans, en, *New England Journal of Medicine* **373,** 895-907, issn : 0028-4793, 1533-4406, http://www.nejm.org/doi/10.1056/NEJMoa1502214 (2022) (sept. 2015).

40. Vinagre, J. *et al.,* Frequency of TERT promoter mutations in human cancers, eng, *Nature Communications* **4,** 2185, issn : 2041-1723 (2013).

41. Heidenreich, B. *et al.,* Telomerase reverse transcriptase promoter mutations in primary cutaneous melanoma, en, *Nature Communications* **5,** 3401, issn : 2041-1723, http://www.nature.com/articles/ncomms4401 (2022) (mai 2014).

42. Colebatch, A. J., Dobrovic, A. & Cooper, W. A., TERT gene : its function and dysregulation in cancer, en, *Journal of Clinical Pathology* **72,** 281-284, issn : 0021-9746, 1472-4146, https://jcp.bmj.com/content/72/4/281 (2022) (avr. 2019).

43. Stein, L., Creating a bioinformatics nation, en, *Nature* **417,** 119-120, issn : 0028-0836, 1476-4687, http://www.nature.com/articles/417119a (2022) (mai 2002).

44. Earl, D. *et al.,* Assemblathon 1 : A competitive assessment of de novo short read assembly methods, en, *Genome Research* **21,** 2224-2241, issn : 1088-9051, http://genome.cshlp.org/lookup/doi/10.1101/gr.126599.111 (2022) (déc. 2011).

45. Lindner, R. & Friedel, C. C., A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq, en, *PLoS ONE* **7** (éd. Salzberg, S. L.) e52403, issn : 1932-6203, https://dx.plos.org/10.1371/journal.pone.0052403 (2022) (déc. 2012).

46. Purcell, S. *et al.,* PLINK : a tool set for whole-genome association and population-based linkage analyses, eng, *American Journal of Human Genetics* **81,** 559-575, issn : 0002-9297 (sept. 2007).

47. Liu, Q., Fang, H., Wang, X., Wang, M., Li, S., Coin, L. J. M., Li, F. & Song, J., DeepGenGrep : a general deep learning-based predictor for multiple genomic signals and regions, en, *Bioinformatics* (éd. Valencia, A.) btac454, issn : 1367-4803, 1460-2059, https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac454/6633307 (2022) (juill. 2022).

48. Umarov, R. K. & Solovyev, V. V., Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks, eng, *PloS One* **12,** e0171410, issn : 1932-6203 (2017).

49. Fernandez-Castillo, E., Barbosa-Santillán, L. I., Falcon-Morales, L. & Sánchez-Escobar, J. J., Deep Splicer : A CNN Model for Splice Site Prediction in Genetic Sequences, eng, *Genes* **13,** 907, issn : 2073-4425 (mai 2022).

50. Wucher, V. *et al.,* FEELnc : a tool for long non-coding RNA annotation and its application to the dog transcriptome, eng, *Nucleic Acids Research* **45,** e57, issn : 1362-4962 (mai 2017).

51. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J., A general framework for estimating the relative pathogenicity of human genetic variants, en, *Nature Genetics* **46,** 310-315, issn : 1061-4036, 1546-1718, http://www.nature.com/articles/ng.2892 (2022) (mars 2014).

52. Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R. & Campbell, C., An integrative approach to predicting the functional effects of non-coding and coding sequence variation, *Bioinformatics* **31,** 1536-1543, issn : 1367-4803, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426838/ (2022) (mai 2015).

53. Moyon, L., Berthelot, C., Louis, A., Nguyen, N. T. T. & Roest Crollius, H., Classification of non-coding variants with high pathogenic impact, eng, *PLoS genetics* **18,** e1010191, issn : 1553-7404 (avr. 2022).

54. *FINSURF - HOME* https://www.finsurf.bio.ens.psl.eu/ (2022).

55. Nehme, E. *et al.,* DeepSTORM3D : dense 3D localization microscopy and PSF design by deep learning, eng, *Nature Methods* **17,** 734-740, issn : 1548-7105 (juill. 2020).

56. Arbelle, A. & Raviv, T. R., *Microscopy cell segmentation via adversarial neural networks* in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, Washington, DC, avr. 2018), 645-648, isbn : 978-1-5386-3636-7, https://ieeexplore.ieee.org/document/8363657/ (2022).

57. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K. & Troyanskaya, O. G., Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, eng, *Nature Genetics* **50,** 1171-1179, issn : 1546-1718 (août 2018).

58. Singh, R., Lanchantin, J., Sekhon, A. & Qi, Y., Attend and Predict : Understanding Gene Regulation by Selective Attention on Chromatin, https://arxiv.org/abs/1708.00339 (2022) (2017).

59. Kelley, D. R., Snoek, J. & Rinn, J. L., Basset : learning the regulatory code of the accessible genome with deep convolutional neural networks, en, *Genome Research* **26,** 990-999, issn : 1088-9051, 1549-5469, http://genome.cshlp.org/lookup/doi/10.1101/gr.200535.115 (2022) (juill. 2016).

60. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J., Deep learning : new computational modelling techniques for genomics, en, *Nature Reviews Genetics* **20,** 389-403, issn : 1471-0056, 1471-0064, http://www.nature.com/articles/s41576-019-0122-6 (2022) (juill. 2019).

61. *Advances in neural information processing systems 27 : 28th Annual Conference on Neural Information Processing Systems 2014 [(NIPS)] ; December 8 - 13, 2014, Montreal, Canada ; [proceedings of the 2014 conference]* eng (éd. Welling, M. & Neural Information Processing Systems Foundation) isbn : 978-1-5108-0041-0 (Curran, Red Hook, NY, 2015).

62. Schiffman, J. D. & Breen, M., Comparative oncology : what dogs and other species can teach us about humans with cancer, eng, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **370,** 20140231, issn : 1471-2970 (juill. 2015).

63. Paoloni, M. & Khanna, C., Translation of new cancer treatments from pet dogs to humans, en, *Nature Reviews Cancer* **8,** 147-156, issn : 1474-175X, 1474-1768, http://www.nature.com/articles/nrc2273 (2022) (fév. 2008).

64. Kirkness, E. F. *et al.,* The dog genome : survey sequencing and comparative analysis, eng, *Science (New York, N.Y.)* **301,** 1898-1903, issn : 1095-9203 (sept. 2003).

65. Derrien, T., Vaysse, A., André, C. & Hitte, C., Annotation of the domestic dog genome sequence : finding the missing genes, en, *Mammalian Genome* **23,** 124-131, issn : 0938-8990, 1432-1777, http://link.springer.com/10.1007/s00335-011-9372-0 (2022) (fév. 2012).

66. Shearin, A. L. & Ostrander, E. A., Leading the way : canine models of genomics and disease, en, *Disease Models & Mechanisms* **3,** 27-34, issn : 1754-8411, 1754-8403, https://journals.biologists.com/dmm/article/3/1-2/27/2387/Leading-the-way-canine-models-of-genomics-and (2022) (jan. 2010).

67. Khanna, C. *et al.,* The dog as a cancer model, en, *Nature Biotechnology* **24,** 1065-1066, issn : 1087-0156, 1546-1696, https://www.nature.com/articles/nbt0906-1065b (2022) (sept. 2006).

68. Frantz, L. A. F. *et al.,* Genomic and archaeological evidence suggest a dual origin of domestic dogs, eng, *Science (New York, N.Y.)* **352,** 1228-1231, issn : 1095-9203 (juin 2016).

69. Botigué, L. R. *et al.,* Ancient European dog genomes reveal continuity since the Early Neolithic, en, *Nature Communications* **8,** 16082, issn : 2041-1723, https://www.nature.com/articles/ncomms16082 (2022) (déc. 2017).

70. Germonpré, M., Lázničková-Galetová, M. & Sablin, M. V., Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic, en, *Journal of Archaeological Science* **39,** 184-202, issn : 03054403, https://linkinghub.elsevier.com/retrieve/pii/S0305440311003499 (2022) (jan. 2012).

71. Ovodov, N. D., Crockford, S. J., Kuzmin, Y. V., Higham, T. F. G., Hodgins, G. W. L. & van der Plicht, J., A 33,000-Year-Old Incipient Dog from the Altai Mountains of Siberia : Evidence of the Earliest Domestication Disrupted by the Last Glacial Maximum, en, *PLoS ONE* **6** (éd. Stepanova, A.) e22821, issn : 1932-6203, https://dx.plos.org/10.1371/journal.pone.0022821 (2022) (juill. 2011).

72. *Fédération Cynologique Internationale* https://www.fci.be/fr/ (2022).

73. Parker, H. G. *et al.,* Genetic Structure of the Purebred Domestic Dog, en, *Science* **304,** 1160-1164, issn : 0036-8075, 1095-9203, https://www.science.org/doi/10.1126/science.1097406 (2022) (mai 2004).

74. Ostrander, E. A., Both Ends of the Leash — The Human Links to Good Dogs with Bad Genes, en, *New England Journal of Medicine* **367,** 636-646, issn : 0028-4793, 1533-4406, http://www.nejm.org/doi/abs/10.1056/NEJMra1204453 (2022) (août 2012).

75. Leroy, G., Phocas, F., Hedan, B., Verrier, E. & Rognon, X., Inbreeding impact on litter size and survival in selected canine breeds, en, *The Veterinary Journal* **203,** 74-78, issn : 10900233, https://linkinghub.elsevier.com/retrieve/pii/S1090023314004559 (2022) (jan. 2015).

76. Bannasch, D. *et al.,* The effect of inbreeding, body size and morphology on health in dog breeds, eng, *Canine Medicine and Genetics* **8,** 12, issn : 2662-9380 (déc. 2021).

77. Bittles, A. H. & Black, M. L., Consanguinity, human evolution, and complex diseases, en, *Proceedings of the National Academy of Sciences* **107,** 1779-1786, issn : 0027-8424, 1091-6490, https://pnas.org/doi/full/10.1073/pnas.0906079106 (2022) (jan. 2010).

78. Giger, U., Sargan, D. R. & McNiel, E. A., Breed-specific hereditary diseases and genetic screening, *Cold Spring Harbor Monograph Series* **44,** 249 (2006).

79. Grall, A. *et al.,* PNPLA1 mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans, eng, *Nature Genetics* **44,** 140-147, issn : 1546-1718 (jan. 2012).

80. Jagannathan, V. *et al.,* Dog10K_Boxer_Tasha_1.0 : A Long-Read Assembly of the Dog Reference Genome, en, *Genes* **12,** 847, issn : 2073-4425, https://www.mdpi.com/2073-4425/12/6/847 (2022) (mai 2021).

81. Edwards, R. J. *et al.,* Chromosome-length genome assembly and structural variations of the primal Basenji dog (Canis lupus familiaris) genome, en, *BMC Genomics* **22,** 188, issn : 1471-2164, https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07493-6 (2022) (déc. 2021).

82. Halo, J. V. *et al.,* Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes, en, *Proceedings of the National Academy of Sciences* **118,** e2016274118, issn : 0027-8424, 1091-6490, https://pnas.org/doi/full/10.1073/pnas.2016274118 (2022) (mars 2021).

83. Wang, C. *et al.,* A novel canine reference genome resolves genomic architecture and uncovers transcript complexity, en, *Communications Biology* **4,** 185, issn : 2399-3642, http://www.nature.com/articles/s42003-021-01698-x (2022) (déc. 2021).

84. Dreger, D. L., Rimbault, M., Davis, B. W., Bhatnagar, A., Parker, H. G. & Ostrander, E. A., Whole genome sequence, SNP chips and pedigree structure : building demographic profiles in domestic dog breeds to optimize genetic trait mapping, en, *Disease Models & Mechanisms,* dmm.027037, issn : 1754-8411, 1754-8403, https://journals.biologists.com/dmm/article/doi/10.1242/dmm.027037/257207/Whole-genome-sequence-SNP-chips-and-pedigree (2022) (jan. 2016).

85. Barrios, N., González-Lagos, C., Dreger, D. L., Parker, H. G., Nourdin-Galindo, G., Hogan, A. N., Gómez, M. A. & Ostrander, E. A., Patagonian sheepdog : Genomic analyses trace the footprints of extinct UK herding dogs to South America, en, *PLOS Genetics* **18** (éd. Frantz, L.) e1010160, issn : 1553-7404, https://dx.plos.org/10.1371/journal.pgen.1010160 (2022) (avr. 2022).

86. JAGANNATHAN, V. *et al.,* A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves, en, *Animal Genetics* **50,** 695-704, ISSN : 0268-9146, 1365-2052, https://onlinelibrary.wiley.com/doi/10.1111/age.12834 (2022) (déc. 2019).

87. PLASSAIS, J. *et al.,* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology, en, *Nature Communications* **10,** 1489, ISSN : 2041-1723, http://www.nature.com/articles/s41467-019-09373-w (2022) (déc. 2019).

88. HOEPPNER, M. P. *et al.,* An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts, en, *PLoS ONE* **9** (éd. CHADWICK, B. P.) e91172, ISSN : 1932-6203, https://dx.plos.org/10.1371/journal.pone.0091172 (2022) (mars 2014).

89. SUTTER, N. B. & OSTRANDER, E. A., Dog star rising : the canine genetic system, en, *Nature Reviews Genetics* **5,** 900-910, ISSN : 1471-0056, 1471-0064, https://www.nature.com/articles/nrg1492 (2022) (déc. 2004).

90. DAVIS, B. W. & OSTRANDER, E. A., Domestic Dogs and Cancer Research : A Breed-Based Genomics Approach, en, *ILAR Journal* **55,** 59-68, ISSN : 1084-2020, https://academic.oup.com/ilarjournal/article-lookup/doi/10.1093/ilar/ilu017 (2022) (jan. 2014).

91. HANAHAN, D. & WEINBERG, R. A., The Hallmarks of Cancer, en, *Cell* **100,** 57-70, ISSN : 00928674, https://linkinghub.elsevier.com/retrieve/pii/S0092867400816839 (2022) (jan. 2000).

92. *Institut Curie - Centre de recherche et traitement du cancer en France* https://curie.fr/ (2022).

93. PROUTEAU, A. *et al.,* Canine Oral Melanoma Genomic and Transcriptomic Study Defines Two Molecular Subgroups with Different Therapeutical Targets, eng, *Cancers* **14,** 276, ISSN : 2072-6694 (jan. 2022).

94. WONG, K. *et al.,* Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma, eng, *Nature Communications* **10,** 353, ISSN : 2041-1723 (jan. 2019).

95. *phyloP Tutorial* http://compgen.cshl.edu/phast/phyloP-tutorial.php (2022).

96. SOLOVYEV, V. V. & SHAHMURADOV, I. A., PromH : Promoters identification using orthologous genomic sequences, eng, *Nucleic Acids Research* **31,** 3540-3545, ISSN : 1362-4962 (juill. 2003).

97. SCHNEIDER, V. A. *et al.,* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, en, *Genome Research* **27,** 849-864, ISSN : 1088-9051, 1549-5469, http://genome.cshlp.org/lookup/doi/10.1101/gr.213611.116 (2022) (mai 2017).

98. Axelsson, E., Webster, M. T., Ratnakumar, A., The LUPA Consortium, Ponting, C. P. & Lindblad-Toh, K., Death of *PRDM9* coincides with stabilization of the recombination landscape in the dog genome, en, *Genome Research* **22,** 51-63, issn : 1088-9051, http://genome.cshlp.org/lookup/doi/10.1101/gr.124123.111 (2022) (jan. 2012).

99. *THE DOG GENOME ANNOTATION (DoGA) PROJECT* https://doggenomeannotation.org (2022).

100. Nicholas, F. W., Online Mendelian Inheritance in Animals (OMIA) : a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals, *Nucleic Acids Research* **31,** 275-277, issn : 13624962, https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg074 (2022) (jan. 2003).

101. Le Béguec, C. *et al.,* Characterisation and functional predictions of canine long non-coding RNAs, en, *Scientific Reports* **8,** 13444, issn : 2045-2322, http://www.nature.com/articles/s41598-018-31770-2 (2022) (déc. 2018).

102. Natarajan, K. N., en, in *Single Cell Methods* (éd. Proserpio, V.) 133-153 (Springer New York, New York, NY, 2019), isbn : 978-1-4939-9239-3 978-1-4939-9240-9, http://link.springer.com/10.1007/978-1-4939-9240-9_9 (2022).

103. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y. & Snoek, J., Sequential regulatory activity prediction across chromosomes with convolutional neural networks, en, *Genome Research* **28,** 739-750, issn : 1088-9051, 1549-5469, http://genome.cshlp.org/lookup/doi/10.1101/gr.227819.117 (2022) (mai 2018).

104. Park, P. J., ChIP–seq : advantages and challenges of a maturing technology, en, *Nature Reviews Genetics* **10,** 669-680, issn : 1471-0056, 1471-0064, http://www.nature.com/articles/nrg2641 (2022) (oct. 2009).

105. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J., ATAC-seq : A Method for Assaying Chromatin Accessibility Genome-Wide, en, *Current Protocols in Molecular Biology* **109,** issn : 1934-3639, 1934-3647, https://onlinelibrary.wiley.com/doi/10.1002/0471142727.mb2129s109 (2022) (jan. 2015).

106. Song, L. & Crawford, G. E., DNase-seq : A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells, en, *Cold Spring Harbor Protocols* **2010,** pdb.prot5384, issn : 1940-3402, 1559-6095, 1559-6095, http://www.cshprotocols.org/lookup/doi/10.1101/pdb.prot5384 (2022) (fév. 2010).

107. Noguchi, S. *et al.,* FANTOM5 CAGE profiles of human and mouse samples, en, *Scientific Data* **4,** 170112, issn : 2052-4463, http://www.nature.com/articles/sdata2017112 (2022) (déc. 2017).

108. Chakravarty, D. *et al.,* OncoKB : A Precision Oncology Knowledge Base, en, *JCO Precision Oncology,* 1-16, issn : 2473-4284, https://ascopubs.org/doi/10.1200/PO.17.00011 (2022) (nov. 2017).

109. Cunningham, F. *et al.,* Ensembl 2022, en, *Nucleic Acids Research* **50,** D988-D995, issn : 0305-1048, 1362-4962, https://academic.oup.com/nar/article/50/D1/D988/6430486 (2022) (jan. 2022).

110. Herrero, J. *et al.,* Ensembl comparative genomics resources, en, *Database* **2016,** baw053, issn : 1758-0463, https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw053 (2022) (2016).

111. *nCounter Panels & Assays* en-US, https://nanostring.com/products/ncounter-assays-panels/ (2022).

112. Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. & Smit, A. F., *RepeatModeler2 : automated genomic discovery of transposable element families* en, preprint (Genomics, nov. 2019), http://biorxiv.org/lookup/doi/10.1101/856591 (2022).

113. Karolchik, D., The UCSC Genome Browser Database, *Nucleic Acids Research* **31,** 51-54, issn : 13624962, https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg129 (2022) (jan. 2003).

114. Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R. C., Kundaje, A. & Mahony, S., Domain-adaptive neural networks improve cross-species prediction of transcription factor binding, en, *Genome Research* **32,** 512-523, issn : 1088-9051, 1549-5469, http://genome.cshlp.org/lookup/doi/10.1101/gr.275394.121 (2022) (mars 2022).

115. Quinlan, A. R. & Hall, I. M., BEDTools : a flexible suite of utilities for comparing genomic features, en, *Bioinformatics* **26,** 841-842, issn : 1460-2059, 1367-4803, https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033 (2022) (mars 2010).

116. Kent, W. J., BLAT–the BLAST-like alignment tool, eng, *Genome Research* **12,** 656-664, issn : 1088-9051 (avr. 2002).

117. O'Shea, K. & Nash, R., An Introduction to Convolutional Neural Networks, https://arxiv.org/abs/1511.08458 (2022) (2015).

118. Abadi, M., *TensorFlow : learning functions at scale* en, in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming* (ACM, Nara Japan, sept. 2016), 1-1, isbn : 978-1-4503-4219-3, https://dl.acm.org/doi/10.1145/2951913.2976746 (2022).

119. *GenOuest bioinformatics – Development, expertise and resources for bioinformatics* en-US, https://www.genouest.org/ (2022).

120. *2021 IEEE Congress on Evolutionary Computation (CEC).* English, OCLC : 1284943445, isbn : 978-1-72818-393-0 978-1-72818-394-7, https://ieeexplore.ieee.org/servlet/opac?punumber=9504682 (2022) (IEEE., 2021).

121. *Neural networks : tricks of the trade* 2nd ed (éd. Montavon, G., Orr, G. & Müller, K.-R.) *Lecture notes in computer science* **7700,** OCLC : ocn828098376, isbn : 978-3-642-35288-1 (Springer, Heidelberg, 2012).

122. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P., Integrative Genomics Viewer (IGV) : high-performance genomics data visualization and exploration, eng, *Briefings in Bioinformatics* **14,** 178-192, issn : 1477-4054 (mars 2013).

123. Chen, L., Fish, A. E. & Capra, J. A., Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties, en, *PLOS Computational Biology* **14** (éd. Ma, J.) e1006484, issn : 1553-7358, https://dx.plos.org/10.1371/journal.pcbi.1006484 (2022) (oct. 2018).

124. Kelley, D. R., Cross-species regulatory sequence activity prediction, en, *PLOS Computational Biology* **16** (éd. Ma, J.) e1008050, issn : 1553-7358, https://dx.plos.org/10.1371/journal.pcbi.1008050 (2022) (juill. 2020).

125. Berkes, C. A. & Tapscott, S. J., MyoD and the transcriptional control of myogenesis, eng, *Seminars in Cell & Developmental Biology* **16,** 585-595, issn : 1084-9521 (oct. 2005).

126. Villar, D. *et al.,* Enhancer Evolution across 20 Mammalian Species, en, *Cell* **160,** 554-566, issn : 00928674, https://linkinghub.elsevier.com/retrieve/pii/S0092867415000070 (2022) (jan. 2015).

127. Castro-Mondragon, J. A. *et al.,* JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles, en, *Nucleic Acids Research* **50,** D165-D173, issn : 0305-1048, 1362-4962, https://academic.oup.com/nar/article/50/D1/D165/6446529 (2022) (jan. 2022).

128. Zaheer, M. *et al.,* Big Bird : Transformers for Longer Sequences, https://arxiv.org/abs/2007.14062 (2020).

129. Avsec, Ž. *et al.,* Effective gene expression prediction from sequence by integrating long-range interactions, en, *Nature Methods* **18,** 1196-1203, issn : 1548-7091, 1548-7105, https://www.nature.com/articles/s41592-021-01252-x (2022) (oct. 2021).

130. Landrum, M. J. *et al.,* ClinVar : improving access to variant interpretations and supporting evidence, eng, *Nucleic Acids Research* **46,** D1062-D1067, issn : 1362-4962 (jan. 2018).

ANNEXES

Annexe I : Communication orale et article JOBIM 2022

Gene Expression prediction using Deep Learning

Camille KERGAL[1], Marie-Dominique GALIBERT[1], Catherine ANDRÉ[1], DoGA CONSORTIUM[¶], Christophe HITTE[1] and Thomas DERRIEN[1]

[1] Univ Rennes, CNRS, IGDR - UMR 6290, F-35000 Rennes, France

[¶] Members of the DoGA Consortium are listed in the Acknowledgements section of the manuscript

Corresponding Author: camille.kergal@univ-rennes1.fr

Abstract

One fundamental question in biology consists in predicting gene expression based on DNA sequence alone. To this aim, deep artificial neural networks have been recently shown to be powerful methods to predict the regulatory activity of a nucleic acid sequence and, in fine*, to assess the impact of regulatory mutations on gene expression. Yet for comparative genomic/transcriptomic studies, it is not clear whether predictive sequence models of gene expression in one species could be easily generalized to other species. The tool Basenji proposes a deep-learning approach using Convolutional Neural Networks (CNN) to predict human gene expression. We used and adapted Basenji to train a dog-specific model of gene expression using a comprehensive set of canine CAGE data (Cap Analysis of Gene Expression) produced by the DoGA consortium (n=116 experiments) corresponding to 37 core tissues. We first showed that the dog model reached similar performance than in humans with high correlations between true expression levels and predicted ones in all samples included in our model (Pearson correlations median =0.66 [min=0.34; max=0.75]). Next, we selected a subset of matched human/dog tissues, to compare the expression predictions of >1,300 orthologous dog cancer genes based on either the dog (within-species predictions) or human (cross-species predictions) models. We show that the within-species model led to higher prediction performance than the cross-species model (Pearson r = 0.65, r= 0.41, respectively). We then evaluated several genomic sequence features that could be associated with the model's effectiveness. We showed that GC content and TE content correlate significantly with the decrease in performance. Finally, given the interest of the dog as a biomedical model for cancer studies, we determined the promoter regions of 1,300 human-dog orthologous cancer genes, from which we will leverage the power of our approach to predict mutations impacting gene expression. Our model is available through github (https://github.com/ckergal/BLIMP) and will be usable via a user-friendly galaxy instance.*

Keywords Deep Learning, gene expression, regulatory variant, comparative oncology, dog

1. Introduction

Deep learning (DL) algorithms have recently attracted a lot of attention in genomics and transcriptomics since they promise to extract biological knowledge from large datasets generated by high throughput sequencing technologies in a data-driven manner [1–3]. Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) DL approaches are particularly successful in modeling regulatory elements, outperforming traditional machine learning methods [4]. Well established DL-based tools are now able to predict the amount of RNA produced in a particular cell or tissue by learning DNA motifs such as Transcription Factor Binding Sites (TFBS) regulating gene expression levels [1]. In human genomics, the Basenji tool suite uses a CNN-based approach to learn to predict thousands of epigenetic and transcriptomic profiles only based on the human genome sequence as input [5]. Given that these computational models are fine-tuned to predict the regulatory

activity of a specific DNA sequence in any tissues used as learning, they offer the benefit to evaluate the impact of non-coding regulatory variants on the level of gene expression in a tissue-specific manner, and more particularly for variants localized in essential DNA motifs learnt by the tools. This has many implications given that most known genomic loci, identified by genome wide-association studies (GWAS) to be associated with a specific trait or disease, are localized outside of protein-coding genes [6] *i.e* in regulatory regions such as promoters and/or enhancers. Using CNN-based approaches thus allows to annotate the impact of non-coding variations that may for instance create novel or alter existing TFBSs and modify gene expression.

It is yet debated whether neural networks used to train models in one species can be easily generalized to other species or whether species-specific neural networks, specifically tuned with hyperparameters (HP) optimization, would provide better predictive power. For instance, Chen *et al.* trained CNN classifiers to identify enhancer sequences based on ChIP-seq data in both human and mouse genomes [4]. Although the authors concluded that their binary classifiers generalized relatively well across species, they recognized that the features learnt by the CNN were difficult to interpret. Conversely, Cochran *et al.* recently showed that cross-species DL-based models (between human and mouse data) for TFBSs identification consistently display lower performance than within-species models [7]. For gene expression predictions, the author of Basenji tried to jointly use the mouse and human genomes to train a multi-task convolutional neural network in order to predict RNA abundance from 6,959 CAGE data (Cap Analysis of Gene Expression) from both species [8]. While this strategy led to higher performance than single training, it required to *a priori* define orthologous sequences of 131 kb between human and mouse genome that can be aligned which only represent ~40-45% of both genomes. In addition, while mouse expression data are abundant and may include specific conditions which are difficult or unethical to obtain in humans, they may not recapitulate gene expression in natural conditions.

As part of our Canine Genetics Team work in comparative oncology between human and dog, we used the Basenji framework to train and optimize a canine-specific model of gene expression, using the most comprehensive set of canine CAGE data produced by the Dog Genome Annotation Project (DoGA) consortium (www.doggenomeannotation.org). We showed that the canine-specific model has comparable performance to the human Basenji model. We then performed gene expression predictions of a canine cancer gene panel and showed that we achieved consistently higher performance when predicted with the canine model (within-species predictions) rather than using the human model (cross-species predictions). The ultimate goal of our work is to use the canine model to predict the impact of non-coding genome variations on gene expression and thus, prioritize regulatory variants associated with diseases and phenotypical traits in the dog species.

2. Materials and Methods

2.1. DoGA consortium CAGE data

In total, we used 116 samples of dog CAGE profiles (Cap Analysis of Gene Expression), which are used by the Basenji algorithm to model transcription start sites (TSS). CAGE sequencing technology allows to quantify the amount of RNA in a given biological sample at a given state and thus can explain promoter usage due to its ability to map capped 5'-ends of transcripts (TSS) with high accuracy [9]. Those 116 canine expression profiles represent 37 distinct tissues and were collected through a collaboration with the Dog Genome Annotation Project (DoGA) consortium. Sequencing of the sample was established according to the FANTOM protocol with a median number of 7.6 millions reads per sample.

2.2. Human dataset description

Human gene expression prediction model was established by Kelley et al. [5]. Training was done on a total set of 5,313 quantitative sequencing assays performed on human samples. Among those samples, 674 DNase-seq and 10 ATAC-seq experiments mapped DNA accessibility, 3,991 CHIP-seq profiled transcription factor binding or histone modifications and 638 corresponded to CAGE expression profiles. In our work, we focused on test predictions from the human model of CAGE samples only, in order to compare with similar sequencing technology as in the dog prediction model. Data processing was described in *Kelley 2020* [8]. In our comparative study between human and dog, we focused on a subset of samples of 19 human CAGE expression profiles matching dog tissues from our dog prediction model. These samples were downloaded from the FANTOM database and are available at:https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/basic/human.tissue.hCAGE/.

2.3. Reference genome assemblies

For human data analyses, the hg38 (GRCh38) genome assembly version was used as the reference assembly. For the dog, the recent canFam4 genome assembly was used as the reference genome for mapping CAGE-seq samples and for all data processing steps of the Basenji algorithm.

2.4. Dog CAGE data Processing

In order to train and use a prediction model with the Basenji tool, data processing is necessary. For the 116 dog CAGE expression data, all BAM files were processed with the *bam_cov.py* script from the Basenji tool suite in order to transform the alignments into normalized BigWig coverage tracks, as described in *Kelley et al.* [8]. The global architecture of Basenji consists in two stages of several convolutional layers with the first stage being composed of seven blocks of convolutional layers framed with batch normalization, GELU activation function and max pooling aggregation function aiming at capturing relevant DNA motifs from the input sequence in each CAGE sample. The second stage is composed of 11 blocks of dilated convolutional layers in order to spread information across the sequence and therefore, to model long range interactions. In our application, we extracted 17,400 non-overlapping sequences of 131,072 bp ($=2^{17}$) from the canFam4 reference genome and further randomly distributed these sequences into three sets: train, valid and test. For valid and test sets, we assigned 10% of sequences to each and 80% for the train set as done in the Human and Mouse model [8]. Each 131 kb sequence is then aggregated into 1,024 * 116 coefficients, each one representing the coverage summed of non-overlapping 128 bp windows in the 116 CAGE expression data. Once the canine model has been trained with the 116 samples, predictions made by the model are evaluated with the remaining 10% of genome sequences (test set) by comparing them with the experimentally measured expression levels. Then, to assess the performance of the model, Pearson correlations are computed between both values (experimental and predicted).

2.5. Cancer gene panel

In order to evaluate the performance and the usability of the DL models in a comparative oncology strategy between human and dog, we established a panel of genes described to be involved in cancer and evolutionary conserved between both species. We first used a set of 1,039 human cancer genes proposed by the OncoKB database (May 2019 update) [10] and kept genes with high-confident canine orthologs according to the Ensembl Compara database (v98), as described in [11]. Then, we also included 672 genes extracted from the pan-cancer gene panel of the Nanostring company, representing orthologous genes between human and dog involved in the immune response to immuno-oncology therapies. Taking the union of these two complementary resources, a final set of 1,317 genes known to be involved in cancers and orthologous between human and dog were used to assess the performance of the model and to predict the impact of regulatory mutations in their promoter sequences (here defined as the 1,024 bp window in 5' of the Transcription Start Site - TSS). From all promoter sequences of these 1,317 genes, several features were computed such as the GC percentage and the

content in transposable elements (TE). For the latter, the genomic positions of SINEs (Short Interspersed Nuclear Elements) as defined by RepeatMasker were extracted from UCSC [12] and intersected with the genomic coordinates of the cancer genes promoters using bedtools intersect version 2.25 [13]. Evolutionary conservation was analyzed by sequence comparison of the 1,024 bp promoter sequences using BLAT v35. BLAT is widely used in comparative genomics optimized for pairwise DNA-sequence alignment and was set as with parameters -minScore=10 -minIdentity=50. BLAT scores, calculated according to aligned length and sequence similarity were used to evaluate sequence alignments.

3. Results

3.1. Predicting canine gene expression using CNN

We used the Basenji framework with default hyperparameters to predict 116 CAGE (Cap Analysis of Gene Expression) data using the reference dog genome (canFam4) as input (see Methods). For each CAGE sample included in the model, we computed the Pearson correlation coefficients between the predicted expression level of the test set sequences and the experimentally measured ones from CAGE data. Using this strategy, the model (thereafter called the "DOG NO OPT" model") achieved a median Pearson correlation across tissues of 0.64 (min=0.30 for pancreas, max=0.74 for neurohypophysis) slightly lower than the human model (median R = 0.66, range=0.18 to 0.76). In order to improve the robustness of the dog model, we optimized two hyperparameters (HP) e.g. L2_scale (as a normalizing HP) and learning rate, as it has been shown that these two HP are to be optimized primarily for smaller datasets. With a grid search strategy, we produced 42 models with the same input data but with different combinations of the two HP values (**Fig 1.A**).

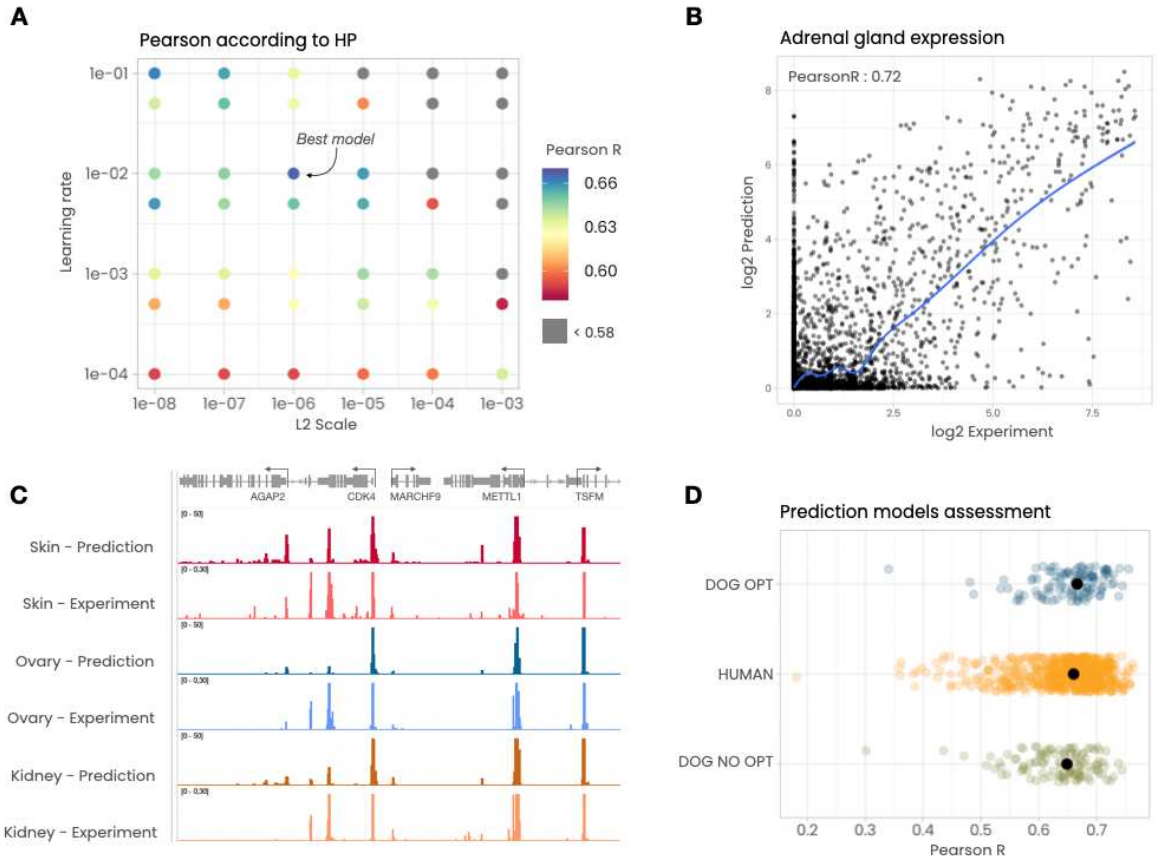

Fig 1. Canine model assessment **A)** Performance of the canine model as measured by Pearson correlations with respect to different values of two hyperparameters (Learning rate and L2 Scale). **B)**

Scatter plot of real/experimental (x-axis) versus predicted (y-axis) log2 sequence expressions in the adrenal gland CAGE sample. The blue line represents the Generalized Additive Model (GAM) regression between experiment and prediction values. **C)** IGV (Integrative Genomics Viewer) [14] representation of a dog canFam4 genomic region encompassing gene track (top layer) and from top to bottom predicted (dark) and experimental (light) CAGE tracks of skin (red), ovary (blue) and kidney (orange) samples. **D)** Pearson coefficients in all tissues for non-optimized dog model (DOG NO OPT), human (HUMAN) and dog model after optimization (DOG OPT). Median Pearson R are represented as black dots.

All resulting models were then ranked according to their median Pearson coefficients and the model whose HP set led to the more accurate predictions was selected. For each sample, we could derive correlations between true and predicted expression values of the tested sequences as illustrated for the adrenal gland corresponding sample in **Fig 1.B**. We also used a complementary strategy to visualize the quality of the model by generating genome-wide tracks of predicted canine CAGE profiles and comparing them with experimentally derived CAGE profiles (**Fig 1.C**). Combining all samples, the final model (called "Optimized model or DOG OPT") achieved a median correlation of 0.66 (min=0.34, max=0.75) (**Fig 1.D**).

We also compared our canine model performance with the human prediction model from the original Basenji tool described by *Kelley D.R. 2020* [8]. In order not to bias the comparison with different sequencing technologies, only values predicted by the human model from CAGE samples were used for the comparison with the canine prediction of CAGE gene expression (**Fig 1.D**). Altogether, the dog optimized led to slightly higher performance than the human model.

3.2. Comparison of within-species versus cross-species models of gene expression

At the transcriptional level, human and dog protein-coding gene expression profiles are globally conserved. It has been shown that orthologous gene expression profiles cluster preferentially by tissues rather than by species [15, 16]. Comparative transcriptomic studies motivated us to investigate whether learned DNA motifs predictive of gene expression in one species could also be used to predict gene expression in another species. In other words, given a test set of canine genes, we computed the performance of the dog model (within-species predictions) with respect to the human model (cross-species predictions) in order to predict their expression levels in matched tissues between both species. We used 1,317 canine genes conserved in human (see Methods) and for each similar tissue, we measured the performance of both the dog and the human model to predict their expression levels **(Fig 2.A)**.

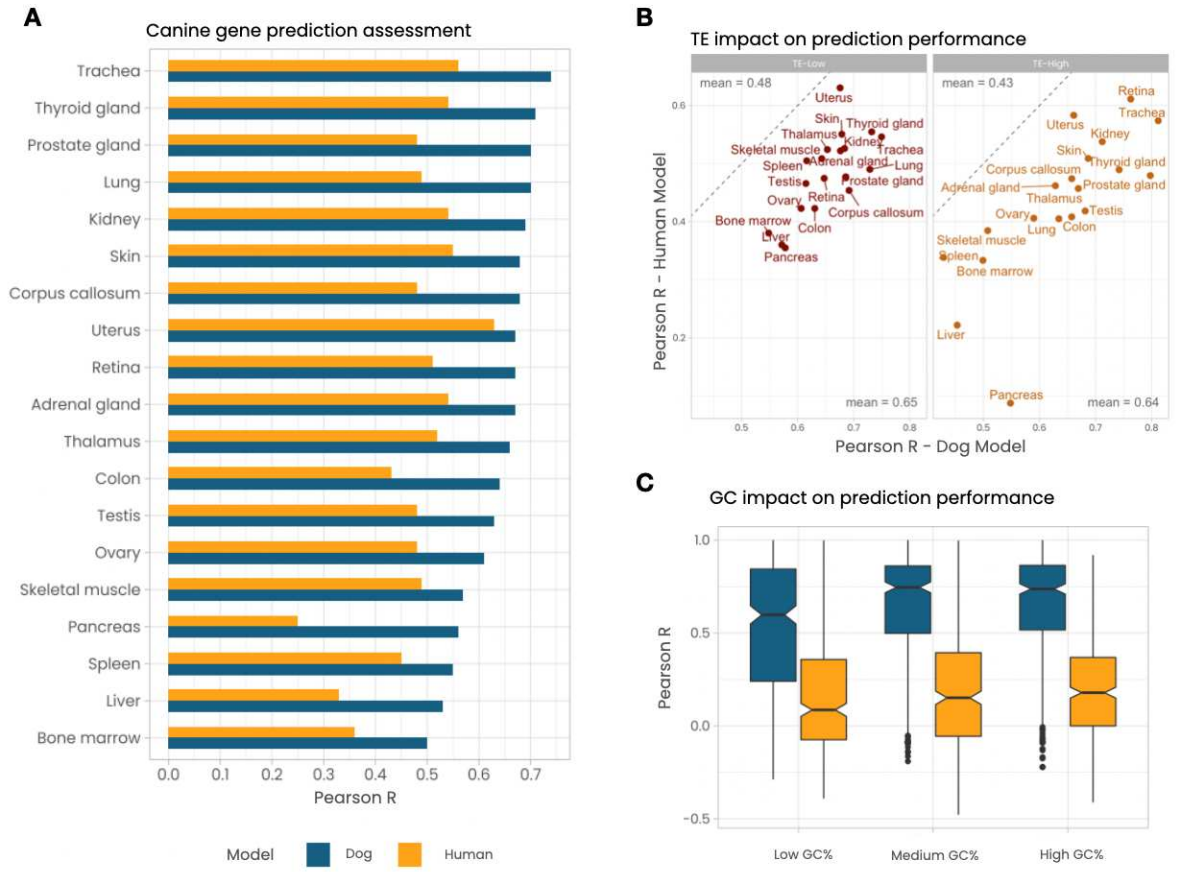

Fig 2. Within-species versus cross-species prediction. A) Performance of the DOG model (within-species) (blue) versus the HUMAN model (cross-species) (yellow) for gene expression of canine cancer genes across 19 matched dog/human tissues. For each tissue, we computed the Hotelling-Williams test to assess the significance of the difference between the model predictions. Each p-value is highly significant. **B)** Influence of gene promoter contents in Transposable Element (TE) (low = left panel; high = right panel) for within-species prediction (x-axis) versus cross-species prediction (y-axis); dotted diagonal represents the line where y=x . **C)** Influence on Pearson correlation of gene promoter content in GC (low, medium and high according to ⅓ and ⅔ percentiles GC%) for within-species prediction (blue boxes) and cross-species prediction (yellow boxes).

This study highlighted that in all of the 19 tissues considered, the dog model always exhibited higher Pearson coefficients than those obtained with the human model (median R_{within} = 0.67 versus R_{cross} = 0.49). Next, we investigated the genomic features that could explain the lower predictive power of the human model. For the 1,317 canine promoter sequences, we computed several features such as GC content, evolutionary conservation and content in transposable elements (TEs) (see Methods). For the latter, we categorized genes as whether they contain canine SINE elements or not in their promoter sequences and computed Pearson coefficients separately for these two sets of genes for the two models. The performance of the dog model was not significantly different for genes having or not TE in their promoters (R=0.65 versus R=0.64 respectively) **(Fig 2.B)**. This could be expected given that the dog model was trained on the entire dog genome sequence which included TE elements. Conversely, the human model performed poorly for dog genes containing TE in their promoters as compared to genes without TEs (R=0.43 versus R=0.48 respectively), probably highlighting canine-specific SINEs important for gene expression regulation but not learnt by the human model. To assess the impact of GC content of gene promoter on prediction performance, gene promoter sequences were split in three categories (low, medium and high GC) with respect to ⅓ and ⅔ percentiles GC **(Fig 2.C)**. With both prediction models, performance was improved for promoters

exhibiting high and medium GC content although the difference was more pronounced with the within-species model compared to the cross-species model (R_{cross} = 0.12, R_{cross} = 0.15 and R_{cross} = 0.18, R_{within}= 0.57, R_{cross} = 0.74 and R_{cross} = 0.73 for low, medium and high GC content, respectively). Interestingly in dogs, it is known that gene promoters are highly enriched in GC content, in part due to the loss of the *PRDM9* gene [17]. Finally, we also compared the influence of sequence conservation in the performance of the cross-species model. As for GC content, gene promoters were separated as lowly and highly conserved with respect to median blat score (see Methods). As expected, performance of the cross-species model was found lower for genes with lower level of sequence conservation than those with a higher level of sequence conservation (R_{cross}=0.34 and 0.45, respectively).

4. Conclusion - Discussion

We have developed a predictive model that can process a genome sequence to better understand gene expression in dogs. We present a CNN-based canine-specific model of gene expression using the most comprehensive set of CAGE data provided by the DoGA consortium. We showed that this dog model has better performance for predicting dog gene expression compared to a cross-species strategy using a human trained model. Hence, we provide a canine-specific model that can be used to predict gene expression and their variations for the largest collection of tissue (n=37) currently available. This computational tool is designed to be used by the scientific community that relies on the dog genetic model to decipher complex diseases and phenotypes mapping. Although the cross-species model displays lower performance, under specific conditions (GC and TE content, sequence conservation), a sub-optimal use of the cross-species model can be envisaged to extend its application to the several hundreds of tissues and cell lines included in the human model.

While this CNN-based model reached good correlations between predicted and experimentally measured expression levels, recent deep learning architecture involving transformers have been shown to outperform CNN particularly for capturing longer interaction between genes and their regulatory elements [18]. It could thus be interesting to apply such approaches for dog gene expression prediction.

The ultimate goal of gene expression models consists in learning the DNA regulatory code responsible for gene expression in a tissue-specific manner and thus to categorize the complexity of non-coding mutations and prioritize their impacts and roles. In our work, we have defined and extracted the gene promoter sequence of 1,317 genes involved in cancers. Using an *in silico* saturation mutagenesis analysis [19] of the promoter sequences, we seek to identify the regulatory variations between humans and dogs that will be predicted by both models that *a priori* significantly alter gene expression levels in both species. This aspect will be covered in an extended version of the work.

Acknowledgements

Authors thank the bioinformatics core facility Genouest (https://www.genouest.org/) and the Dog Genome Annotation (DoGA) Consortium (César L. Araujo, Milla Salonen, Riika Sarviaho, Julia Niskanen, Sruthi Hundi, Jenni Puurunen, Sini Sulkama, Sini Karjalainen, Antti Sukura, Pernilla Syrjä, Niina Airas, Henna Pekkarinen, Ilona Kareinen, Anna Knuuttila, Heli Nordgren, Karoliina Hagner, Tarja Pääkkönen, Kaarel Krjutskov, Sini Ezer, Shintaro Katayama, Masahito Yoshihara, Auli Saarinen, Matthias Hörtenhuber, Amitha Raman, Irene Stevens, Maria Kaukonen, Ileana B. Quintero, Abdul Kadir Mukarram, Marjo K. Hytönen, Kaisa Kyöstilä, Meharji Arumili, Carsten O. Daub, Juha Kere, Hannes Lohi) for the CAGE data. David R. Kelley for helpful suggestion during elaboration of the dog prediction model (http://www.davidrkelley.com/)

References

[1] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature Genetics*, 2018, doi: 10.1038/s41588-018-0295-5.

[2] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nat Genet*, vol. 50, no. 8, pp. 1171–1179, Aug. 2018, doi: 10.1038/s41588-018-0160-6.

[3] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nat Rev Genet*, vol. 20, no. 7, pp. 389–403, Jul. 2019, doi: 10.1038/s41576-019-0122-6.

[4] L. Chen, A. E. Fish, and J. A. Capra, "Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties," *PLoS Comput Biol*, vol. 14, no. 10, p. e1006484, Oct. 2018, doi: 10.1371/journal.pcbi.1006484.

[5] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks.," *Genome research*, vol. 28, no. 5, pp. 739–750, 2018, doi: 10.1101/gr.227819.117.

[6] L. A. Hindorff *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009, doi: 10.1073/pnas.0903103106.

[7] K. Cochran *et al.*, "Domain-adaptive neural networks improve cross-species prediction of transcription factor binding," *Genome Res.*, vol. 32, no. 3, pp. 512–523, Mar. 2022, doi: 10.1101/gr.275394.121.

[8] D. R. Kelley, "Cross-species regulatory sequence activity prediction," *PLOS Computational Biology*, vol. 16, no. 7, p. e1008050, Jul. 2020, doi: 10.1371/journal.pcbi.1008050.

[9] M. Lizio *et al.*, "Monitoring transcription initiation activities in rat and dog," *Sci Data*, vol. 4, no. 1, p. 170173, Nov. 2017, doi: 10.1038/sdata.2017.173.

[10] D. Chakravarty *et al.*, "OncoKB: A Precision Oncology Knowledge Base," *JCO Precis Oncol*, vol. 2017, Jul. 2017, doi: 10.1200/PO.17.00011.

[11] J. Herrero *et al.*, "Ensembl comparative genomics resources," *Database : the journal of biological databases and curation*, vol. 2016, pp. 1–17, 2016, doi: 10.1093/database/bav096.

[12] J. Navarro Gonzalez *et al.*, "The UCSC Genome Browser database: 2021 update," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1046–D1057, Jan. 2021, doi: 10.1093/nar/gkaa1070.

[13] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *BIOINFORMATICS APPLICATIONS NOTE*, vol. 26, no. 6, pp. 841–842, 2010, doi: 10.1093/bioinformatics/btq033.

[14] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).

[15] C. Le Béguec *et al.*, "Characterisation and functional predictions of canine long non-coding RNAs.," *Scientific reports*, vol. 8, no. 1, p. 13444, Sep. 2018, doi: 10.1038/s41598-018-31770-2.

[16] A. Breschi, T. R. Gingeras, and R. Guigó, "Comparative transcriptomics in human and mouse," *Nature Reviews Genetics*, vol. 18, no. 7, pp. 425–440, 2017, doi: 10.1038/nrg.2017.19.

[17] C. Wang *et al.*, "A novel canine reference genome resolves genomic architecture and uncovers transcript complexity," *Communications Biology*, vol. 4, no. 1, p. 185, Dec. 2021, doi: 10.1038/s42003-021-01698-x.

[18] Ž. Avsec *et al.*, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nat Methods*, pp. 1–8, Oct. 2021, doi: 10.1038/s41592-021-01252-x.

[19] S. Nair, A. Shrikumar, J. Schreiber, and A. Kundaje, "fastISM: Performant in-silico saturation mutagenesis for convolutional neural networks," *Bioinformatics*, p. btac135, Mar. 2022, doi: 10.1093/bioinformatics/btac135.

Annexe II : Article de collaboration

En annexe II, je reporte un article où je suis co-auteure, en raison de ma contribution aux analyses statistiques des travaux présentés. En effet au cours de ma thèse j'ai réalisé et apporté ma contribution aux reflexions des analyses statistiques de plusieurs projets. Cet article ne fait pas partie de mon propre projet de thèse et est donc reporté en annexe.

cancers

MDPI

Article

Canine Oral Melanoma Genomic and Transcriptomic Study Defines Two Molecular Subgroups with Different Therapeutical Targets

Anais Prouteau [1], Stephanie Mottier [1], Aline Primot [1], Edouard Cadieu [1], Laura Bachelot [1], Nadine Botherel [1], Florian Cabillic [2], Armel Houel [1], Laurence Cornevin [2], Camille Kergal [1], Sébastien Corre [1], Jérôme Abadie [3], Christophe Hitte [1], David Gilot [1], Kerstin Lindblad-Toh [4,5], Catherine André [1], Thomas Derrien [1,*] and Benoit Hedan [1,*]

1 IGDR—UMR 6290, CNRS, University of Rennes 1, 35000 Rennes, France; prouteauanais@gmail.com (A.P.); stephanie.mottier@univ-rennes1.fr (S.M.); aline.primot@free.fr (A.P.); edouard.cadieu@univ-rennes1.fr (E.C.); laura.bachelot@univ-rennes1.fr (L.B.); nadine.botherel@univ-rennes1.fr (N.B.); armel.houel@univ-rennes1.fr (A.H.); camille.kergal@univ-rennes1.fr (C.K.); sebastien.corre@univ-rennes1.fr (S.C.); christophe.hitte@univ-rennes1.fr (C.H.); david.gilot@univ-rennes1.fr (D.G.); catherine.andre@univ-rennes1.fr (C.A.)

2 Laboratoire de Cytogénétique et Biologie Cellulaire, CHU de Rennes, INSERM, INRA, University of Rennes 1, Nutrition Metabolisms and Cancer, 35000 Rennes, France; florian.cabillic@chu-rennes.fr (F.C.); laurence.cornevin@univ-rennes1.fr (L.C.)

3 Laboniris, Department of Biology, Pathology and Food Sciences, Oniris, 44300 Nantes, France; jerome.abadie@oniris-nantes.fr

4 Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; kersli@broadinstitute.org

5 Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 24 Uppsala, Sweden

* Correspondence: thomas.derrien@univ-rennes1.fr (T.D.); benoit.hedan@univ-rennes1.fr (B.H.); Tel.: +33-2-23-23-43-19 (B.H.)

Simple Summary: In humans, mucosal melanoma (MM) is a rare and aggressive cancer. The canine model is frequently and spontaneously affected by MM, thus facilitating the collection of samples and the study of its genetic bases. Thanks to an integrative genomic and transcriptomic analysis of 32 canine MM samples, we identified two molecular subgroups of MM with a different microenvironment and structural variant (SV) content. We demonstrated that SVs are associated with recurrently amplified regions, and identified new candidate oncogenes (*TRPM7*, *GABPB1*, and *SPPL2A*) for MM. Our findings suggest the existence of two MM molecular subgroups that could benefit from dedicated therapies, such as immune checkpoint inhibitors or targeted therapies, for both human and veterinary medicine.

Abstract: Mucosal melanoma (MM) is a rare, aggressive clinical cancer. Despite recent advances in genetics and treatment, the prognosis of MM remains poor. Canine MM offers a relevant spontaneous and immunocompetent model to decipher the genetic bases and explore treatments for MM. We performed an integrative genomic and transcriptomic analysis of 32 canine MM samples, which identified two molecular subgroups with a different microenvironment and structural variant (SV) content. The overexpression of genes related to the microenvironment and T-cell response was associated with tumors harboring a lower content of SVs, whereas the overexpression of pigmentation-related pathways and oncogenes, such as *TERT*, was associated with a high SV burden. Using whole-genome sequencing, we showed that focal amplifications characterized complex chromosomal rearrangements targeting oncogenes, such as *MDM2* or *CDK4*, and a recurrently amplified region on canine chromosome 30. We also demonstrated that the genes *TRPM7*, *GABPB1*, and *SPPL2A*, located in this CFA30 region, play a role in cell proliferation, and thus, may be considered as new candidate oncogenes for human MM. Our findings suggest the existence of two MM molecular subgroups that may benefit from dedicated therapies, such as immune checkpoint inhibitors or targeted therapies, for both human and veterinary medicine.

check for updates

Citation: Prouteau, A.; Mottier, S.; Primot, A.; Cadieu, E.; Bachelot, L.; Botherel, N.; Cabillic, F.; Houel, A.; Cornevin, L.; Kergal, C.; et al. Canine Oral Melanoma Genomic and Transcriptomic Study Defines Two Molecular Subgroups with Different Therapeutical Targets. *Cancers* **2022**, *14*, 276. https://doi.org/10.3390/cancers14020276

Academic Editors: Ludwig M. Heindl and Jochen Sven Utikal

Received: 25 November 2021
Accepted: 27 December 2021
Published: 6 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: mucosal melanoma; dog model; oncogenes; immune checkpoint inhibitors; chromosomal rearrangements; *CDK4*; *MDM2*

1. Introduction

Melanoma is the deadliest skin cancer in humans, with an incidence of 80,000 cases per year in the USA [1]. Mucosal melanoma (MM), a rare clinical entity, accounts for 1–2% [2,3] of all melanomas; however, this rate is approximately 22% in certain Asian countries where cutaneous melanoma (CM) occurs less frequently [4]. MM is caused by melanocytes that reside in the mucous membranes of the respiratory, gastrointestinal, and urogenital tracts, and occurs mainly in the head and neck (31–55%), anorectal (17–24%), and vulvovaginal (18–40%) regions [3]. Compared to CM, MM is highly aggressive and has a less favorable prognosis [2,5,6], with a 5-year survival rate of 20–35% based on the disease location and stage [2,3,6,7]. Treatment options include surgical resection and/or radiation therapy to achieve locoregional control; however, the prognosis remains poor. Some advanced cases of MM may benefit from targeted therapy, such as the usage of KIT inhibitors or immunotherapy using checkpoint inhibitors [2,3,8–10]. However, the response to immunotherapy is highly variable, and patients with MM have a lower response rate than those with CM [3,8].

In recent years, large genomic studies involving whole exome sequencing (WES) and, more recently, whole genome sequencing (WGS) have been conducted to examine the mechanisms of MM tumor initiation, progression, and metastasis, and to find more suitable therapies [11–18]. These studies have shown that MM presents a lower mutation burden and has a greater load of structural variants (SVs) and copy number alterations (CNAs) than CM [11,12,14,15]. In MM, SV and CNA frequently involve known oncogenes, such as *CDK4*, *MDM2*, and *TERT* [14–16,18]. Two recent studies have identified two MM subgroups based on the pattern of complex clustered SVs [14,15]. In oral melanomas comprising one such pattern, some of the SVs were linked to poor outcomes [15]. Since MM is a rare entity in humans, popular breeds of dog patients with MM can be evaluated to better understand the molecular characteristics of these MM subgroups and to reveal relevant therapeutic targets.

In the last decade, dog models have emerged as being unique, spontaneous, and immunocompetent for human cancers [19–24], particularly for MM [13,22,23]. Canine MM is the most frequently occurring oral malignancy in dogs and shares several clinical, biological, and genetic features with its human counterpart. As in humans, canine MM tends to rapidly metastasize to distant organs and responds poorly to chemotherapy. Immunotherapy in dogs has been developed to create a melanoma vaccine; however, the results remain controversial, with highly variable responses [23,25,26]. Recently, the development of anti PD-L1 [27,28] and anti-PD1 antibodies [26] for canine patients with MM has proven to be effective in pilot studies. Genetically, canine MM mimics human MM with a predominance of somatic CNAs and complex SVs, and a relatively low tumor mutation burden [13,24,29]. At the transcriptomic level, only a few studies in canine MM tumors involving a limited number of samples have identified deregulation in multiple pathways, such as MAPK/ERK, PI3K/AKT, "cytokine receptor interaction", "ECM receptor interaction", and "focal adhesion" [30–32].

In this study, we examined the genetic basis of MM in a spontaneous canine model in specific breeds that are frequently affected by MM by combining genomic and transcriptomic sequencing experiments together with functional validations. Unsupervised clustering analysis of RNAseq data from 32 canine MM samples highlighted the existence of two molecular subgroups of MM with differential expressions: one characterized by immune and microenvironment signatures, and the second by the overexpression of oncogenes (such as *MDM2*, *CDK4*, and *TERT*, which were characterized by the enrichment of complex chromosomal rearrangements). The effect on cell proliferation of genes, which is

frequently altered through chromosomal rearrangements, was explored in both canine and human MM cell lines. The results of the study suggest the existence of two MM molecular subgroups that could benefit from distinct therapies and lead to further translational studies in human and veterinary oncology.

2. Materials and Methods

2.1. Sample Collection

Blood and tissue biopsy samples from dogs with oral melanoma were collected for this study through Cani-DNA (http://dog-genetics.genouest.org, accessed on 18 November 2021) and the Canine Comparative Oncology Genomics Consortium Biological Resource Centers BRC (www.ccogc.org, accessed on 1 July 2021) (61 and 18 cases, respectively) (Supplementary Materials Table S1). Oral melanoma diagnoses were confirmed based on the histopathological analysis (JA) with adequate immunostaining (expression of the S100, MelanA, PNL2 markers). DNA and RNA were extracted, as previously described [21,33]. Blood and tissue sampling was performed by veterinarians on privately owned dogs during the course of the health follow-up, with the owner's consent.

2.2. Canine Cell Lines

The Bear (accession numbers CVCL_OD14) and CML10 (accession number CVCL_IZ11) cell lines were obtained from Dr. J. Modiano (Colorado State University, Fort collins, CO, USA). The cell lines were cultured at 37 °C in RPMI 1640 medium (Gibco, Amarillo, TX, USA), supplemented with 10% fetal bovine serum and 0.2% primocin (Invivogen, San Diego, CA, USA). Five other cell lines, obtained from fresh canine oral melanoma samples (Dog-OralMel-18249, Dog-OralMel-18395, Dog-OralMel-18333, Dog-OralMel-18848, and Dog-OralMel-18657), were developed in the laboratory (Table S1). These cell lines were cultured in DMEM/F-12 medium (Gibco), supplemented with 10% fetal bovine serum and 0.2% primocin. After ten passages, RPMI 1640 was used. All cell lines were tested for mycoplasma using the MycoAlertTM Plus kit (Lonza, Rockland, ME, USA), and were found to be mycoplasma-free.

2.3. Fluorescence In Situ Hybridization

Fluorescence in situ hybridization analyses were performed on chromosome preparations generated from the cell lines using the following conventional techniques: colcemid arrest, hypotonic treatment, and methanol/glacial acetic acid fixation, as described previously [34]. The following bacterial artificial chromosome (BAC) clones were used: CH82-199H02 and CH82-179B09 for the *MDM2* region, CH82-213B06 and CH82-204K11 for the *CDK4* region, CH82-99P23 and CH82-60O16 for the CFA 7 region (58.4 Mb to 58.9 Mb), CH82-1E17 and CH82-40I15 for the region containing *GABPBP1*, and *USP8* and *TRPM7* amplifications (https://bacpacresources.org/, accessed on 1 July 2021). These BAC clones were labeled using green-dUTP (Abbott Molecular, Des Plaines, IL, USA) and Cy3-dCTP (Amersham Biosciences, Chalfont, UK). The slides were analyzed by an experienced cytogeneticist (FC) using a fluorescence microscope (Axioskop2, Axio Imager Z2, Zeiss, Göttingen, Germany) and Isis imaging software (Metasystems, Altlussheim, Germany). At least 100 non-overlapping tumor nuclei were examined in this study.

2.4. RNAseq Clustering and Signature Analysis

For the 32 canine MM samples, polyadenylated RNAs were extracted, sequenced, and analyzed, as previously described [33]. All RNAseq fastq files are available in the SRA under the BioProject PRJNA749900. Briefly, the pipeline used the "canFam3.1-plus" annotation as the reference annotation [35], and the canFam3.1 assembly version as the reference genome [36]. Based on the protocol described by Djebali et al. [37], FASTQ reads were aligned to the transcriptome and genome using the STAR program (v2.5.0a) [38]. Finally, gene expression levels were estimated as raw counts (unnormalized) using the RSEM program (v1.2.25) [39] for each sample individually, and subsequently merged to

obtain a matrix expression file (with genes in rows and samples in columns). This matrix of expression was then normalized and transformed across all samples in order to stabilize the variance using the DESEq2 (v1.22.2) [40] function *vstcounts* with the option "blind = TRUE".

To select the genes that would provide the most information for clustering analysis, we used 6000 of the most variable genes, that is, those with the highest median absolute deviation. Then, the nonnegative matrix factorization (NMF) algorithm from the NMF R package (v0.22.0) [41] was applied for different values of *k* clusters/ranks (*k* = 2–6). To determine the best *k* cluster, the *nmfEstimateRank* function was used (Appendix A Figure A1), and the consensus matrix method (Silhouette) identified *k* = 2 as the optimal number of clusters. Finally, we employed the NMF function *extractFeatures* to obtain cluster-specific gene signatures for the two clusters (Table S2). These gene lists were then used as the input for the gprofiler2 (v0.1.9) R program [42] with the "*gost*" function, with the "organism" set to "cfamiliaris" to perform a gene set enrichment analysis over multiple databases (Gene Ontology, KEGG, CORUM). Finally, heatmaps were plotted using the complexHeatmap software (v2.3.1) [43] to integrate and visualize multiple sources of information (gene signatures, sample clustering, SV content, and oncogene mutational status) (https://github.com/tderrien/Prouteau_et_al, accessed on 18 November 2021).

2.5. Whole Genome Sequencing

DNA from four canine oral melanoma cell lines and the blood samples of two affected dogs were extracted for whole genome sequencing (WGS), as previously described [21]. WGS was performed with the BGI sequencing platform (BGI, Shenzhen, China), using BGISEQ-500 short-read sequencing as previously described [44]. Briefly, 1000 ng of genomic DNA was quantified using a Qubit 3.0 fluorometer (Life Technologies, Paisley, UK) and sheared using an E220 Covaris instrument (Covaris Inc., Woburn, MA, USA). Sizes were selected using a Vahstm DNA Clean beads kit (Vazyme, Nanjing, China) to an average size of 200–400 bp. The selected fragments were end repaired, and 3′ adenylated and BGISEQ-500 platform-specific adaptors were ligated to the A-tailed fragments. The ligated fragments were purified and amplified by PCR. Finally, circularization was performed to generate single-stranded DNA circles. After quantification, the libraries were loaded onto a sequencing flow cell and processed for 100 bp paired-end sequencing on the BGISEQ-500 platform.

2.6. Low Pass Sequencing

Low-pass sequencing of 42 formalin-fixed paraffin-embedded (FFPE) samples of melanoma was performed by Psomagen (Rockville, MD, USA). Briefly, the input DNA quality was verified using gel electrophoresis, and the quantity was measured using the Picogreen assay (Thermo Scientific, Waltham, MA, USA); 2 ng of DNA was prepared in 30 ul of buffer and used for library construction using the Nextera DNA Flex Library Kit (Illumina, San Diego, CA, USA), according to the manufacturer's guidelines. The size of the final DNA libraries was then validated using the TapeStation D1000 ScreenTape (Agilent, Santa Clara, CA, USA) and D1000 reagents (Agilent, Santa Clara, CA, USA). The quantity was measured using the Picogreen assay (Thermo Scientific, Waltham, MA, USA), and the molar concentration was calculated using both sources. The libraries were normalized to 2 nM, pooled in equimolar volume, and then loaded on the flow cell from the Novaseq S4 300 cycle kit (Illumina, San Diego, CA, USA) and the XP-4lane kit (Illumina, San Diego, CA, USA). The prepared flow cell and SBS cartridge from the Novaseq S4 300 cycle kit (Illumina, San Diego, CA, USA) were inserted into the Novaseq 6000 system and sequenced using 151-10-10-151 running parameters, reaching a mean depth of 0.48 X.

2.7. Mapping

After sequencing, the raw reads were filtered (adapter sequences, contamination, and low-quality reads were removed) according to the manufacturer's guidelines. Sequence data were then aligned to the dog reference genome (assembly version canFam3.1) using

BWA-MEM (version 0.7.17) [45], and PCR duplicate reads were removed using Picard tools (version 2.18.23) (http://broadinstitute.github.io/picard/, accessed on 1 July 2021). The read data were processed according to the GATK best practices; specifically, the base quality score recalibration was assessed using GATK4 (version 4.0.12).

2.8. Somatic Variant Calling

The Mutect2 tool from the GATK4 software was then used to call somatic SNVs and short INDELS against a panel of normal (PON), comprising matched normal samples, and against germline variants from 722 dog genomes (https://data.broadinstitute.org/vgb/Ostrander_VCFs/722g.990.SNP.INDEL.chrAll.vcf.gz, accessed on 1 July 2021). Variant annotation was performed using the VEP program [46] with the EnsEMBL "Canis familiaris" annotation (v. 95).

2.9. Somatic Copy Number Variant Calling

The somatic copy number variants (CNVs) were determined using the "R" Package DNAcopy [47]. The canine genome was split into bins with a window size of 10 kb, or based on the exome target regions for WGS and WES. The number of reads was normalized to the total number of reads per sample. The log2 ratio with the corresponding germinal DNA, when available (or with a pool of three normal DNA for BEAR or CML10 cell lines), was estimated for each window. Several windows were merged (segmentation) into a larger segment with a significance threshold of 1×10^4, and the copy number of the segments was estimated.

2.10. SV Calling

Based on the aligned bam files from the BWA-MEM software, four structural variant types were identified: breakends (BND), deletions (DEL), duplications (DUP), and inversions (INV) using Delly software [48] (version v0.8.3). Only somatic SVs with the flags "PASS" and "PRECISE" were retained. Finally, circular representations of the SVs and the gain/loss from the DNA copies along the canine chromosomes were measured using the circlize R package [49] (version 0.4.10) (https://github.com/tderrien/Prouteau_et_al, accessed on 18 November 2021).

2.11. Modeling Chromothripsis Events

To assess the patterns of chromothripsis, we tested for the enrichment of chromosome-specific SVs by modeling the association between SV breakpoints and chromosome length using hypergeometric distributions (or a hypermetric distribution). More precisely, we defined "x" as the number of SV breakpoints on a specific chromosome out of a total of "k"-detected breakpoints, where "m" was the chromosome size and "N" was the size of the canine genome. The computed p-value was the probability of "x" SVs in the total number of SVs, calculated using dhyper ($x = x$, $m = m$, $n = N - m$, $k = k$) in R. Statistical significance was set at $p < 10^{-3}$. The clustering of the SV breakpoints was tested as previously described [50]. Chromothripsis regions were detected with Shatterproof software [51], using the copy number and BND information along with the standard parameters.

2.12. Classification of Dogs with "Low" versus "High" Content of Structural Variants (SV)

High/low SV classification was defined using the copy number status of segments resulting from the WES data. SV samples were defined as "high" if they had an abundance of copy number states oscillating on at least one chromosome (copy number states $\geq$10 states) [14], with at least one major amplification (copy number $\geq$ 90 percentile of whole genome amplifications). This classification was verified using manual curation.

2.13. Targeted Copy Number and Expression Analysis

High-throughput quantitative PCR (qPCR) was performed using the SmartChip system (Takara Bio, Kusatsu, Japan) on 47 oral melanoma samples to detect focal amplifications

on tumor DNA on CFA 10 and CFA 30. We used primer pairs targeting two genes on CFA 10 (*MDM2* and *CDK4*) and four genes on CFA 30 (*TRPM7*, *USP8*, *SPPL2A*, and *GABPB1*) (Table S3). A primer pair targeting a region of CFA 9 was used as an internal control, and each experiment was performed using DNA from an unaffected dog as an external control. qPCR was performed on tumor DNA samples after pre-amplification using the SYBR green PCR master mix (Thermo Fisher Scientific, Waltham, MA, USA) on the 7900HT Fast Real-Time PCR System (Applied Biosystems, Waltham, MA, USA), using standard procedures. Each sample was measured thrice, and the relative amounts of the sequences were determined using the ΔΔCt method.

RNA was extracted from the 47 tumor samples and reverse transcription was performed on 1 µg of RNA from tumorous or healthy tissue using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Waltham, MA, USA), according to the manufacturer's instructions. RT-qPCR was performed using the same target genes. *SAP130* was used as the control housekeeping gene, and a pool of data from non-affected oral tissues of four dogs was used as the external control. Each sample was measured thrice, and the relative amounts of the sequences were determined using the ΔΔCt method.

2.14. Western Blot

Cellular protein extracts from cell lines were prepared using a cell lysis buffer containing 18 mmol/L Tris-HCl, pH 7.5; 135 mmol/L NaCl, 0.9 mmol/L EDTA; 0.9 mmol/L EGTA and supplemented with 1 mmol/L PMSF; 1× EDTA-free cocktail protease inhibitor (Roche Diagnostics, Bale, SWISS); 30 mmol/L sodium fluoride, 40 mmol/L glycero phosphate; 1 mmol/L sodium orthovanadate; and 0.5% Triton X-100.

Protein concentrations were determined using the BCA protein assay (Sigma-Aldrich, St. Louis, MO, USA) with bovine serum albumin as a standard. Protein samples were denatured for 10 min at 95 °C, and equal amounts of cell proteins (50 µg) were subjected to 10% SDS-PAGE and transferred onto nitrocellulose membranes (Amersham-GEH Life, Buckinghamshire, UK). The membranes were probed with the appropriate antibodies. The primary antibodies used were: anti-CDK4 (559693, BD Pharmingen, San Diego, CA, USA), anti-MDM2 (clone 2A10, ref MABE281, Merckmillipore, Darmstadt, Germany), and anti–ERK (sc-94, Santa Cruz Biotechnology, Dallas, TX, USA). Horseradish peroxidase-conjugated secondary antibodies were purchased from Jackson ImmunoResearch (West Grove, PA, USA). Signals were detected using a LAS-4000 Imager (Fuji Photo Film, Tokyo, Japan).

2.15. Transfection with Specific and Control siRNAs and Cell Proliferation Assays

The siRNAs (80 nM) for CDK4, GABPB1, TRPM7, USP8, SPPL2A, and control siRNA (IDT) were transfected into the Bear cell line using Lipofectamine 2000 (Invitrogen, Waltham, MA, USA), according to the manufacturer's recommendations (Tables S1 and S3). CDK4 siRNA was used as the positive control. The siRNAs (10 nM) for GABPB1, TRPM7, and SPPL2A, as well as the control siRNA (IDT), were transfected into two human cell lines (HMV-2/CVCL_1282 -Merck-, WM3211/CVCL_6797 -Rockland-) using Lipofectamine 2000 (Invitrogen, Waltham, MA, USA) following the manufacturer's recommendations (Tables S1 and S3). Cells were seeded in 6-well plates in 1 mL of RPMI 1640 medium (Gibco, Amarillo, TX, USA), supplemented with 10% fetal bovine serum and 0.2% primocin (Invivogen, Waltham, MA, USA) at a density of 1×10^6 cells, and incubated at 37 °C. The medium was changed 6 h post-transfection to avoid toxicity.

Cell proliferation was evaluated in 96-well plates, 72 h after transfection with an initial density of 30,000 cells per well, using a methylene blue colorimetric assay. Briefly, the cells were fixed for 30 min in 90% ethanol, removed, dried, and subsequently stained for 30 min using 1% methylene blue dye in borate buffer. The fixed cells were washed 4–5 times using tap water, and 100 µL of 0.1 N HCl was added to each well. Cell density was analyzed using a spectrophotometer at 620 nm.

3. Results

3.1. Transcriptome Landscape of Canine Oral Melanoma Highlights Two Distinct Subgroups

The ability of the expression profiles of the 32 canine MM tumors in predicting the prognostic subtypes of MM was first investigated (Table S1). Using unsupervised clustering based on the nonnegative matrix factorization (NMF) algorithm (see Section 2), we observed that the canine tumor samples could be separated into two distinct clusters (group 1, 12 samples; group 2, 20 samples) based on the underlying structure of their expression data (Figure 1A). Since NMF organizes both the samples and genes, it also allows us to define a list of genes that represent the signatures of the two clusters (Table S2). Using gene set enrichment analysis with the gprofiler tool [42], we showed that the 384 genes with the group 1 signature (Table S3) were significantly associated with immune-related functional gene sets (Biological Process Term "Immune Response" $p = 1.6 \times 10^{-24}$, "cytokine-mediated signaling pathway" $p = 2.05 \times 10^{-15}$) (Table S4). This suggests that the microenvironment differs between both groups, with the MM in group 1 presenting a higher degree of immune cell infiltration. This was confirmed by the overexpression of genes encoding components of the T-cell receptor (TCR) complex, such as *CD3E*, *CD3D*, *CD3G*, and *CD247* (Figure A2). Moreover, in this group, genes related to cytotoxic functions (such as granzyme B), the IFN-γ pathway (such as *IRF1*), or to the presence of an ongoing immune response and a cytokine-rich microenvironment (such as *IL18*, *CCL3*, *CCL4*, *CCL13*, and *CCL22*) [52] were significantly overexpressed (Figure A2). The ratio between immune cell types with an immunostimulatory or immunosuppressive function seemed to be more important than the absolute number of immune cell types in determining the antitumorigenic vs. protumorigenic role of the microenvironment [52]. Therefore, we determined the expression ratio of proinflammatory cytokines (IFN-γ, IL-1A, IL-1B, and IL-2), versus immunosuppressive molecules (IL-10, IL-11, IL-4, and TGFB1) in both groups. This ratio was significantly increased in group 1, suggesting a relative increase in pro-immunogenic responses in this group ($p = 0.03$, Student's t-test; Figure 1B). This profile is concordant with that of "hot immune" tumors [53,54], and the overexpression of immune checkpoint genes such as *CTLA4* (log2fold change = 2.78; adjusted $p = 1.22 \times 10^{-5}$), *TIM3* (log2fold change = 1.6; adjusted p-value $= 2.2 \times 10^{-9}$), *LAG3* (log2fold change = 1.3; adjusted p-value $= 5.8 \times 10^{-3}$), or immunomodulating cytokines (including IL2RG). This suggests that this MM group may benefit from treatment with immune checkpoint inhibitors [53,54]. In addition, several studies have involved the phenotype switching of melanoma cells as an escape route to CM-targeted therapies using BRAF inhibitors [55,56]. Under the control of the microenvironment or intrinsic cell factors, melanoma cells can acquire a resistance to targeted therapies by switching from a proliferative to an invasive state [57,58]. These changes to an aggressive phenotype are associated with dedifferentiation (from a melanocytic/transitory to a neural crest-like/undifferentiated phenotype) and an epithelial-to-mesenchymal-like (EMT-like) transition that promotes metastasis [59]. Several genes associated with the acquisition of an invasive-dedifferentiated-EMT-like phenotype [57,60] were also overexpressed in group 1 (Table S2).

In contrast, melanin metabolic processes and pigmentation pathways were identified in the second group ($n = 20$ samples), i.e., the biological process term "melanin biosynthetic process" $p = 7.6 \times 10^{-4}$, "melanocyte differentiation" $p = 9.9 \times 10^{-3}$, "pigmentation" $p = 1.8 \times 10^{-4}$) (Table S4). Gene signatures from group 2 corresponded to a proliferative and differentiated phenotype of melanoma (*DCT*, *MLANA*, *TYR*) [57,61]. This was reflected by the overexpression of the melanocyte-specific transcription factor MITF (log2fold change = 2.2, adjusted p-value $= 1.0 \times 10^{-5}$) (Figure 1C), which is known to control the proliferation, migration, and invasion of melanoma cells in CMs. Interestingly, even if these CMs are highly proliferative, they are highly sensitive to targeted therapies, such as those using BRAF inhibitors [57]. This group presented with the overexpression of well-recognized cancer driver genes, such as *TERT* (log2fold change = 1.5, adjusted p-value = 0.028), *KIT* (log2fold change = 2.1, adjusted p-value $= 5.0 \times 10^{-5}$), or the oncogene *POM121* (log2fold

change = 0.67, adjusted *p*-value = 0.011), which has recently been linked to poor prognosis in human oral MM [15] (Figure 1C).

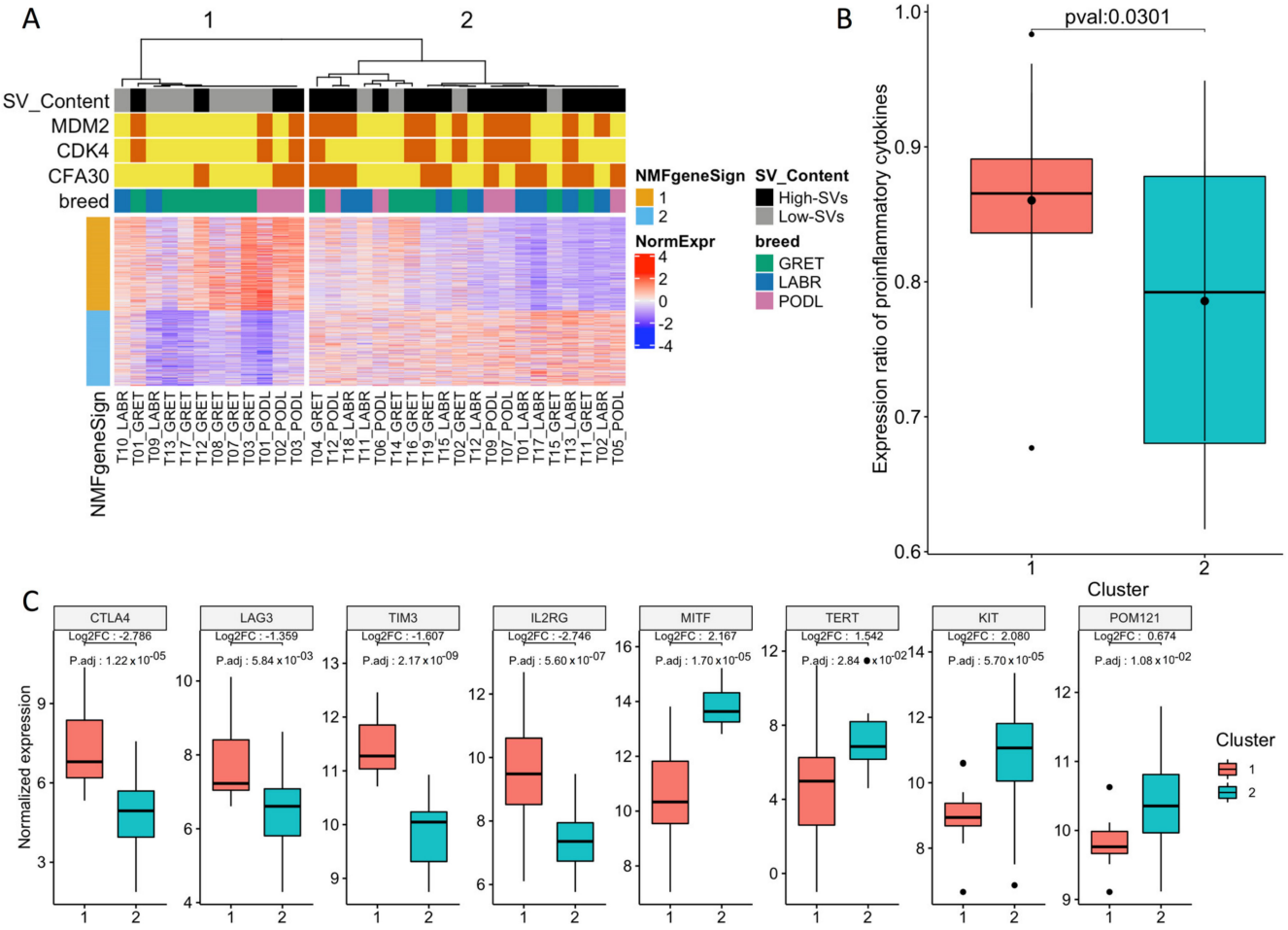

Figure 1. Transcriptomic analysis of 32 canine MM samples. (**A**) NMF clustering of expression data identified two subgroups that overlap with SV content. The first subgroup is characterized by the expression of immune response and cytokine-mediated signaling pathways and a low SV content, whereas the second subgroup is characterized by the expression of melanin metabolic processes and pigmentation pathways and is enriched in SV as well as focal amplifications of MDM2, CDK4, or CFA30:17Mb region. (**B**) Expression ratio of proinflammatory cytokines (IFN-γ, IL-1A, IL-1B, and IL-2) and immunosuppressive molecules (IL-10, IL-11, and TGFB1) significantly differs between the two groups ($p = 0.03$, Student test). (**C**) Expression of immune checkpoint genes or known oncogenes according to the transcriptomic classification. The following immune checkpoint genes CTLA4, LAG3, TIM3, and IL2RG are overexpressed in group 1, while the oncogenes MITF, TERT, KIT, and POM121 are overexpressed in group 2.

SVs involving *MDM2* and *CDK4* genes are hallmarks of human and canine MM [14–16,18], and a recent study in humans pointed out two different MM subgroups based on SV profiling [14]. Using WES data of the corresponding canine MM, we analyzed their genomic SV profiles (Section 2) and classified canine tumors into "low SV" ($n = 12$) or "high SV" ($n = 20$) (Figure 1A) based on the distribution of SV features (number, intensity, and clustering). Interestingly, the two subgroups defined by the transcriptomic analysis significantly overlapped with the two groups defined according to SV status (exact Fisher test $p = 9.5 \times 10^{-3}$). The first transcriptome subgroup, that is, "group 1" in this study, which was characterized by an immune signature, comprised tumors with "low SV," while the second transcriptome subgroup contained more "high SV" tumors. These results are

concordant with previous studies in human cancers, showing that chromosomal instability and CNAs are associated with a decrease in cytotoxic immune cell infiltration [52] and resistance to checkpoint inhibitors [52,53,62]. Our findings suggest that oral melanomas behave similarly and may evolve through two different paths driven by SV somatic changes that would lead to marked differences in therapeutic options: the first group containing "hot immune" tumors with a lower numbers of SVs would likely benefit from immunotherapy, while the second group, which contains "cold immune" tumors with a higher number of SVs would respond better to therapies targeting amplified oncogenes, such as treatment using CDK4/6 inhibitors.

Given that SVs have been associated with resistance to immunotherapy in human cutaneous melanoma [53] and with a poor outcome in both human [15] and canine oral MM [24], we then focused on the characterization of these SVs in canine MM as models for therapies in human oral MM.

3.2. "High SV" MM Are Characterized by Focal Amplifications and Numerous Chromosomal Translocations

To refine the annotation of SVs, we performed WGS on four canine MM cell lines (Bear, CML10, Dog-OralMel-18333, and Dog-OralMel-18395 cell lines) at an average depth of 30X. Then, the Delly program [48] was used to catalog a total of 5906 SVs (DEL, DUP, INV, INS, and BND) on the four cell lines (Table S5) (see Section 2). Three canine cell lines showed strong focal amplifications of CFA 30 and/or CFA 10, which are recurrently observed in canine MM [13,24,29,63,64] (Figure 2). The Bear cell line showed focal amplifications of the CFA 30, targeting the 16–17 Mb region (up to 67 copies) and of the CFA 10 targeting *MDM2* and *CDK4* region (up to 46 copies). The Dog-OralMel-18333 cell line was characterized by focal amplifications targeting the same region of CFA 30 (up to 27 copies) and CFA 7 (up to 13 copies), whereas the Dog-OralMel-18395 cell line presented strong focal amplifications on CFA 10 (up to 41 copies) targeting *MDM2* and *CDK4*. In contrast, the last cell line, CML10, did not show any of these alterations. These SV profiles suggest that the CML10 cell line belongs to MM group 1 while the three others to the group 2.

Interestingly, when analyzing the SV distribution in the tumor genome, we found that the regions harboring focal amplifications comprised significant amounts of several types of SVs, such as INVs, DELs, and insertions (Supplementary Table S6 and Figure 2). In particular, these focal amplifications were found to display inter-or intra-chromosomal translocations with significant breakpoint clustering, similar to what is seen in human MM, particularly oral MM, which presents inter-and intra-chromosomal translocations between or within HSA 5 (*TERT*) and HSA 12 (*CDK4*) [14,16]. Thus, for the three canine MM cell lines with strong focal amplifications, the amplified regions presenting breakpoint clustering (see Methods) were detected as potential chromothripsis regions using the Shatterproof tool (Table S6). For the three cell lines harboring these focal amplifications, inter-or intra-chromosomal translocations were explored using FISH.

It showed large genomic regions with repeated amplifications and fusions of CFA 30 and CFA 10 for the Bear, CFA 30 and CFA 7 for the Dog-OralMel-18333, and intra CFA10 for the Dog-OralMel-18395 cell lines (Figure 2C,F,I). Although we could not find the presence of double-minute chromosomes to explain the highly elevated copy numbers using FISH analysis, the three cell lines harboring strong focal amplifications presented features of chromothripsis, as defined by Korbel et al. [50], that is, breakpoint clustering and irregular oscillating copy number states.

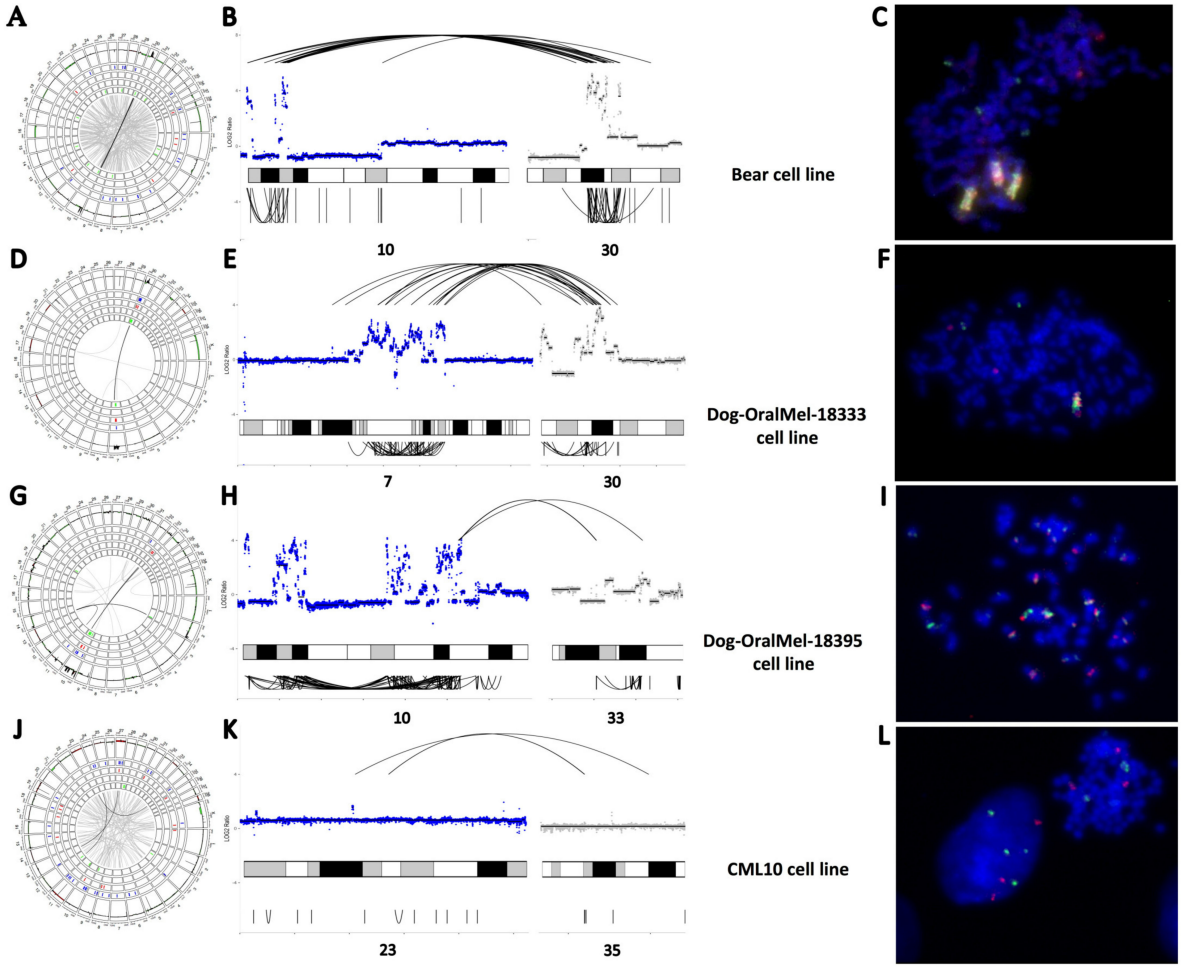

Figure 2. Structural variants (SVs) of the four canine MM cell lines. (**A**–**C**) SVs identified through WGS in the cell line Bear. (**A**). Circos plot representing the distribution of SVs along dog chromosomes with, from external to internal layers, CNA gains/losses (in dark red/green), deletions, duplications, insertions, and inversions in blue, light red, orange, and green, respectively. Interchromosomal break-ends (BND) are represented by gray lines connecting chromosomes with a color intensity corresponding to the number of reads validating the SV. (**B**). Focus on CFA 10 and CFA 30 present the focal amplifications and clusters of SVs (BND and INV). Copy numbers in log2ratio are represented as inter (top) and intra-chromosomal (bottom) SVs. (**C**). Fluorescence in situ hybridization (FISH) analysis of Bear cell line targeting CFA 10 (MDM2 region) and CFA 30 (TRPM7 region) in green and red, respectively, showing derivative chromosomes compatible with a chromothripsis-like event. (**D**–**F**). SVs of the cell line Dog-OralMel-18333. (**D**). Circos plot representing CNV, BND, DEL, and INV across the genome. (**E**). Focus on CFA 7 and CFA 30 presenting the focal amplifications and cluster of SVs (BND, DEL, DUP and INV). (**F**). FISH analysis of Dog-OralMel-18333 cell line targeting CFA 7 and CFA 30 (TRPM7 region) in green and red, respectively, showing derivative chromosomes compatible with a chromothripsis-like event. (**G**–**I**). SVs of the cell line Dog-OralMel-18395. (**G**). Circos plot representing CNV, BND, DEL, and INV across the genome. (**H**). Focus on CFA 10 and CFA 33 presenting the focal amplifications and cluster of SVs (BND, DEL, DUP, and INV). (**I**). FISH analysis of Dog-OralMel-18395 cell line targeting CFA 10 (MDM2 region) and CFA 10 (CDK4 region) in green and red, respectively, showing derivative chromosomes compatible with a chromothripsis-like event. (**J**–**L**). SVs of the cell line CML10. (**J**). Circos plot representing CNV, BND, DEL, and INV across the genome. (**K**). Focus on CFA 23 and CFA 35 presenting interchromosomal rearrangement. (**L**). FISH analysis of CML10 cell line targeting CFA 10 (MDM2 region) and CFA 30 (TRPM7 region) in green and red, respectively, showing 3 to 4 spots compatible with a tetraploid state without massive genomic rearrangements. Image (**A**,**D**,**G**,**J**) are available in the Supplementary Material (Figures A3–A6).

Thus, the results of this study suggest that focal amplifications of the *CDK4* and *MDM2* genes, as well as those in the 16–17 Mb region of CFA 30 in canine MM, reflect major complex inter-and intra-chromosomal rearrangements. These massive DNA rearrangements, clustered with high amplifications and deletions, result in long derivative chromosomes arising from cataclysmic events compatible with chromothripsis [14,50]. These recurrent focal amplifications are expected to drive the oncogenic process of MM.

3.3. Exploring Candidate Oncogenes on CFA 30

3.3.1. Correlation between Copy Numbers (CNA) and Expression of Candidate Oncogenes

Chromothripsis-like events often involve the amplification of oncogenes to promote tumor progression [65,66]. In canine MM, these recurrent amplifications target well-recognized oncogenes, such as *CDK4,* which are involved in the early phase of the cell cycle, and *MDM2,* whose protein inhibits p53, as well as several candidate oncogenes localized on CFA 30 amplifications, which remain unidentified in human MM. Based on the intersection of genomic intervals defining recurrent CFA30 amplifications in the WES data from 32 dogs, we highlighted the minimal recurrently amplified region, containing four genes that may be relevant candidate oncogenes: *TRPM7* is involved in the PI3K/AKT and MAPK pathways, as well as in EMT in ovarian and lung cancers [67–70]; *GABPB1* encodes the ß subunit of a transcription factor and is linked to cell proliferation and poor outcome in renal cell carcinoma [71]; *USP8* has been linked to tumorigenesis and poor prognosis in several cancers [72,73]; and *SPPL2A* plays a role in innate and adaptive immunity [74].

We hypothesized that the tumor advantages conferred by these focal amplifications were associated with the overexpression of the oncogenes targeted by these amplifications. Thus, we first checked whether the expression levels of the candidate oncogenes of CFA 30 (*TRPM7, USP8, GABPB1*, and *SPPL2A*) were correlated with the copy number of these genes. With this aim, we analyzed the copy numbers and gene expression of the candidate oncogenes in 47 novel canine MM samples by performing qPCR and RT-qPCR. A significant positive correlation between gene expression and copy number was found for all four CFA 30 candidate genes, with *SPPL2A* having the strongest correlation (Spearman coefficient: 0.76, $p = 6.9 \times 10^{-9}$). Concerning oncogenes located on CFA 10, the expression levels of *MDM2* and *CDK4* were positively correlated with DNA copy numbers (Spearman coefficient: 0.68, $p = 8.9 \times 10^{-7}$ and Spearman coefficient: 0.65, $p = 1 \times 10^{-6}$, respectively) (Figure 3A). Finally, Western blot analyses using cross-reactive available antibodies showed that the *MDM2* and *CDK4* focal amplifications were linked to a higher amount of protein (Figure 3B).

Thus, amplifications and chromosomal rearrangements of CFA 10 and CFA 30 led to a clear overexpression of the targeted genes and protein levels, suggesting the acquisition of new cell capacities in connection with the proliferative and aggressive features of MM cells.

3.3.2. Effect of Candidate Oncogenes on MM Cell Proliferation

To evaluate the effect of the altered expression of the CFA30 genes on cancer cell proliferation, we assessed the proliferation capability of canine MM cancer cells after silencing *TRPM7, SPPL2A, GABPB1*, and *USP8* genes using colorimetric cell proliferation assays. Expression analysis showed that the siRNA experiments reduced the mRNA expression levels of the four oncogenes by 50–72% (Table S7), and the silencing of the *TRPM7*, *SPPL2A*, and *GABPB1* genes significantly decreased cell proliferation from 18–28% (Figure 3C). These results suggest that the combined amplification of the CFA 30 candidate oncogenes might have a cumulative effect on MM cell proliferation.

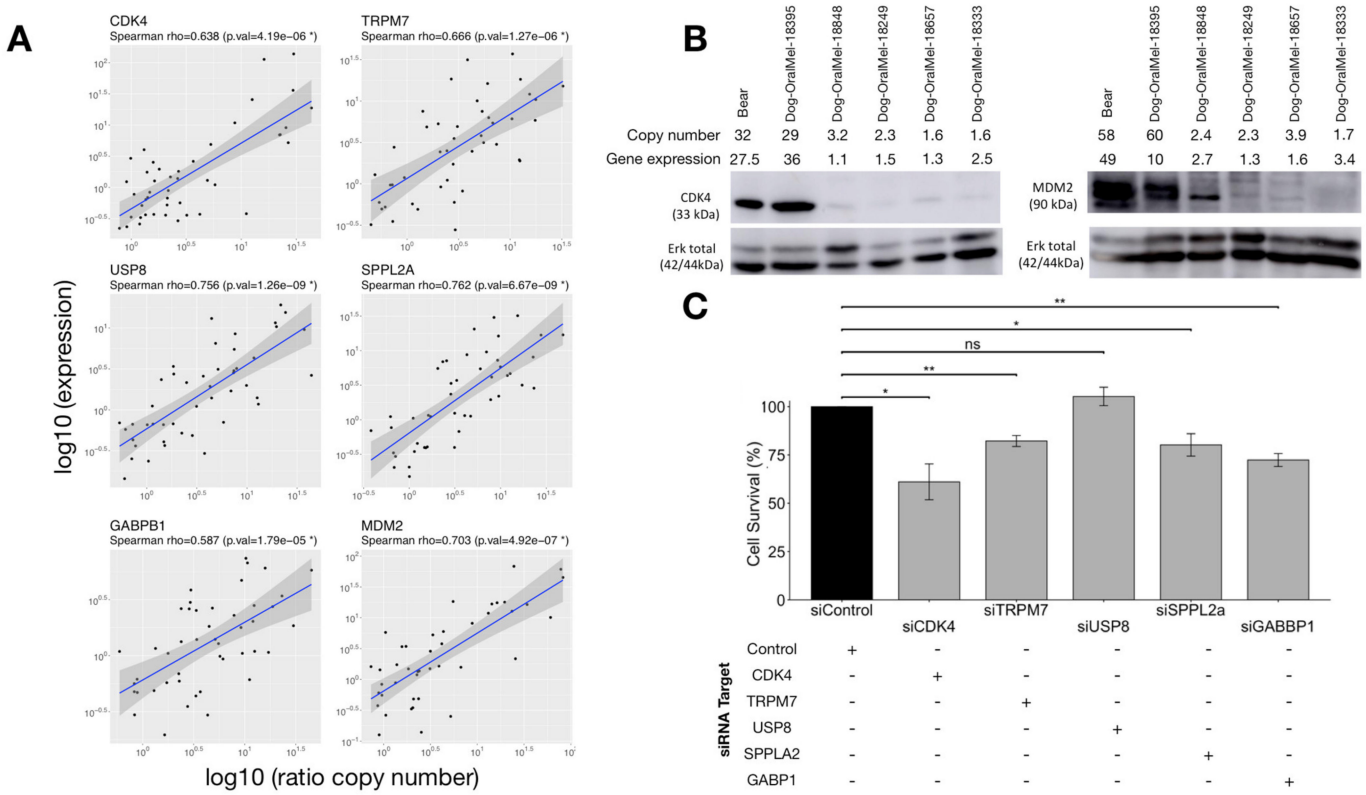

Figure 3. Expression data of the CFA 10 and CFA30 candidate genes at mRNA and protein levels and impact on cell proliferation. (**A**). For the CFA 10 genes (MDM2, CDK4) and CFA 30 genes (GABPB1, TRPM7, SPPL2A and USP8), expression was defined by RT-qPCR and was significantly correlated to copy number in tumor tissue (n = 46). (**B**). Western blot analysis in six canine MM cell lines showing a higher amount of protein in cell lines having high copy numbers of CDK4 and MDM2. (**C**). Cell proliferation assay (Bear cell line) showing the effect of the knockdown (siRNA) of candidate oncogenes from CFA 30 and CDK4 on cell survival. The knockdown of CDK4, TRPM7, GABPB1, and SPPL2A had an impact on cell proliferation in comparison with the control siRNA (t-test, with * corresponding to p-values <0.05 and ** to p-values <0.01).

3.3.3. Effect of Candidate Oncogenes on Survival of Canine MM

Recent studies have pointed out that complex clustered SVs in human MM or specific recurrent SVs (i.e., focal amplifications of CFA30) in canine MM are linked to poor outcomes [15,24]. To explore the impact of the SV profiles on canine MM, the SV profiles were determined using low-pass sequencing on 42 canine MM FFPE samples with available survival data (part of a previously published cohort by Prouteau et al.) [24]. In this cohort, the minimal region of CFA 30 amplification was delimited between 16.2 Mb and 16.8 Mb. This region was amplified in 64.3% of cases and contained *GABPB1*, *USP8*, *TRPM7*, and *SPPL2A* genes. While high SV content in canine MM is associated with a poor outcome (one-sided log-rank p-value = 0.025), a high SV is less significantly associated with a poor outcome than is the amplification of CFA30 candidate oncogenes (one-sided log-rank p-value = 0.0012) (Figure A7).

3.4. Exploring the Value of CFA 30 Oncogenes in Human Melanomas

The canine model of MM allowed us to identify new oncogenes; thus, we explored the involvement of these canine CFA 30 oncogenes in human cancers, more specifically, in melanomas, according to published data. Since *TRPM7* and *USP8* CRISPR inactivation has an impact on the majority of human cell lines (703/990 and 960/990 for TRPM7 and USP8, respectively), these genes are considered to be common essential genes (https://depmap.

org/, accessed on 1 July 2021). Moreover, a high expression of TRPM7, USP8, GABPB1, or SPPL2A was shown to be associated with shorter survival times in cutaneous melanomas (Log-rank TRPM7, p-value = 0.005; Log-rank USP8, p-value = 0.017; Log-rank GABPB1, p-value = 0.019; and Log-rank GABPB1, p-value = 0.022) (https://www.proteinatlas.org, accessed on 1 July 2021). Interestingly, *USP8* knockdown suppressed cell growth, survival, and migration in cutaneous melanoma [75]. In addition, *TRPM7* expression levels have been shown to be associated with an invasive behavior and metastatic potential in cutaneous melanoma cell line [76] and is also expected to act as a protector in both melanocyte physiology and in melanoma cells [77].

Human MM and acral melanomas are characterized by a higher SV content and a lower tumor mutational burden than UV-induced cutaneous melanomas; thus, they are more likely to present strong focal amplifications of the corresponding genes. In the Cancer Genome Atlas, we found one case of acral melanoma (TCGA-ER-A19T-01) harboring a strong focal amplification on chromosome 15 (orthologous to canine chromosome 30) encompassing the candidate oncogenes *SPPL2A*, *USP8*, *TRPM7*, and *GABPB1*. This amplification of 29 copies of these genes was associated with high gene overexpression (16.8-, 22.7-, and 30.2-fold for *TRPM7*, *USP8*, and *SPPL2A*, respectively). In another cohort of 34 patients with acral melanomas, we found one tumor with an intrachromosomal rearrangement on chromosome 15 targeting the genes *TRPM7* and *MYO5A* [78]. Furthermore, two recent studies involving MM cases that performed WGS on 65 and 67 tumor samples showed that chromosome 15 frequently had a deletion in the proximal part, and more rarely, amplifications targeting the candidate orthologous region identified in the present study [14,15]. Moreover, in a cohort of 65 patients with only oral melanoma, *TRPM7* was one of the 48 significantly mutated genes [15].

To confirm the importance of these genes in non-UV-induced melanomas in humans, we evaluated the proliferation capability of two human melanoma cell lines, one mucosal melanoma cell line (HMV-2/CVCL_1282), and one acral melanoma cell line (WM3211/CVCL_6797). Using siRNA to induce silencing of the *TRPM7*, *SPPL2A*, and *GABPB1* genes, we showed that their expression was decreased by 61–69% (Table S8), which also led to a significant decrease in cell proliferation from 10.8–27.4% (Figure A8).

These findings confirm that the candidate genes of the canine CFA30 16.2–16.8 Mb region/Human CFA15 50.3–50.9 Mb region are also involved in a subset of human non-UV-induced mucosal and acral melanomas. The results of this study suggest that the therapeutic target potential of these genes in human MM, especially those of the oral cavity, should be explored.

4. Discussion

In the last decade, dog models have been established as relevant models for clinical and genetic studies of human MM [13,23,24,79–81]. The present work provides a better understanding of the genomic and transcriptomic profiles associated with canine MM by identifying two molecular subgroups of canine patients differing in their transcriptomic profiles and SV content. Thus, these findings suggest different tumor microenvironments for each subgroup, each requiring different types of therapies. In addition, this work allowed the identification of new candidate oncogenes for MM, which should be of interest to human oncology.

The first molecular subgroup contained a majority of tumors with a relatively low SV content, with overexpression of genes related to the microenvironment, particularly the T-cell response and cytotoxic functions. Thus, the likely presence of an effective T-cell infiltrate in tumors in this group, in addition to the overexpression of immune checkpoint proteins, such as CTLA-4, TIM3, and LAG3, favors a response to immunotherapy, especially when using checkpoint inhibitors [52–54,82]. Interestingly, cell-type-specific enrichment analysis from our MM bulk RNAseq data with the xCell program [83] confirmed that group 1 samples contained an enrichment of immune cells, such as dendritic cells, macrophages M1 and lymphocytes T (Figure A9). While several studies have linked high tumor muta-

tional burden to a better response of immunotherapy [62,84,85], two recent publications have shown that CNA/SV levels are also predictive of a response to immunotherapy in different cancer types. Among these cancers, cutaneous melanomas harboring a high level of CNAs have the poorest response rate to immunotherapy [52,53]. Ock et al. also suggested that these alterations (CNA/SV) are stronger predictors of the response to immune checkpoints than tumor mutational burden [53]. Following these studies, we suggest that the first molecular subgroup with low SV identified in this study may respond to immunotherapy, while the other subgroup with higher SV rates may respond better to targeted therapies (e.g., anti-CDK4). This hypothesis is further reinforced by the results of a previous study by Ock et al., who observed that the differentially expressed genes between the two groups were enriched in genes associated with response to immunotherapy in human cancers according to [53]. We found similarities in terms of the expression modifications (up or downregulation) in the genes differing in expression between the two canine MM subgroups and the immunotherapy responders and non-responders in the study by Ock et al. ($p = 1.0 \times 10^{-11}$) [53] (Table S9). However, several genes associated with the acquisition of an invasive, dedifferentiated, EMT-like phenotype were also overexpressed in this group. While a favorable tumor immune microenvironment is critical for effective immunotherapy, tumor cells could exploit the dedifferentiation program to resist immunotherapy. Further studies combining therapeutic trials with single-cell RNAseq would be very relevant to better assess the benefit of immunotherapies in this group.

In a recent study, Newell et al. also identified two distinct subgroups of human MM based on the degree of localized complex rearrangements [14]. While, here, we classified MM using a transcriptomic-based strategy, we confirmed the existence of two MM subgroups with different SV profiles targeting similar driver genes (*MDM2*, *CDK4*). The association of different SV profiles between the two molecular subgroups of MM could reflect the involvement of different DNA repair mechanism alterations and, thus, different tumorigenesis pathways. The second subgroup was characterized by the presence of recurrent focal high amplifications and numerous chromosomal rearrangements ("high SV"). This chromosomal instability has already been described in human melanomas and is associated with poor outcomes in cutaneous and oral melanomas [15,86]. This aggressiveness was also observed in canine oral melanoma harboring focal amplification on CFA 30 (median survival time of 159 days vs. 317 days for MM with CFA 30 amplification or no CFA30 amplification, respectively, $p = 5 \times 10^{-5}$) [24], and we also demonstrated that dogs with MM carrying a high SV burden had a poorer prognosis (one-sided $p = 0.0255$) than those carrying a lower SV burden. While the molecular subgroup with high SV exhibits a higher mitotic index (mean = 21.8 mitosis vs. 9.8 for the high-SV group and the low-SV group, respectively, Mann–Whitney test, p-value = 0.022), further studies, including data on clinical staging, are needed to refine the correlation between the molecular subgroups and known clinical/histological prognostic factors.

Focal amplification in melanomas results from complex genomic catastrophes previously described as chromothripsis in humans [14,15,86] and dogs [29]; however, most events are too complex to confidently assign to one particular type of mutational mechanism [14]. The WGS of four canine MM cell lines showed that the focal amplifications had complex rearrangements reminiscent of chromothripsis, with chained or clustered breakpoints localized to a subset of chromosomes in regions that also contained copy number oscillations. The formation of double minutes is expected in chromothripsis to explain the high amplification of driver oncogenes. The FISH analysis in this study revealed large, homogeneously stained derivative chromosomes from CFA 10 and/or CFA 30, but no double-minute chromosomes. Similar patterns of high *MDM2*/*CDK4* amplification with large derivative chromosomes were observed in human glioblastoma, and were assigned to chromothripsis, with aggregation of double minute chromosomes [66]. In human acute myeloid leukemia, such catastrophic events have been linked to somatic *TP53* mutations and, thus, p53 dysfunction [87]. In dogs, approximately 50% of MM harbor an amplification of the *TP53* inhibitor, MDM2 [24], and this alteration is mutually exclusive with

TP53 mutations [29]. Similarly, inactivated p53 mutations [11–15,17,88–91] and *MDM2* amplifications [13–16] have been observed in human mucosal and acral melanoma, suggesting that p53 pathway dysregulation may be crucial in non-UV-induced melanoma development [29].

In human melanomas, recurrent focal amplifications are frequent (over 50%) and significantly less common in cutaneous melanomas. The focal amplification of 5p15 (*TERT* gene locus) is mainly observed in oral MM [92]; recurrent interchromosomal translocations between HSA 5 (*TERT*) and HSA 12 (*CDK4*) were observed in 39.6% of cases [15]. Similarly, over 50% of canine oral MM cases harbored recurrent focal amplification of *CDK4/MDM2* oncogenes. However, unlike human MM, these amplifications did not co-occur with *TERT* DNA amplification, even though TERT RNA was significantly over-expressed in the "high SV" cases (log2ratio = 2.54, p-corrected = 6.7×10^{-3}). These results suggest that focal amplifications of the *CDK4/MDM2* oncogenes in human and canine MM drive the same oncogenic pathway activation by distinct mechanisms. Promoter mutations enhance *TERT* expression in human cutaneous melanoma [93], and these mutations create a de novo binding site for the transcription factor *GABP* heterotetramer activating *TERT* transcription [94,95]. In dogs, *GABPB1* is overexpressed due to CFA 30 amplification; no *TERT* promoter mutations were observed in the four canine cell lines characterized by WGS. Similarly, Hendricks et al. (2018) did not find any *TERT* mutations in five canine MM cases analyzed using WGS to explain *TERT* overexpression. Other mechanisms could be involved in gene overexpression, such as epigenetic regulation or activation of the NFKB or JAK/STAT pathways [93,96,97].

Regarding the focal amplification of CFA 30 (16–17 Mb region), the corresponding amplification of HSA 15 has already been observed in few human melanomas [78,86] (https://www.cancer.gov/tcga, accessed on 1 July 2021), and was associated with overexpression of the amplified genes (https://www.cancer.gov/tcga, accessed on 1 July 2021). Nevertheless, based on the high recurrence of this region amplifications in canine MM (~50% [24]), the CFA 30 region should contain important oncogenes for non-UV-induced melanoma development. The following lines of evidence support this hypothesis: (i) in dogs, the degree of expression of candidate oncogenes of CFA 30 is correlated with the copy number; (ii) these genes appear to be essential in a majority of human cell lines (*TRPM7* and *USP8*); and (iii) they are associated with a shorter survival time in cutaneous melanomas. In the present study, we showed that *TRPM7*, *SPPL2A*, and *GABPB1* are involved in the proliferation of canine and human non-UV-induced melanomas. We hypothesized that genes included in the CFA 30 amplification have a cumulative effect on the tumoral transformation of melanocytes, and further studies are required to explore their roles, such as in tumor proliferation, adhesion, and migration [98]. Interestingly, a recent study using transcriptomic profiling of canine cancers identified *TRPM7* as a biomarker of melanoma and *SPPL2A* as a highly relevant target [99]. *TRPM7* thus appears to be an important gene for non-UV-induced melanomas, as it was recently found to be (i) involved in several other human cancers [98]; (ii) significantly mutated in human oral melanomas [15]; and (iii) involved in intrachromosomal translocation in acral melanoma in a previous study [78]. The importance of these genes for UV-induced melanomas has to be further explored by silencing or inactivating their expression in cutaneous cell lines in order to assess whether these genes could also be relevant therapeutic targets in UV-induced melanomas.

5. Conclusions

Domestic dogs are increasingly being considered as a genetic system for the study of human cancers with underlying genetic components, owing to the strong breed-related predisposition to cancers. In this study, we identified two molecular subgroups of canine oral MM, differing in their SV content and gene expression profiles (immune microenvironment vs. pigmentation pathway and oncogenes), which may correspond to different therapeutic options. Moreover, this model allowed us to identify several oncogenes that are relevant for MM development, particularly *TRPM7*, which was recently shown to be involved in rare

human non-UV-induced melanomas. These results suggest that it is important to permit the sub-setting of tumors to improve tumor stratification for clinical trials, which will likely have differing clinical outcomes, and to provide the optimum therapy depending on this classification for both canine and human medicine.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14020276/s1, Table S1: Tumor specimens that have been included in this study, Table S2: liste of genes link to NMF gene signature of cluster 1 and cluster 2, Table S3: Primers that have been included in this study and their applications, Table S4: Pathway enrichment analysis with gene signature of clusters 1 and 2., Table S5: SV identified on the four canine cell lines of OMM, Table S6: Description of SV on the four canine cell lines of OMM, Table S7: Mean expression of genes targeted by siRNA on the canine cell line Bear, Table S8: Mean expression of genes targeted by siRNA on the human WM3211 and HMV2 cell lines, Table S9: Expresssion level in canine OMM cluster 1 of genes involved in human immune signature score, Western blot of CDK4, MDM2, and ERK on the cell lines (Figure 3B). The whole blots and the densitometry readings of the band of interest are included.

Author Contributions: Conceptualization, A.P. (Anais Prouteau), C.A., T.D. and B.H.; methodology, L.B., D.G. and F.C.; validation, A.P. (Anais Prouteau), S.M., A.P. (Aline Primot), E.C., L.B., A.H., L.C., C.K., T.D. and B.H.; formal Analysis, A.P. (Anais Prouteau), E.C., S.C., J.A., C.H., T.D. and B.H.; investigation, A.P. (Anais Prouteau), T.D. and B.H.; resources, N.B. and K.L.-T.; data curation, S.M. and T.D.; writing—original draft preparation, A.P. (Anais Prouteau), C.A., T.D. and B.H.; writing—review and editing, S.C., J.A., K.L.-T., D.G.; supervision, T.D. and B.H.; project administration, C.A.; funding acquisition, C.A. and K.L.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CNRS and in the frame of the French Plan Cancer 2009–2013, by Aviesan/INSERM MTS 2012-08. A. Prouteau (salary and direct costs) was funded through the FHU CAMIn (CHU-Univ Rennes-Brittany region & Ligue Nationale contre le cancer). The sample collection performed through the French Cani-DNA BRC was funded by the CRB-Anim infrastructure ANR-11-INBS-0003 in the frame of the Investing for the Future Program (PIA1).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of COMITÉ NATIONAL DE RÉFLEXION ETHIQUE SUR L'EXPÉRIMENTATION ANIMALE (protocol code CE07-2020-11-CA and date of approval: 23 November 2020).

Informed Consent Statement: Not applicable.

Data Availability Statement: The RNASeq data are available in SRA under the BioProject PRJNA749900. WGS of canine cell lines are available at PRJNA779870 and low-pass FFPE fastq files are available under the SRA bioproject PRJNA780881. WES fastq files are available in the SRA under the bioproject PRJNA786469.

Acknowledgments: The authors thank the veterinarians, who collected samples from melanoma affected dogs, together with their clinical data, and who provided anatomopathological diagnoses, especially Olivier Albaric, Laetitia Dorso and Florian Chocteau (Laboniris, Oniris, Ecole Nationale Vétérinaire de Nantes, France), Marie-Odile SEMIN (LAPV, Amboise, France), Caroline Laprie (Vet-Histo, Marseille, France), Marie Lagadic (Idexx Alfort, France), and Frédérique Degorce-Rubiale (LAPVSO, Toulouse, France). We also thank the dog owners for their general health data and dog samples. Melanoma data were gathered through the Cani-DNA BRC (http://dog-genetics.genouest.org, accessed on 1 July 2021), which is part of the CRB-Anim PIA1 infra-structure, ANR-11-INBS-0003. The authors thank A. Fautrel (Platform H2P2 Biosit Rennes, France) for support with FFPE samples. The authors especially thank Patrick Devauchelle and Pauline de Fornel, (MICEN-Vet, Cretail, France) as well as Didier Lanore (Bordeaux) for their advice on veterinary oncology, and Marie-Dominique Galibert (IGDR, Rennes, France) for her interest and follow-up. We also warmly thank Clotilde de Brito, Laetitia Lagoutte, Annabelle Garand, Anne Sophie Guillory and Frederique Allais (IGDR, Rennes, France) for their input on canine oral melanomas, and Stéphane Dréano (IGDR, Rennes, France) for Sanger sequencing. The authors also thank Jaime Modiano (Colorado State University) who provided the 2 oral melanoma cell lines (Bear -CVCL_OD14 and CML10 CVCL_IZ11). We are most grateful to Biogenouest Genomics, the Human & Environmental Genomics

core facility of Rennes (Biosit/OSUR) for its technical support, the GenOuest Bioinformatic core facility (https://www.genouest.org, accessed on 1 July 2021) for storing sequencing data, hosting the Cani-DNA website, and for providing the computing infrastructure.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

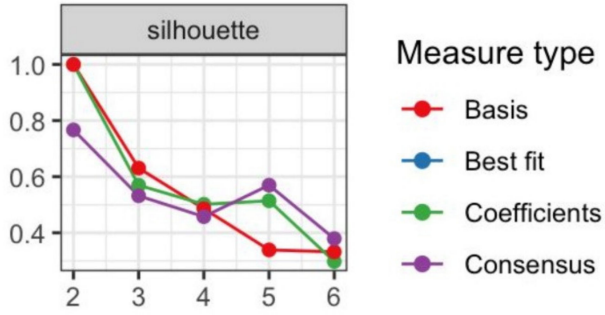

Figure A1. Estimation of the best K cluster using the silhouette method.

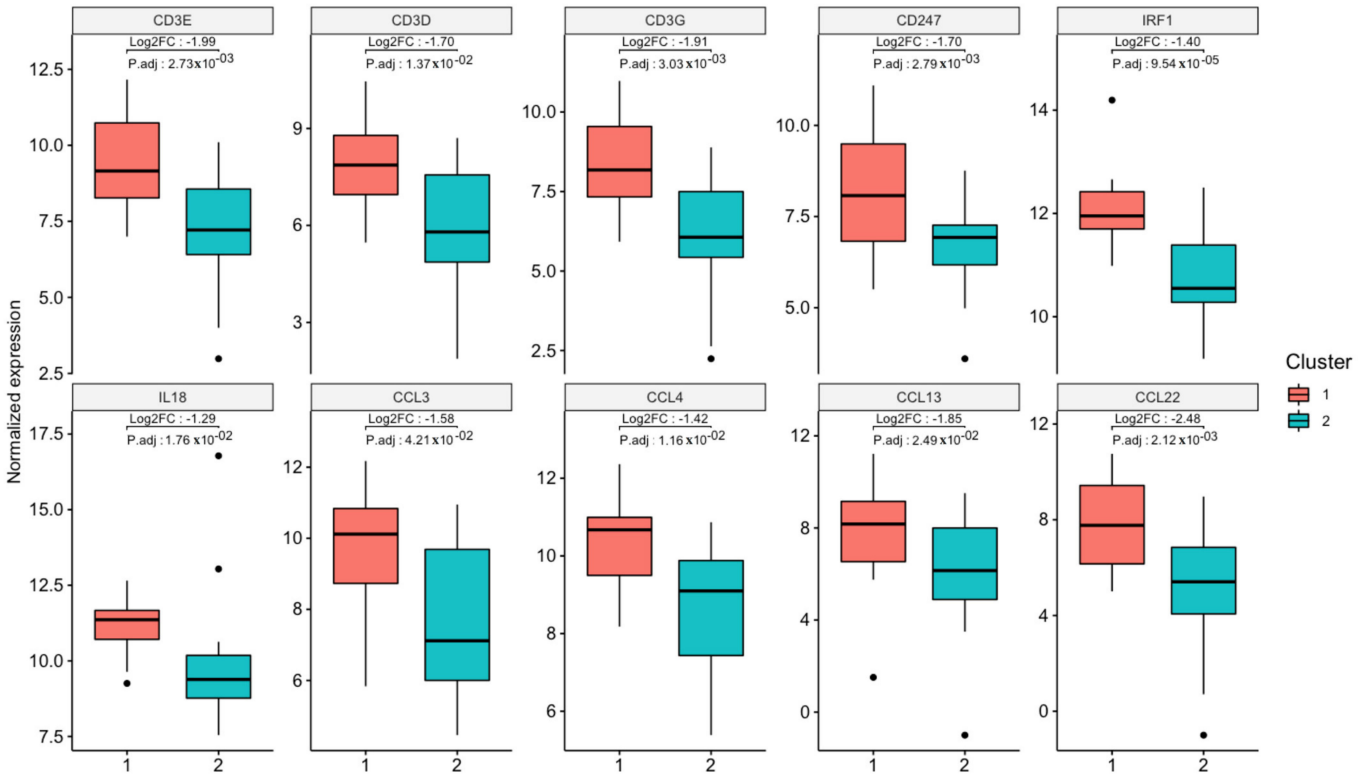

Figure A2. Expression of genes linked to infiltration of the immune cells according to the clustering groups. The following genes are overexpressed in group 1, reflecting significant immune cell infiltration in this group: CD3E, CD3D, CD3G, CD247, IRF1, IL18, CCL3, CCL4, CCL13 and CCL22.

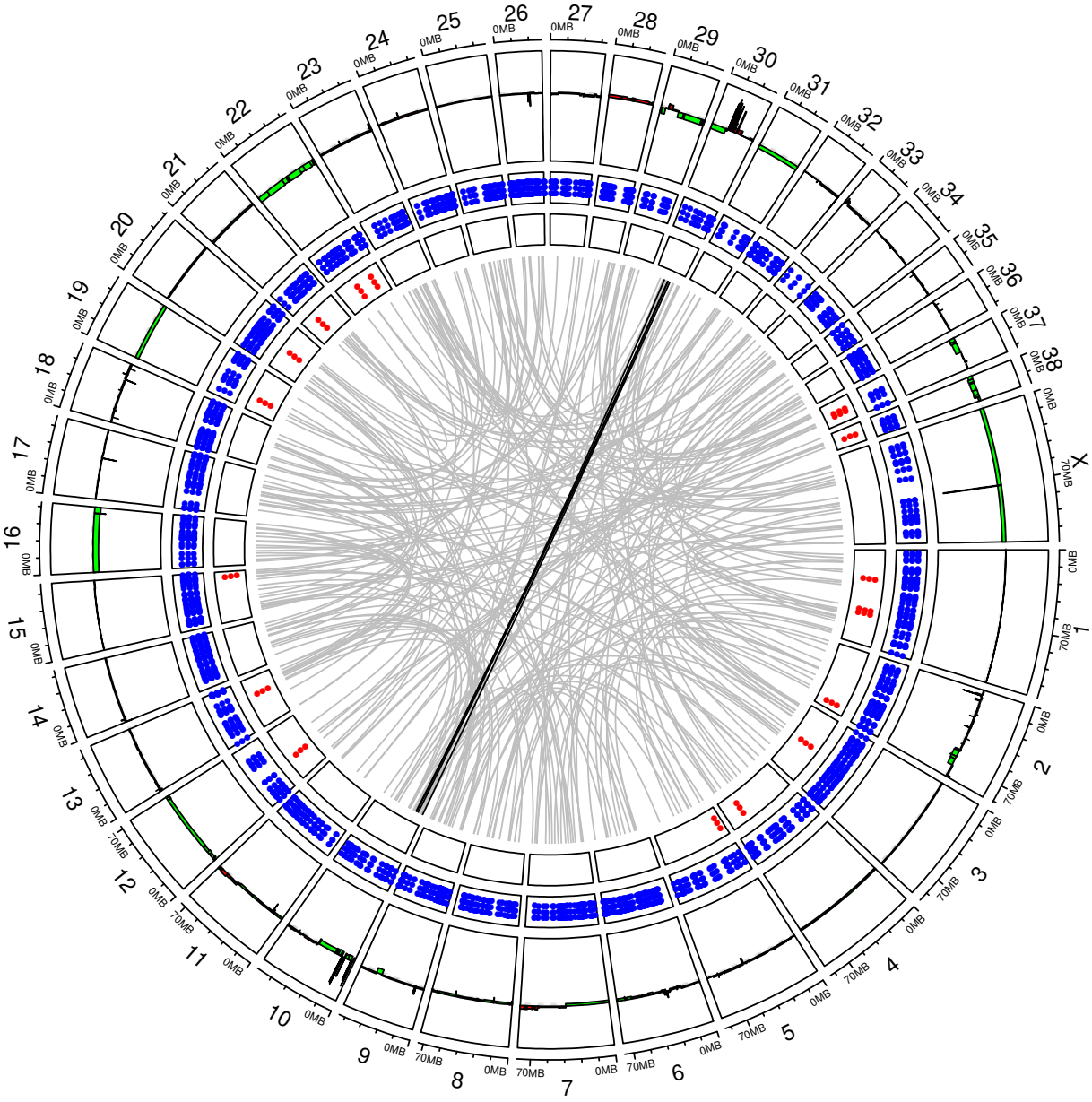

Figure A3. Circos plot representing the distribution of SVs of cell line Bear along dog chromosomes with, from external to internal layers, CNA gains/losses (in dark red/green), deletions, duplications, insertions, and inversions in blue, light red, orange, and green, respectively. Interchromosomal break-ends (BND) are represented by gray lines connecting chromosomes with a color intensity corresponding to the number of reads validating the SV.

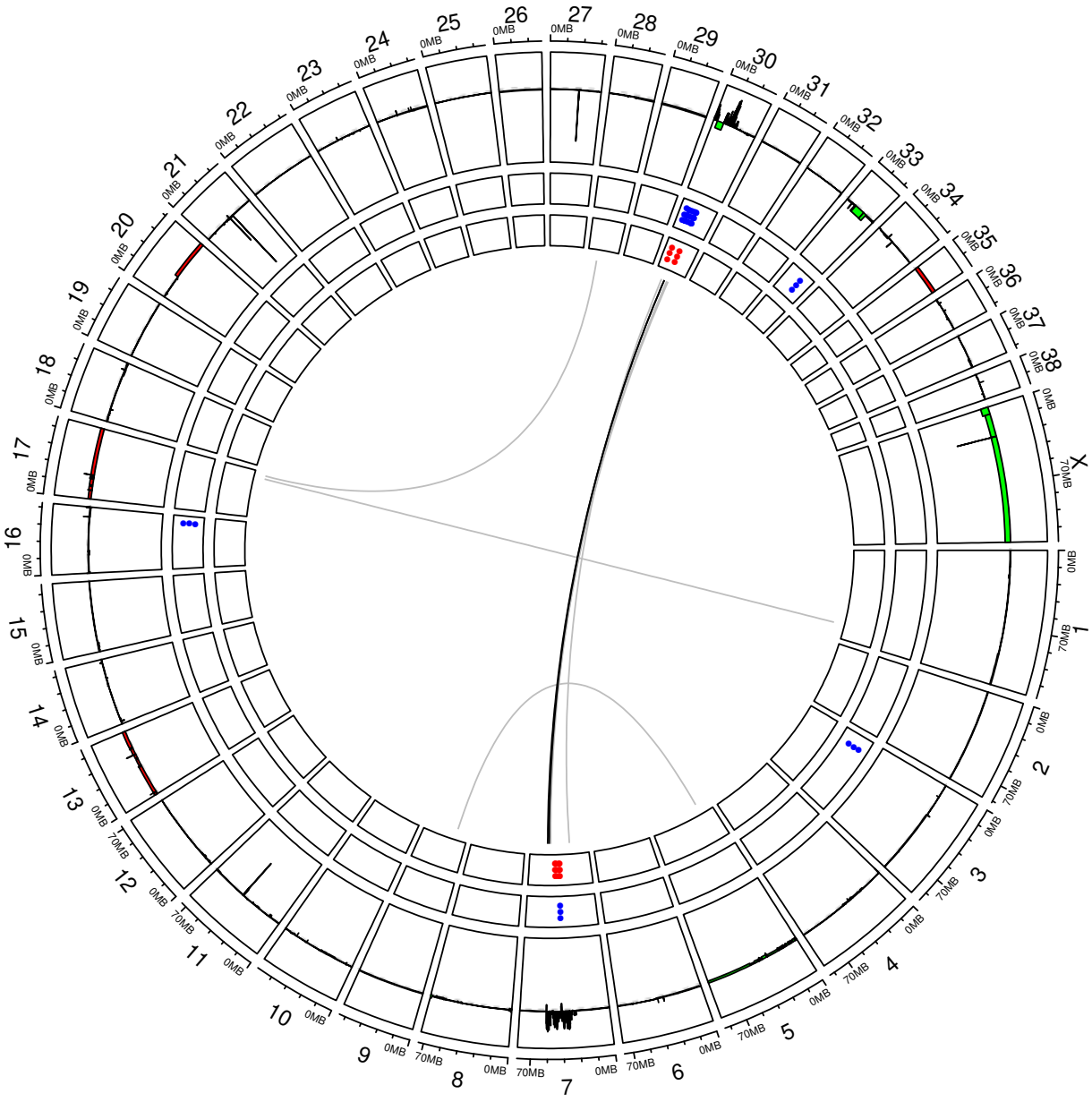

Figure A4. Circos plot representing the distribution of SVs of Dog-OralMel-18333 cell line along dog chromosomes with, from external to internal layers, CNA gains/losses (in dark red/green), deletions, duplications, insertions, and inversions in blue, light red, orange, and green, respectively. Interchromosomal break-ends (BND) are represented by gray lines connecting chromosomes with a color intensity corresponding to the number of reads validating the SV.

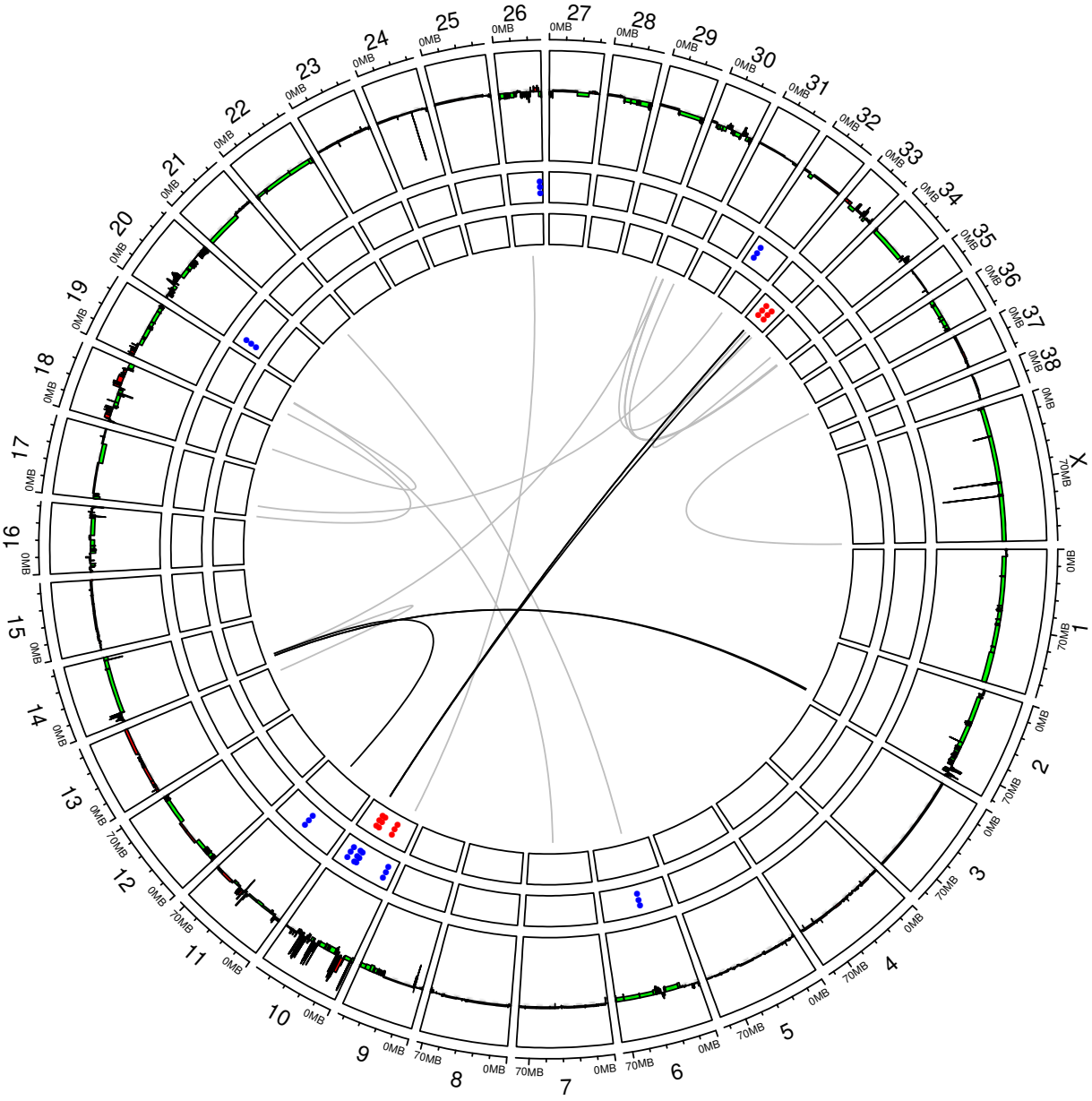

Figure A5. Circos plot representing the distribution of SVs of Dog-OralMel-18395 cell line along dog chromosomes with, from external to internal layers, CNA gains/losses (in dark red/green), deletions, duplications, insertions, and inversions in blue, light red, orange, and green, respectively. Interchromosomal break-ends (BND) are represented by gray lines connecting chromosomes with a color intensity corresponding to the number of reads validating the SV.

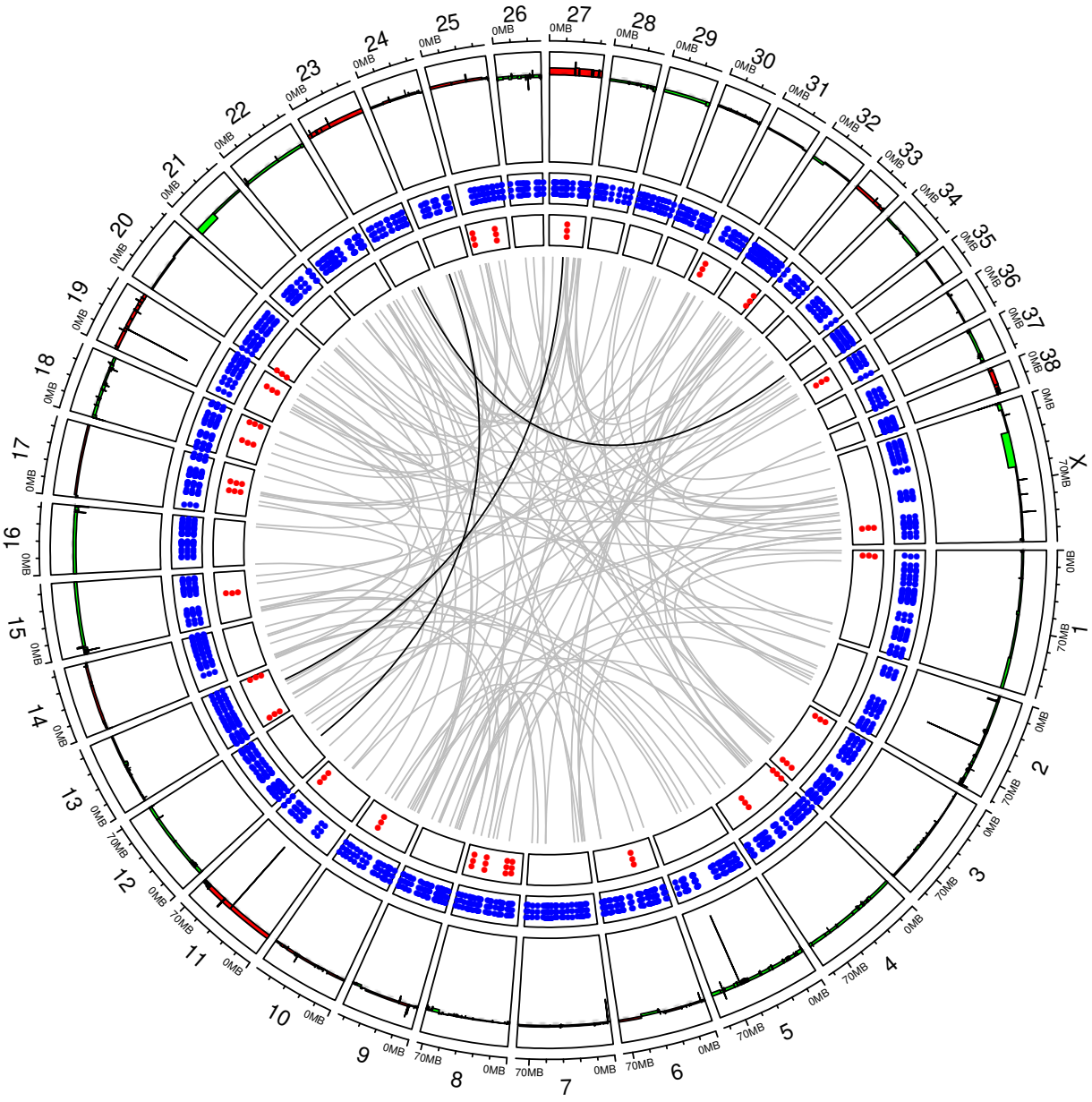

Figure A6. Circos plot representing the distribution of SVs of cell line CML10 along dog chromosomes with, from external to internal layers, CNA gains/losses (in dark red/green), deletions, duplications, insertions, and inversions in blue, light red, orange, and green, respectively. Interchromosomal break-ends (BND) are represented by gray lines connecting chromosomes with a color intensity corresponding to the number of reads validating the SV.

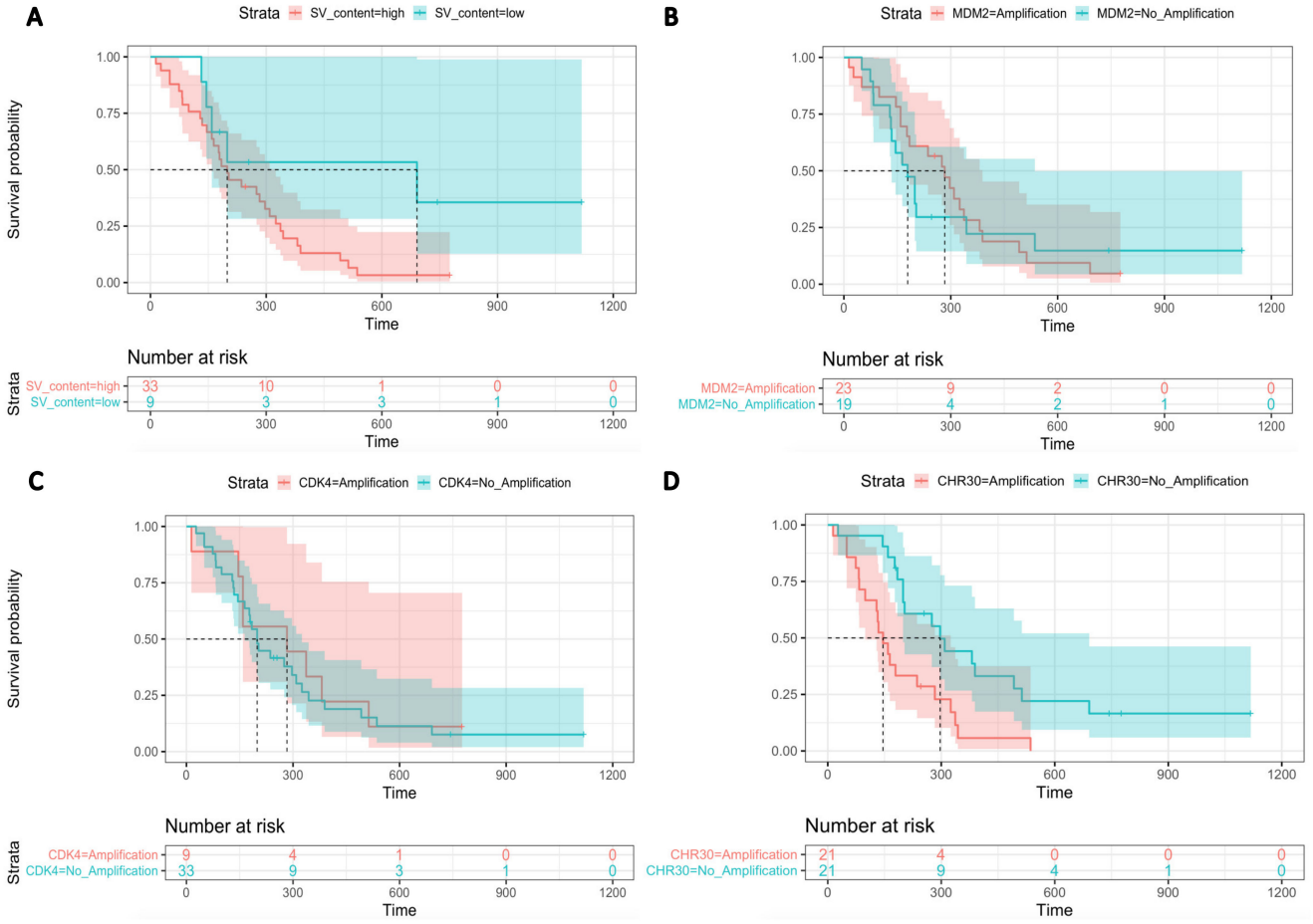

Figure A7. Kaplan Meier Survival Curves for death from MM according to SV profiles (**A**) (one sided log-rank $p = 0.025$), MDM2 amplification (**B**) (one sided log-rank $p = 0.39$), CDK4 amplification (**C**) (one sided log-rank $p = 0.34$) and CHR30 amplification (**D**) (one sided log-rank $p = 0.0012$).

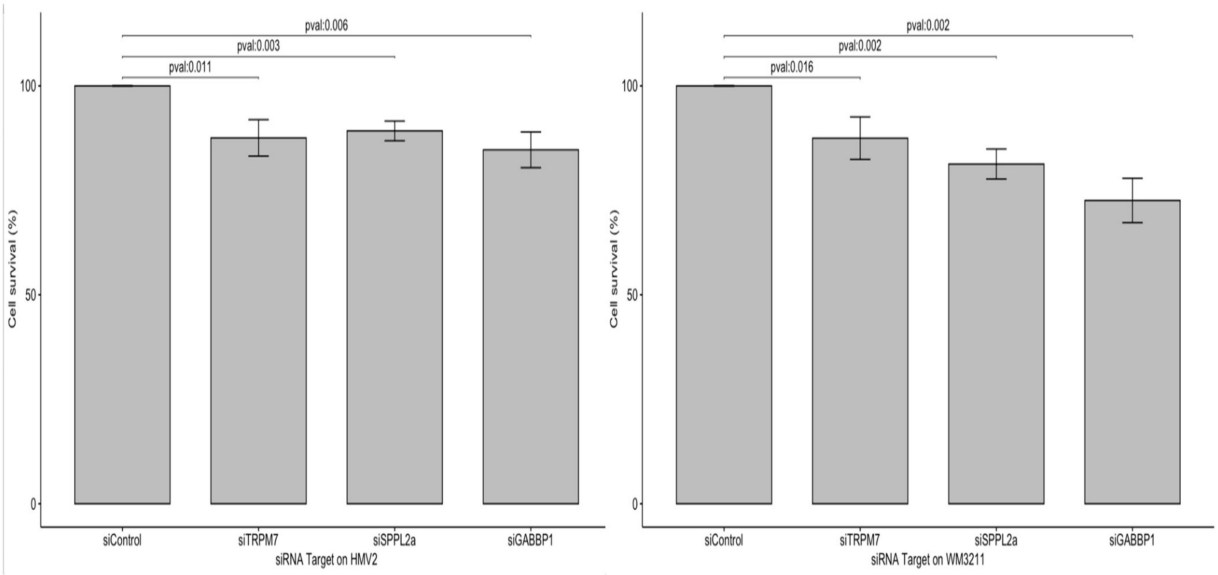

Figure A8. Cell proliferation assay on two human non-UV induced melanoma cell lines (HMV2 and WM3211) showing the effect of the knockdown (siRNA) of TRPM7, GABPB1 and SPPL2A genes on cell survival.

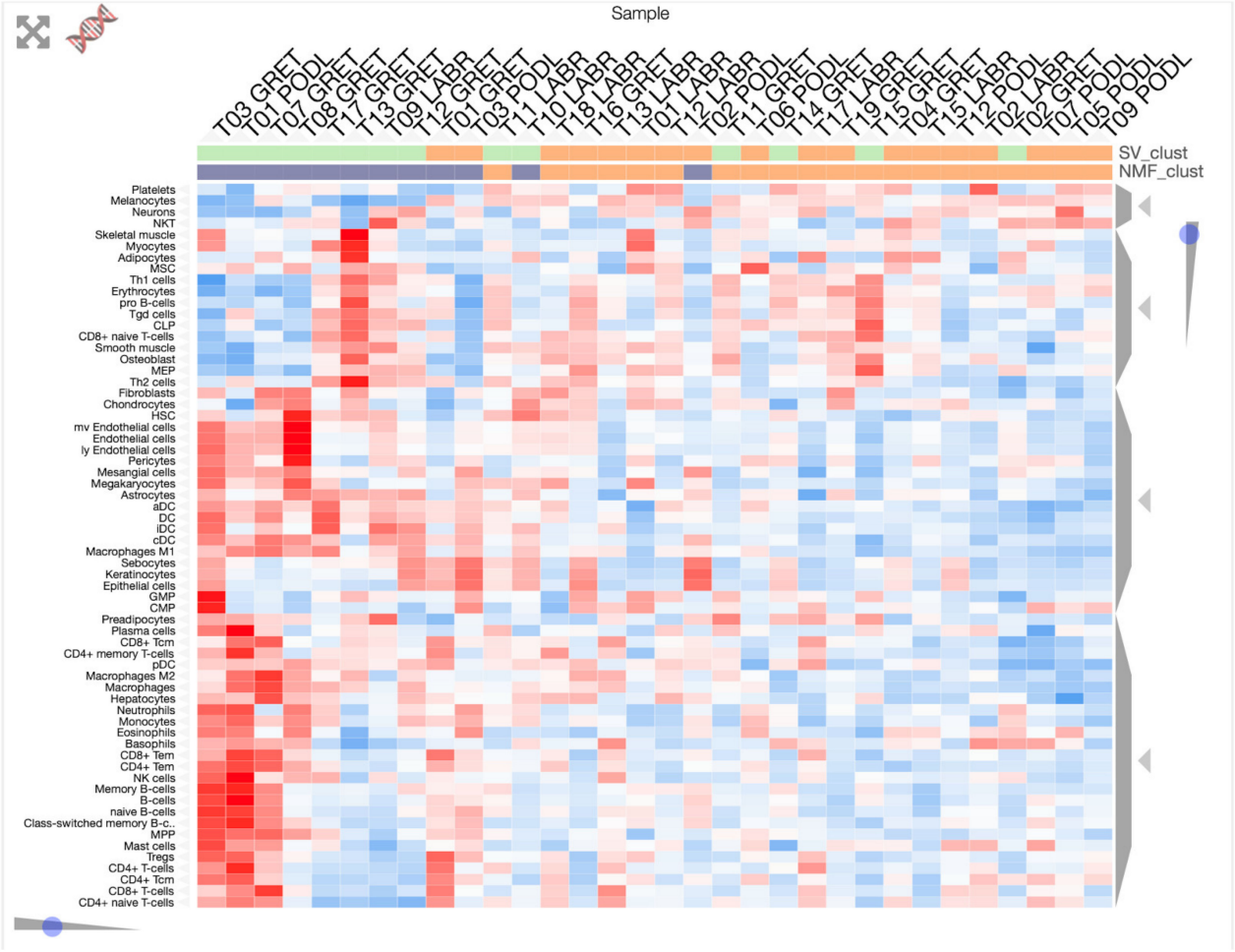

Figure A9. Cell types enrichment analysis in MM. Cell type enrichment analysis were performed with xCell program [83] with 64 immune and stroma cell types and MM are then clustered using clustergrammer [100]. Enrichment for immune cells such as dendritic cells, macrophages M1 and lymphocytes T is observed in the group 1 with low SV content (SV_clust = clustering of MM samples based on SV content with group1 in light green and group2 in orange, NMF_clust = clustering of MM samples based on NMF with RNASeq data with group1 in purple and group2 in orange).

References

1. Williams, M.D. Update from the 4th Edition of the World Health Organization Classification of Head and Neck Tumours: Mucosal Melanomas. *Head Neck Pathol.* **2017**, *11*, 110–117. [CrossRef] [PubMed]
2. Yde, S.S.; Sjoegren, P.; Heje, M.; Stolle, L.B. Mucosal Melanoma: A Literature Review. *Curr. Oncol. Rep.* **2018**, *20*, 28. [CrossRef]
3. Lerner, B.A.; Stewart, L.A.; Horowitz, D.P.; Carvajal, R.D. Mucosal Melanoma: New Insights and Therapeutic Options for a Unique and Aggressive Disease. *Oncology* **2017**, *31*, e23–e32. [PubMed]
4. Chi, Z.; Li, S.; Sheng, X.; Si, L.; Cui, C.; Han, M.; Guo, J. Clinical presentation, histology, and prognoses of malignant melanoma in ethnic Chinese: A study of 522 consecutive cases. *BMC Cancer* **2011**, *11*, 85. [CrossRef]
5. Kuk, D.; Shoushtari, A.N.; Barker, C.A.; Panageas, K.S.; Munhoz, R.R.; Momtaz, P.; Ariyan, C.E.; Brady, M.S.; Coit, D.G.; Bogatch, K.; et al. Prognosis of Mucosal, Uveal, Acral, Nonacral Cutaneous, and Unknown Primary Melanoma from the Time of First Metastasis. *Oncologist* **2016**, *21*, 848–854. [CrossRef]
6. Bishop, K.D.; Olszewski, A.J. Epidemiology and survival outcomes of ocular and mucosal melanomas: A population-based analysis. *Int. J. Cancer* **2014**, *134*, 2961–2971. [CrossRef] [PubMed]
7. Tacastacas, J.D.; Bray, J.; Cohen, Y.K.; Arbesman, J.; Kim, J.; Koon, H.B.; Honda, K.; Cooper, K.D.; Gerstenblith, M.R. Update on primary mucosal melanoma. *J. Am. Acad. Dermatol.* **2014**, *71*, 366–375. [CrossRef]
8. D'Angelo, S.P.; Hamid, O.A.; Tarhini, A.; Schadendorf, D.; Chmielowski, B.; Collichio, F.A.; Pavlick, A.C.; Lewis, K.D.; Weil, S.C.; Heyburn, J.; et al. A phase 2 study of ontuxizumab, a monoclonal antibody targeting endosialin, in metastatic melanoma. *Investig. New Drugs* **2018**, *36*, 103–113. [CrossRef] [PubMed]

9. Schaefer, A.; Sachpekidis, C.; Diella, F.; Doerks, A.; Kratz, A.-S.; Meisel, C.; Jackson, D.B.; Soldatos, T.G. Public Adverse Event Data Insights into the Safety of Pembrolizumab in Melanoma Patients. *Cancers* **2020**, *12*, 1008. [CrossRef]
10. Wang, Y.; Gu, T.; Tian, X.; Li, W.; Zhao, R.; Yang, W.; Gao, Q.; Li, T.; Shim, J.-H.; Zhang, C.; et al. A Small Molecule Antagonist of PD-1/PD-L1 Interactions Acts as an Immune Checkpoint Inhibitor for NSCLC and Melanoma Immunotherapy. *Front. Immunol.* **2021**, *12*, 654463. [CrossRef]
11. Furney, S.; Turajlic, S.; Stamp, G.; Thomas, J.M.; Hayes, A.; Strauss, D.; Gavrielides, M.; Xing, W.; Gore, M.; Larkin, J.; et al. The mutational burden of acral melanoma revealed by whole-genome sequencing and comparative analysis. *Pigment. Cell Melanoma Res.* **2014**, *27*, 835–838. [CrossRef]
12. Hayward, N.K.; Wilmott, J.S.; Waddell, N.; Johansson, P.A.; Field, M.A.; Nones, K.; Patch, A.-M.; Kakavand, H.; Alexandrov, L.B.; Burke, H.; et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **2017**, *545*, 175–180. [CrossRef] [PubMed]
13. Wong, K.; van der Weyden, L.; Schott, C.R.; Foote, A.; Constantino-Casas, F.; Smith, S.; Dobson, J.M.; Murchison, E.P.; Wu, H.; Yeh, I.; et al. Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nat. Commun.* **2019**, *10*. [CrossRef]
14. Newell, F.; Kong, Y.; Wilmott, J.S.; Johansson, P.A.; Ferguson, P.M.; Cui, C.; Li, Z.; Kazakoff, S.H.; Burke, H.; Dodds, T.J.; et al. Whole-genome landscape of mucosal melanoma reveals diverse drivers and therapeutic targets. *Nat. Commun.* **2019**, *10*. [CrossRef] [PubMed]
15. Zhou, R.; Shi, C.; Tao, W.; Li, J.; Wu, J.; Han, Y.; Yang, G.; Gu, Z.; Xu, S.; Wang, Y.; et al. Analysis of Mucosal Melanoma Whole-Genome Landscapes Reveals Clinically Relevant Genomic Aberrations. *Clin. Cancer Res.* **2019**, *25*, 3548–3560. [CrossRef]
16. Lyu, J.; Song, Z.; Chen, J.; Shepard, M.J.; Song, H.; Ren, G.; Li, Z.; Guo, W.; Zhuang, Z.; Shi, Y. Whole-exome sequencing of oral mucosal melanoma reveals mutational profile and therapeutic targets. *J. Pathol.* **2018**, *244*, 358–366. [CrossRef]
17. Hintzsche, J.D.; Gorden, N.T.; Amato, C.M.; Kim, J.; Wuensch, K.E.; Robinson, S.E.; Applegate, A.J.; Couts, K.L.; Medina, T.M.; Wells, K.R.; et al. Whole-exome sequencing identifies recurrent SF3B1 R625 mutation and comutation of NF1 and KIT in mucosal melanoma. *Melanoma Res.* **2017**, *27*, 189–199. [CrossRef]
18. Broit, N.; Johansson, P.A.; Rodgers, C.B.; Walpole, S.T.; Newell, F.; Hayward, N.K.; Pritchard, A.L. Meta-Analysis and Systematic Review of the Genomics of Mucosal Melanoma. *Mol. Cancer Res.* **2021**, *19*, 991–1004. [CrossRef] [PubMed]
19. LeBlanc, A.K.; Breen, M.; Choyke, P.; Dewhirst, M.; Fan, T.M.; Gustafson, D.L.; Helman, L.J.; Kastan, M.B.; Knapp, D.W.; Levin, W.J.; et al. Perspectives from man's best friend: National Academy of Medicine's Workshop on Comparative Oncology. *Sci. Transl. Med.* **2016**, *8*, 324ps5. [CrossRef]
20. Dewhirst, M.W.; Page, R.L. Editorial: Emerging Translational Opportunities in Comparative Oncology with Companion Canine Cancers. *Front. Oncol.* **2020**, *10*, 270. [CrossRef] [PubMed]
21. Ulvé, R.; Rault, M.; Bahin, M.; Lagoutte, L.; Abadie, J.; De Brito, C.; Coindre, J.-M.; Botherel, N.; Rousseau, A.; Wucher, V.; et al. Discovery of Human-Similar Gene Fusions in Canine Cancers. *Cancer Res.* **2017**, *77*, 5721–5727. [CrossRef] [PubMed]
22. Van Der Weyden, L.; Patton, E.E.; Wood, G.; Foote, A.K.; Brenn, T.; Arends, M.J.; Adams, D.J. Cross-species models of human melanoma. *J. Pathol.* **2015**, *238*, 152–165. [CrossRef]
23. Prouteau, A.; André, C. Canine Melanomas as Models for Human Melanomas: Clinical, Histological, and Genetic Comparison. *Genes* **2019**, *10*, 501. [CrossRef] [PubMed]
24. Prouteau, A.; Chocteau, F.; De Brito, C.; Cadieu, E.; Primot, A.; Botherel, N.; Degorce, F.; Cornevin, L.; Lagadic, M.A.; Cabillic, F.; et al. Prognostic value of somatic focal amplifications on chromosome 30 in canine oral melanoma. *Vet. Comp. Oncol.* **2019**, *18*, 214–223. [CrossRef] [PubMed]
25. Verganti, S.; Berlato, D.; Blackwood, L.; Amores-Fuster, I.; Polton, G.A.; Elders, R.; Doyle, R.; Taylor, A.; Murphy, S. Use of Oncept melanoma vaccine in 69 canine oral malignant melanomas in the UK. *J. Small Anim. Pract.* **2017**, *58*, 10–16. [CrossRef]
26. Igase, M.; Nemoto, Y.; Itamoto, K.; Tani, K.; Nakaichi, M.; Sakurai, M.; Sakai, Y.; Noguchi, S.; Kato, M.; Tsukui, T.; et al. A pilot clinical study of the therapeutic antibody against canine PD-1 for advanced spontaneous cancers in dogs. *Sci. Rep.* **2020**, *10*. [CrossRef]
27. Maekawa, N.; Konnai, S.; Takagi, S.; Kagawa, Y.; Okagawa, T.; Nishimori, A.; Ikebuchi, R.; Izumi, Y.; Deguchi, T.; Nakajima, C.; et al. A canine chimeric monoclonal antibody targeting PD-L1 and its clinical efficacy in canine oral malignant melanoma or undifferentiated sarcoma. *Sci. Rep.* **2017**, *7*, 8951. [CrossRef]
28. Maekawa, N.; Konnai, S.; Nishimura, M.; Kagawa, Y.; Takagi, S.; Hosoya, K.; Ohta, H.; Kim, S.; Okagawa, T.; Izumi, Y.; et al. PD-L1 immunohistochemistry for canine cancers and clinical benefit of anti-PD-L1 antibody in dogs with pulmonary metastatic oral malignant melanoma. *NPJ Precis. Oncol.* **2021**, *5*, 10. [CrossRef]
29. Hendricks, W.P.D.; Zismann, V.; Sivaprakasam, K.; Legendre, C.; Poorman, K.; Tembe, W.; Perdigones, N.; Kiefer, J.; Liang, W.; DeLuca, V.; et al. Somatic inactivating PTPRJ mutations and dysregulated pathways identified in canine malignant melanoma by integrated comparative genomic analysis. *PLoS Genet.* **2018**, *14*, e1007589. [CrossRef]
30. Fowles, J.S.; Denton, C.L.; Gustafson, D.L. Comparative analysis of MAPK and PI3K/AKT pathway activation and inhibition in human and canine melanoma. *Vet. Comp. Oncol.* **2013**, *13*, 288–304. [CrossRef]
31. Rahman, M.; Lai, Y.; Husna, A.A.; Chen, H.; Tanaka, Y.; Kawaguchi, H.; Hatai, H.; Miyoshi, N.; Nakagawa, T.; Fukushima, R.; et al. Transcriptome analysis of dog oral melanoma and its oncogenic analogy with human melanoma. *Oncol. Rep.* **2019**, *43*, 16–30. [CrossRef]

32. Di Palma, S.; McConnell, A.; Verganti, S.; Starkey, M. Review on Canine Oral Melanoma: An Undervalued Authentic Genetic Model of Human Oral Melanoma? *Vet. Pathol.* **2021**, *58*, 881–889. [CrossRef] [PubMed]
33. Hitte, C.; Le Béguec, C.; Cadieu, E.; Wucher, V.; Primot, A.; Prouteau, A.; Botherel, N.; Hédan, B.; Lindblad-Toh, K.; André, C.; et al. Genome-Wide Analysis of Long Non-Coding RNA Profiles in Canine Oral Melanomas. *Genes* **2019**, *10*, 477. [CrossRef] [PubMed]
34. Hédan, B.; Thomas, R.; Motsinger-Reif, A.; Abadie, J.; André, C.; Cullen, J.; Breen, M. Molecular cytogenetic characterization of canine histiocytic sarcoma: A spontaneous model for human histiocytic cancer identifies deletion of tumor suppressor genes and highlights influence of genetic background on tumor behavior. *BMC Cancer* **2011**, *11*, 201. [CrossRef]
35. Wucher, V.; Legeai, F.; Hédan, B.; Rizk, G.; Lagoutte, L.; Leeb, T.; Jagannathan, V.; Cadieu, E.; David, A.; Lohi, H.; et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **2017**, *45*, e57. [CrossRef]
36. Hoeppner, M.P.; Lundquist, A.; Pirun, M.; Meadows, J.; Zamani, N.; Johnson, J.; Sundström, G.; Cook, A.; Fitzgerald, M.G.; Swofford, R.; et al. An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts. *PLoS ONE* **2014**, *9*, e91172. [CrossRef] [PubMed]
37. Djebali, S.; Wucher, V.; Foissac, S.; Hitte, C.; Corre, E.; Derrien, T. Bioinformatics Pipeline for Transcriptome Sequencing Analysis. In *Advanced Structural Safety Studies*; Springer: Singapore, 2017; Volume 1468, pp. 201–219.
38. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
39. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
40. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
41. Gaujoux, R.; Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **2010**, *11*, 367. [CrossRef]
42. Kolberg, L.; Raudvere, U.; Kuzmin, I.; Vilo, J.; Peterson, H. gprofiler2—An R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **2020**, *9*, 709. [CrossRef]
43. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
44. Patch, A.-M.; Nones, K.; Kazakoff, S.H.; Newell, F.; Wood, S.; Leonard, C.; Holmes, O.; Xu, Q.; Addala, V.; Creaney, J.; et al. Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. *PLoS ONE* **2018**, *13*, e0190264. [CrossRef]
45. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
46. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.* **2016**, *17*, 122. [CrossRef]
47. Seshan, V.E.; Olshen, A. *DNAcopy: DNA Copy Number Data Analysis*, version 1.68.0; R Package: Madison, WI, USA, 2016.
48. Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A.M.; Benes, V.; Korbel, J. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, i333–i339. [CrossRef]
49. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [CrossRef]
50. Korbel, J.O.; Campbell, P.J. Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell* **2013**, *152*, 1226–1236. [CrossRef]
51. Govind, S.K.; Zia, A.; Hennings-Yeomans, P.H.; Watson, J.D.; Fraser, M.; Anghel, C.; Wyatt, A.W.; van der Kwast, T.; Collins, C.C.; McPherson, J.D.; et al. ShatterProof: Operational detection and quantification of chromothripsis. *BMC Bioinform.* **2014**, *15*, 78. [CrossRef] [PubMed]
52. Davoli, T.; Uno, H.; Wooten, E.C.; Elledge, S.J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **2017**, *355*. [CrossRef]
53. Ock, C.-Y.; Hwang, J.-E.; Keam, B.; Kim, S.-B.; Shim, J.-J.; Jang, H.-J.; Park, S.; Sohn, B.H.; Chan-Young, O.; Ajani, J.A.; et al. Genomic landscape associated with potential response to anti-CTLA-4 treatment in cancers. *Nat. Commun.* **2017**, *8*, 1050. [CrossRef]
54. Galon, J.; Bruni, D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* **2019**, *18*, 197–218. [CrossRef] [PubMed]
55. Kemper, K.; de Goeje, P.; Peeper, D.S.; Van Amerongen, R. Phenotype Switching: Tumor Cell Plasticity as a Resistance Mechanism and Target for Therapy. *Cancer Res.* **2014**, *74*, 5937–5941. [CrossRef]
56. Hoek, K.S.; Goding, C.R. Cancer stem cells versus phenotype-switching in melanoma. *Pigment Cell Melanoma Res.* **2010**, *23*, 746–759. [CrossRef] [PubMed]
57. Tsoi, J.; Robert, L.; Paraiso, K.; Galvan, C.; Sheu, K.M.; Lay, J.; Wong, D.J.; Atefi, M.; Shirazi, R.; Wang, X.; et al. Multi-stage Differentiation Defines Melanoma Subtypes with Differential Vulnerability to Drug-Induced Iron-Dependent Oxidative Stress. *Cancer Cell* **2018**, *33*, 890–904. [CrossRef]
58. Rambow, F.; Rogiers, A.; Marin-Bejar, O.; Aibar, S.; Femel, J.; Dewaele, M.; Karras, P.; Brown, D.; Chang, Y.H.; Debiec-Rychter, M.; et al. Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* **2018**, *174*, 843–855.e19. [CrossRef] [PubMed]

59. Vandamme, N.; Denecker, G.; Bruneel, K.; Blancke, G.; Akay, Ö.; Taminau, J.; De Coninck, J.; De Smedt, E.; Skrypek, N.; Van Loocke, W.; et al. The EMT Transcription Factor ZEB2 Promotes Proliferation of Primary and Metastatic Melanoma While Suppressing an Invasive, Mesenchymal-Like Phenotype. *Cancer Res.* **2020**, *80*, 2983–2995. [CrossRef]
60. Corre, S.; Tardif, N.; Mouchet, N.; LeClair, H.M.; Boussemart, L.; Gautron, A.; Bachelot, L.; Perrot, A.; Soshilov, A.; Rogiers, A.; et al. Sustained activation of the Aryl hydrocarbon Receptor transcription factor promotes resistance to BRAF-inhibitors in melanoma. *Nat. Commun.* **2018**, *9*, 4775. [CrossRef] [PubMed]
61. Verfaillie, A.; Imrichova, H.; Atak, Z.K.; Dewaele, M.; Rambow, F.; Hulselmans, G.; Christiaens, V.; Svetlichnyy, D.; Luciani, F.; Mooter, L.L.V.D.; et al. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat. Commun.* **2015**, *6*, 6683. [CrossRef] [PubMed]
62. Keenan, T.E.; Burke, K.P.; Van Allen, E.M. Genomic correlates of response to immune checkpoint blockade. *Nat. Med.* **2019**, *25*, 389–402. [CrossRef]
63. Poorman, K.; Borst, L.; Moroff, S.; Roy, S.; Labelle, P.; Motsinger-Reif, A.; Breen, M. Comparative cytogenetic characterization of primary canine melanocytic lesions using array CGH and fluorescence in situ hybridization. *Chromosom. Res.* **2014**, *23*, 171–186. [CrossRef]
64. Brocca, G.; Ferraresso, S.; Zamboni, C.; Martinez-Merlo, E.M.; Ferro, S.; Goldschmidt, M.H.; Castagnaro, M. Array Comparative Genomic Hybridization Analysis Reveals Significantly Enriched Pathways in Canine Oral Melanoma. *Front. Oncol.* **2019**, *9*, 1397. [CrossRef]
65. Rode, A.; Maass, K.K.; Willmund, K.V.; Ernst, A.; Lichter, P. Chromothripsis in cancer cells: An update. *Int. J. Cancer* **2015**, *138*, 2322–2333. [CrossRef] [PubMed]
66. Furgason, J.M.; Koncar, R.F.; Michelhaugh, S.K.; Sarkar, F.H.; Mittal, S.; Sloan, A.E.; Barnholtz-Sloan, J.; Bahassi, E.M. Whole genome sequence analysis links chromothripsis to EGFR, MDM2, MDM4, and CDK4 amplification in glioblastoma. *Oncoscience* **2015**, *2*, 618–628. [CrossRef] [PubMed]
67. Liu, K.; Xu, S.-H.; Chen, Z.; Zeng, Q.-X.; Li, Z.-J.; Chen, Z.-M. TRPM7 overexpression enhances the cancer stem cell-like and metastatic phenotypes of lung cancer through modulation of the Hsp90α/uPA/MMP2 signaling pathway. *BMC Cancer* **2018**, *18*, 1–12. [CrossRef] [PubMed]
68. Liu, L.; Wu, N.; Wang, Y.; Zhang, X.; Xia, B.; Tang, J.; Cai, J.; Zhao, Z.; Liao, Q.; Wang, J. TRPM7 promotes the epithelial–mesenchymal transition in ovarian cancer through the calcium-related PI3K/AKT oncogenic signaling. *J. Exp. Clin. Cancer Res.* **2019**, *38*, 1–15. [CrossRef]
69. Yee, N.S. Role of TRPM7 in Cancer: Potential as Molecular Biomarker and Therapeutic Target. *Pharmaceuticals* **2017**, *10*, 39. [CrossRef]
70. Meng, X.; Cai, C.; Wu, J.; Cai, S.; Ye, C.; Chen, H.; Yang, Z.; Zeng, H.; Shen, Q.; Zou, F. TRPM7 mediates breast cancer cell migration and invasion through the MAPK pathway. *Cancer Lett.* **2013**, *333*, 96–102. [CrossRef]
71. Chen, S.-C.; Yen, M.-C.; Chen, F.-W.; Wu, L.-Y.; Yang, S.-J.; Kuo, P.-L.; Hsu, Y.-L. Knockdown of GA-binding protein subunit β1 inhibits cell proliferation via p21 induction in renal cell carcinoma. *Int. J. Oncol.* **2018**, *53*, 886–894. [CrossRef]
72. Shin, S.; Kim, K.; Kim, H.-R.; Ylaya, K.; Do, S.-I.; Hewitt, S.M.; Park, H.-S.; Roe, J.-S.; Chung, J.-Y.; Song, J. Deubiquitylation and stabilization of Notch1 intracellular domain by ubiquitin-specific protease 8 enhance tumorigenesis in breast cancer. *Cell Death Differ.* **2020**, *27*, 1341–1354. [CrossRef]
73. Yan, M.; Zhao, C.; Wei, N.; Wu, X.; Cui, J.; Xing, Y. High Expression of Ubiquitin-Specific Protease 8 (USP8) Is Associated with Poor Prognosis in Patients with Cervical Squamous Cell Carcinoma. *Med Sci. Monit.* **2018**, *24*, 4934–4943. [CrossRef] [PubMed]
74. Kong, X.-F.; Martinez-Barricarte, R.; Kennedy, J.; Mele, F.; Lazarov, T.; Deenick, E.K.; Ma, C.; Breton, G.; Lucero, K.B.; Langlais, D.; et al. Disruption of an antimycobacterial circuit between dendritic and helper T cells in human SPPL2a deficiency. *Nat. Immunol.* **2018**, *19*, 973–985. [CrossRef]
75. Duan, B.; Wang, C.; Liu, Z.; Yang, X. USP8 is a Novel Therapeutic Target in Melanoma Through Regulating Receptor Tyrosine Kinase Levels. *Cancer Manag. Res.* **2021**, *13*, 4181–4189. [CrossRef]
76. McNeill, M.S.; Paulsen, J.; Bonde, G.; Burnight, E.; Hsu, M.-Y.; Cornell, R.A. Cell Death of Melanophores in Zebrafish trpm7 Mutant Embryos Depends on Melanin Synthesis. *J. Investig. Dermatol.* **2007**, *127*, 2020–2030. [CrossRef]
77. Guo, H.; Carlson, J.A.; Slominski, A. Role of TRPM in melanocytes and melanoma. *Exp. Dermatol.* **2012**, *21*, 650–654. [CrossRef] [PubMed]
78. Liang, W.S.; Hendricks, W.; Kiefer, J.; Schmidt, J.; Sekar, S.; Carpten, J.; Craig, D.W.; Adkins, J.; Cuyugan, L.; Manojlovic, Z.; et al. Integrated genomic analyses reveal frequent TERT aberrations in acral melanoma. *Genome Res.* **2017**, *27*, 524–532. [CrossRef]
79. Hernandez, B.; Adissu, H.A.; Wei, B.-R.; Michael, H.T.; Merlino, G.; Simpson, R.M. Naturally Occurring Canine Melanoma as a Predictive Comparative Oncology Model for Human Mucosal and Other Triple Wild-Type Melanomas. *Int. J. Mol. Sci.* **2018**, *19*, 394. [CrossRef]
80. Simpson, R.M.; Bastian, B.; Michael, H.; Webster, J.D.; Prasad, M.L.; Conway, C.M.; Prieto, V.M.; Gary, J.M.; Goldschmidt, M.; Esplin, D.G.; et al. Sporadic naturally occurring melanoma in dogs as a preclinical model for human melanoma. *Pigment. Cell Melanoma Res.* **2014**, *27*, 37–47. [CrossRef] [PubMed]
81. Gillard, M.; Cadieu, E.; De Brito, C.; Abadie, J.; Vergier, B.; Devauchelle, P.; Degorce, F.; Dréano, S.; Primot, A.; Dorso, L.; et al. Naturally occurring melanomas in dogs as models for non-UV pathways of human melanomas. *Pigment. Cell Melanoma Res.* **2014**, *27*, 90–102. [CrossRef]

82. Sweis, R.; Spranger, S.; Bao, R.; Paner, G.P.; Stadler, W.M.; Steinberg, G.; Gajewski, T.F. Molecular Drivers of the Non–T-cell-Inflamed Tumor Microenvironment in Urothelial Bladder Cancer. *Cancer Immunol. Res.* **2016**, *4*, 563–568. [CrossRef]
83. Aran, D.; Hu, Z.; Butte, A.J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **2017**, *18*, 220. [CrossRef]
84. Chan, T.; Yarchoan, M.; Jaffee, E.; Swanton, C.; Quezada, S.; Stenzinger, A.; Peters, S. Development of tumor mutation burden as an immunotherapy biomarker: Utility for the oncology clinic. *Ann. Oncol.* **2019**, *30*, 44–56. [CrossRef] [PubMed]
85. Goodman, A.M.; Kato, S.; Bazhenova, L.; Patel, S.P.; Frampton, G.M.; Miller, V.; Stephens, P.J.; Daniels, G.A.; Kurzrock, R. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **2017**, *16*, 2598–2608. [CrossRef]
86. Hirsch, D.; Kemmerling, R.; Davis, S.; Camps, J.; Meltzer, P.S.; Ried, T.; Gaiser, T. Chromothripsis and Focal Copy Number Alterations Determine Poor Outcome in Malignant Melanoma. *Cancer Res.* **2013**, *73*, 1454–1460. [CrossRef] [PubMed]
87. Rücker, F.G.; Dolnik, A.; Blätte, T.J.; Teleanu, V.; Ernst, A.; Thol, F.; Heuser, M.; Ganser, A.; Döhner, H.; Döhner, K.; et al. Chromothripsis is linked to TP53 alteration, cell cycle impairment, and dismal outcome in acute myeloid leukemia with complex karyotype. *Haematologica* **2018**, *103*, e17–e20. [CrossRef]
88. Hou, J.Y.; Baptiste, C.; Mbbs, R.B.H.; Tergas, A.I.; Feldman, R.; Jones, N.L.; Chatterjee-Paer, S.; Bus-Kwolfski, A.; Wright, J.D.; Burke, W.M. Vulvar and vaginal melanoma: A unique subclass of mucosal melanoma based on a comprehensive molecular analysis of 51 cases compared with 2253 cases of nongynecologic melanoma. *Cancer* **2017**, *123*, 1333–1344. [CrossRef]
89. Sheng, X.; Kong, Y.; Li, Y.; Zhang, Q.; Si, L.; Cui, C.; Chi, Z.; Tang, B.; Mao, L.; Lian, B.; et al. GNAQ and GNA11 mutations occur in 9.5% of mucosal melanoma and are associated with poor prognosis. *Eur. J. Cancer* **2016**, *65*, 156–163. [CrossRef]
90. Yeh, I.; Jorgenson, E.; Shen, L.; Xu, M.; North, J.P.; Shain, A.H.; Reuss, D.; Wu, H.; Robinson, W.; Olshen, A.; et al. Targeted Genomic Profiling of Acral Melanoma. *JNCI J. Natl. Cancer Inst.* **2019**, *111*, 1068–1077. [CrossRef] [PubMed]
91. Forschner, A.; Hilke, F.-J.; Bonzheim, I.; Gschwind, A.; Demidov, G.; Amaral, T.; Ossowski, S.; Riess, O.; Schroeder, C.; Martus, P.; et al. MDM2, MDM4 and EGFR Amplifications and Hyperprogression in Metastatic Acral and Mucosal Melanoma. *Cancers* **2020**, *12*, 540. [CrossRef] [PubMed]
92. Jurmeister, P.; Wrede, N.; Hoffmann, I.; Vollbrecht, C.; Heim, D.; Hummel, M.; Wolkenstein, P.; Koch, I.; Heynol, V.; Schmitt, W.D.; et al. Mucosal melanomas of different anatomic sites share a common global DNA methylation profile with cutaneous melanoma but show location-dependent patterns of genetic and epigenetic alterations. *J. Pathol.* **2021**. [CrossRef]
93. Akincilar, S.C.; Unal, B.; Tergaonkar, V. Reactivation of telomerase in cancer. *Cell. Mol. Life Sci.* **2016**, *73*, 1659–1670. [CrossRef]
94. Bell, R.J.A.; Rube, H.T.; Kreig, A.; Mancini, A.; Fouse, S.D.; Nagarajan, R.P.; Choi, S.; Hong, C.; He, D.; Pekmezci, M.; et al. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **2015**, *348*, 1036–1039. [CrossRef] [PubMed]
95. Mancini, A.; Xavier-Magalhães, A.; Woods, W.S.; Nguyen, K.-T.; Amen, A.M.; Hayes, J.L.; Fellmann, C.; Gapinske, M.; McKinney, A.M.; Hong, C.; et al. Disruption of the β1L Isoform of GABP Reverses Glioblastoma Replicative Immortality in a TERT Promoter Mutation-Dependent Manner. *Cancer Cell* **2018**, *34*, 513–528.e8. [CrossRef] [PubMed]
96. Barthel, F.P.; Wei, W.; Tang, M.; Martinez-Ledesma, E.; Hu, X.; Amin, S.B.; Akdemir, K.C.; Seth, S.; Song, X.; Wang, Q.; et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **2017**, *49*, 349–357. [CrossRef]
97. Yuan, X.; Larsson, C.; Xu, D. Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: Old actors and new players. *Oncogene* **2019**, *38*, 6172–6183. [CrossRef]
98. Meng, S.; Alanazi, R.; Ji, D.; Bandura, J.; Luo, Z.-W.; Fleig, A.; Feng, Z.-P.; Sun, H.-S. Role of TRPM7 kinase in cancer. *Cell Calcium* **2021**, *96*, 102400. [CrossRef] [PubMed]
99. Tawa, G.J.; Braisted, J.; Gerhold, D.; Grewal, G.; Mazcko, C.; Breen, M.; Sittampalam, G.; LeBlanc, A.K. Transcriptomic profiling in canines and humans reveals cancer specific gene modules and biological mechanisms common to both species. *PLoS Comput. Biol.* **2021**, *17*, e1009450. [CrossRef] [PubMed]
100. Fernandez, N.F.; Gundersen, G.W.; Rahman, A.; Grimes, M.L.; Rikova, K.; Hornbeck, P.; Ma'Ayan, A. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci. Data* **2017**, *4*, 170151. [CrossRef]

UNIVERSITÉ DE RENNES 1

Titre : Méthode d'apprentissage profond pour l'analyse génomique des cancers canins comme modèles des cancers humains

Mot clés : Apprentissage profond, génomique, cancers, espèce modèle

Résumé : Les méthodes d'apprentissage profond (DL) se sont récemment révélées être de puissantes stratégies pour prédire l'activité régulatrice d'une séquence génomique et donc pour, *in fine*, évaluer l'impact des mutations régulatrices sur l'expression des gènes. L'outil Basenji propose une approche DL utilisant des réseaux de neurones convolutifs pour prédire le niveau d'expression de gènes humains. Nous avons adapté ce programme pour entraîner un modèle d'expression génique spécifique au chien et montré que ce modèle de prédiction atteignait des performances similaires à celles observées chez l'homme, avec des corrélations élevées entre les niveaux d'expression réels et ceux prédits (R=0,66). Pour prédire le niveau d'expression de gènes canins, nous démontrons également que l'utilisation du modèle de prédiction canin (approche intra-espèce) aboutit à de meilleures performances que le modèle humain (approche inter-espèce), notamment en lien avec certaines caractéristiques spécifiques aux séquences canines (niveau de GC, d'éléments transposable et conservation évolutive). Le chien étant un modèle naturel pour l'étude des cancers humains, nous avons également exploité ces modèles pour prédire l'impact de mutations non-codantes sur l'expression de gènes impliqués dans les cancers. Nous avons ainsi localisé 1301 mutations communes entre l'homme et le chien, suggérant un rôle fonctionnel dans la régulation de l'expression de gènes impliqués dans les cancers. Finalement, nos modèles et les outils pour les exploiter sont disponibles sur GitHub : https://github.com/ckergal/BLIMP.

Title: Deep learning methods for genomic analysis of canine cancers as models for human cancers

Keywords: Deep learning, genomics, cancers, species model

Abstract: Deep learning (DL) methods have recently been shown to be powerful strategies for predicting the regulatory activity of a genomic sequence and thus for ultimately assessing the impact of regulatory mutations on gene expression. The Basenji tool proposes a DL approach using convolutional neural networks to predict the expression level of human genes. We adapted this program to train a dog-specific gene expression model and showed that this model achieved similar performance to that observed in humans, with high correlations between real and predicted expression levels (R=0.66). To predict the expression level of canine genes, we show that the canine prediction model (within-species approach) leads to better performances than the human model (cross-species approach), particularly due to some specific features of canine sequences (GC content, transposable elements and evolutionary conservation). As the dog is a spontaneous model for human cancers, we used these models to predict the impact of non-coding mutations on the expression of genes involved in cancers. We identified 1301 common mutations to both humans and dogs, suggesting a functional role in the regulation of the expression of genes involved in cancer. Finally, models and tools to exploit them are available on GitHub: https://github.com/ckergal/BLIMP.