Identification des signatures génétiques de la sélection chez le chien

Amaury Vaysse

N° d'ordre :4371 **ANNÉE 2011**

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : BIOLOGIE

Ecole doctorale Vie - Agronomie - Santé

présentée par

Amaury VAYSSE

préparée à l'unité de recherche IGDR UMR 6061 CNRS
Institut de Génétique et Développement de Rennes
Composante universitaire Sciences de la Vie et de l'Environnement

---

Identification des signatures génétiques de la sélection chez le chien

Thèse soutenue à Rennes
le 16 décembre 2011

devant le jury composé de :

Dr. Lluis QUINTANA-MURCI
Directeur de recherche
CNRS URA3012 Institut Pasteur, Paris / *rapporteur*

Dr. Michèle TIXIER-BOICHARD
Directeur de recherche
UMR1313 INRA/AgroParisTech, Paris / *rapporteur*

Pr. Christian DELAMARCHE
Professeur
université de Rennes 1/CNRS UMR 6026,
Rennes / *examinateur*

Dr. Catherine HÄNNI
Directeur de recherche
UMR 5242 CNRS/INRA/Université Claude Bernard
Lyon I/ENS, Lyon / *examinateur*

Dr. Laurent TIRET
Maître de conférences
École vétérinaire de Maisons Alfort/UMR 955
INRA / *examinateur*

Dr. Christophe HITTE
Ingénieur de recherche
CNRS UMR 6061, Rennes / *directeur de thèse*

Remerciements

Je tiens à remercier premièrement Michèle Tixier-Boichard et Lluis Quintana-Murci pour avoir lu attentivement ma thèse, rédigé leurs rapports sur mon travail et autorisé ma soutenance, ainsi que Catherine Hänni, Laurent Tiret et Christian Delamarche pour avoir accepté d'examiner mon travail.

Mon travail de thèse, commencé en mars 2008, fait parti du travail réalisé depuis février 2006 dans l'équipe génétique du chien. Je tiens à remercier chronologiquement les trois principales personnes qui m'ont permis de définir et réaliser mon travail de bioinformaticien/ biostatisticien. Premièrement je remercie Catherine André, qui m'a engagé en 2006 pour réaliser la base de données de la banque d'échantillons canin "caniDNA", Catherine m'a permis de découvrir et réaliser les analyses de liaison génétique de maladies héréditaires. Je remercie aussi Francis Galibert qui fut le premier à me soumettre des problèmes d'ordre statistique, ce qui a initié mon intérêt pour les analyses statistiques des différents projets du laboratoire, notamment l'utilisation du dN/dS sur les récepteurs olfactifs qui à été le premier pas pour analyser cet indice sur les pseudogènes prédit par Thomas Derrien lors de sa thèse, cette analyse a été le point de départ de mon projet de thèse réalisé sous la direction de Christophe Hitte. Christophe Hitte que je remercie pour son encadrement exceptionnel, pendant ma thèse nous étions un binôme de chercheurs dont l'un, inexpérimenté, naturellement encadré par l'autre, nettement plus expérimenté. Je retiendrai de ces 3 grands, la joie de vivre et les qualités humaines de Catherine, la passion de Francis, la rigueur, le sens des priorités scientifiques et de l'humour de Christophe.

Je remercie, en plus des trois grands, tous les membres de l'équipe génétique du chien: tout ceux qui m'ont accueilli, Je pense à Thomas, Patricia, Hélène, Sandrine, Maud, Stéphanie, Thierry, Benoît, Laetitia H; tous ceux qui sont arrivés et reparti pendant mon passage dans l'équipe: Anaïs, Marc, Michaëlle, Matthieu, Clotilde, Noémie, Estèle, Anne-Sophie mais aussi, ceux qui sont restés moins longtemps: Julien, Sébastien, Berline, Leslie, Morgane (Je te mets dans cette catégorie, même si je sais que tu reviendras), Aline (tu faisait partie de l'unité avant d'intégrer l'équipe, mais je te classe quand même ici), Isabelle, Julie, Chloé, Amandine; tous ceux qui avaient quitté l'équipe mais sont revenus parce qu'ils se sentent mieux ici qu'ailleurs, Edouard, Pascale, Richard (même si tu es re-parti); ceux qui sont arrivés après moi et qui font toujours partie de l'équipe génétique du chien, Laetitia L, Naoual, Nadine et tout ceux que j'ai oublié ici. Que j'ai eu directement ou indirectement l'occasion de travailler avec vous, je garde un très bon souvenir de vous tous et quelques amitiés. J'espère n'oublier personne dans cette liste.

Je remercie aussi de manière plus concise tous ceux qui, dans l'IGDR ou à côté, m'ont fait progresser par leurs conseils, leurs avis ou en me soumettant un questionnement informatique ou statistique (ils se reconnaîtront), tous ceux avec qui j'ai eu l'occasion d'échanger des points de vues sur la science et au delà de la science (Lucie, Pierre, Nabila, Nicolas S, Jacques(s), Sébastien(s), Vincent(s) et tant d'autres) avec un remerciement particulier pour Stéphane et Géraldine. Stéphane et Géraldine qui font partie des personnes indispensables à l'IGDR (oui, je sais, "les cimetières sont remplis de personnes indispensables") sans qui les thèses se dérouleraient bien différemment. Les membres du service gestion sont aussi de ceux là et je remercie Nadine et Nathalie pour les tâches administratives des missions et transports.

Parmi ceux qui ont influencé le déroulement de ma thèse, je remercie mon tuteur, Patrick Prunet et les membres de mon comité de thèse Eric Petit et Mathieu Gautier.

Je remercie également Hugues Roest-Crollius et David Enard de l'équipe "Dynamique et organisation des génomes" de l'ENS de Paris avec qui nous avons collaboré sur la partie concernant la sélection naturelle et collaborons toujours sur la recherche de co-occurrence entre les sélections naturelles et artificielles.

Je remercie de même Abhirami Ratnakumar et Matthew Webster de l'université d'Uppsala avec qui nous avons collaboré au sein du consortium européen LUPA (que je remercie également) pour rechercher les signatures de la sélection artificielle et avec qui nous collaborons toujours pour caractériser finement les différences entre races et rechercher les haplotypes liés aux traits phénotypiques de race.

Avant même ma thèse j'ai rencontré et bénéficié de l'encadrement d'individus formidables que je remercie ici: pour la biologie Christine Le Péron, Céline Raguénès-Nicol et Jean-François Hubert et pour l'informatique, je ne citerai que Patrice Burgevin et Pierre Guenot.

D'un point de vue plus personnel, je remercie ma famille et surtout ma mère, Jeanne et mon frère cadet, Gaëtan qui savent ce que c'est de vivre au côté d'un doctorant .....

Enfin, je tiens à remercier "Insert_Your_Name_Here" pour les apports considérables que j'ai reçu de sa part.

Je vais maintenant vous laisser lire la science dans ce manuscrit, mais avant cela, je souhaite soumettre à votre réflexion premièrement la fameuse phrase

« Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution. »

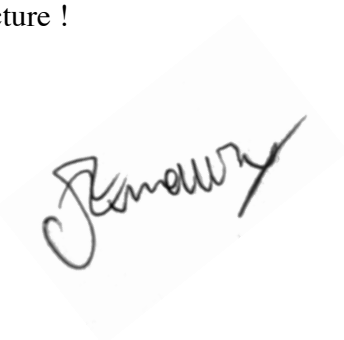
Theodosius Dobzhansky

qui illustre l'intérêt de ce vaste domaine qu'est la biologie de l'évolution et deuxièmement cette phrase qui explique, à mon sens, pourquoi les modèles et indices mathématiques permettent l'étude de l'évolution alors que toute évolution est extrêmement complexe:

« Tout est plus simple qu'on ne peut l'imaginer et en même temps plus enchevêtré qu'on ne saurait le concevoir. »

Johann Wolfgang von Goethe

Bonne lecture !



Résumé

Dès 1859 Charles Darwin s'appuie dans son célèbre ouvrage "*On the Origin of Species by means of natural selection*" sur les variations des espèces domestiquées pour développer son explication du concept de la sélection naturelle. Lorsque l'Homme domestique une espèce, il sélectionne les individus les plus adaptés à ses besoins ou désirs. Ainsi chaque espèce domestiquée est une véritable expérience de manipulation de l'évolution d'une espèce. De ce point de vue, l'espèce canine (*Canis lupus familiaris*) domestiquée il y a environ 15.000 ans à partir du loup gris (*Canis lupus lupus*) est la plus ancienne de ces expériences. L'espèce "chien domestique" est aujourd'hui composé de plus de 350 races qui sont autant d'isolats génétiques issus d'une sélection artificielle drastique et de croisements consanguins pratiqués durant les 2~3 derniers siècles. L'espèce canine représente par conséquent un modèle pour étudier les maladies génétiques ainsi que pour l'étude de l'évolution. En effet deux phases principales ont rythmé l'évolution du chien : une première période dominée par la sélection naturelle au cours de l'évolution des canidés et une seconde, récente avec la création de centaines de races par une sélection artificielle intense appliquée par l'Homme.

J'ai réalisé mon travail de thèse avec pour principaux objectifs 1) d'établir le catalogue complet des gènes canins sous sélection positive dans le contexte phylogénétique de 10 mammifères, afin de rechercher si l'impact de la sélection positive sur le génome canin est différent d'autres mammifères euthériens, et 2) d'identifier les régions de forte différenciation allélique entre races canines qui vont constituer les locus candidats de la sélection artificielle qui ont été influencés par la création des différentes races canines actuelles.

Ces deux parties de mon travail se sont déroulées dans le cadre de deux collaborations. La première collaboration au niveau national avec l'équipe du Dr. Hugues Roest Crollius (équipe DYOGEN ENS Paris) nous a permis d'analyser les événements de sélection naturelle survenus dans la lignée canine en comparaison des événements de sélection naturelle survenus dans les lignées des autres espèces. Nous avons analysé la sélection positive en comparant les séquences codantes de 10.730 gènes en relation d'orthologie de type 1:1 (un seul gène orthologue identifié dans chaque espèce) entre le chien et neuf autres espèces : (i) quatre génomes de primates (Homme, ouistiti, macaque, orang-outan et chimpanzé); (ii) deux génomes de rongeurs (souris et rat); (iii) et deux espèces plus proches du chien (cheval et vache). Le calcul du test statistique LRT nous a permis de définir les gènes sous sélection positive dans chacune des 10 espèces et de constater que le chien présente plus de gènes sous sélection positive en commun avec les Laurasiatheria et les rongeurs qu'à l'attendu.

La réalisation du second objectif de ma thèse s'est inscrit dans le cadre du consortium européen de génétique du chien LUPA (7ème programme cadre européen). Au sein de ce consortium, nous avons collaboré avec l'équipe du Dr Matthew Webster (Université d'Uppsala, Suède). Nous avons utilisé un indice dérivé du Fst sur les données de génotypage de 170.000 marqueurs SNP de 456 chiens répartis en 30 races d'au moins 10 individus. Nous avons déterminé le catalogue des régions de différenciation entre races de chien qui sont à la fois candidates pour être les cibles de la sélection artificielle et pour être responsables des différences phénotypiques fixées entre races.

Ce projet ouvre les perspectives de pouvoir déterminer s'il existe des régions du génome qui sont constamment affectées par les sélections naturelle et artificielle et d'aborder l'espèce canine comme une simulation réduite, mais accélérée de la radiation des mammifères.

Abstract

As early as 1859 Charles Darwin, in his famous "*On the Origin of Species by Means of Natural Selection*", bases its explanation of the concept of natural selection on changes in domesticated species. When Man domesticates a species, he takes possession of it, selects the individuals which are the best adapted for his needs or desires. Thus, each domesticated species is a genuine experience of manipulating the evolution of a species. In this respect, the canine species (*Canis lupus familiaris*), domesticated about 15,000 years ago from the gray wolf (*Canis lupus lupus*) is the oldest of these experiments. The "domestic dog" species is now composed of more than 350 breeds which are all genetic isolates resulting from a drastic artificial selection and inbreeding practices during the 2~3 past centuries. The canine species is therefore a model to study genetic diseases as well as for the study of evolution. Indeed two main phases have paced changes in the dog: a first period dominated by natural selection during the evolution of canids and a second, with the recent creation of hundreds of breeds by intense artificial selection applied by Man.

I did my PhD thesis with the main goals of 1) establishing the complete catalog of canine genes under positive selection within the phylogenetic context of 10 mammals, to investigate whether the impact of positive selection in the canine genome is different than in other eutherian mammals, and 2) identifying regions of high allelic differentiation between breeds that are candidate locus to be under artificial selection pressure at the origin of the creation of dog breeds.

These two parts of my work took place within the framework of two collaborations. The first collaboration - at the national level with the team of Dr. Hugues Roest Crollius (team DYOGEN ENS Paris) - allowed us to analyze the events of natural selection that occurred in the canine lineage in comparison to natural selection events occurring in the lineages of other species. We analyzed positive selection by comparing the coding sequences of 10,730 genes in 1:1 orthology relationship (one orthologous gene identified in each species) between the dog and nine other species: (i) four primate genomes (Human, marmoset, macaque, orangutan and chimpanzee), (ii) the genomes of two rodents (mouse and rat) (iii) and two species closer to the dog (horse and cow). The calculation of the LRT statistical test allowed us to identify the genes under positive selection in each of these 10 species and to find that the dog has more genes under positive selection in common with Laurasiatheria and rodents than expected.

The second part of my thesis took place in the framework of the european dog genetics consortium LUPA (7th European Framework Programme). Within this consortium, we have worked with the team of Dr. Matthew Webster (Uppsala University, Sweden). We used an index derived from the Fst index on the data from the genotyping of 170,000 SNPs in 456 dogs distributed in 30 breeds of at least 10 individuals. We determined the catalog of regions of differentiation between dog breeds. These regions are both candidates to be targets of artificial selection and to be responsible for the fixed phenotypic differences between breeds.

The project call for determining whether there are regions of the genome that are constantly affected by both natural and artificial selection and to consider the canine species as a reduced but accelerated simulation of mammals radiation.

Table des matières

Table des illustrations

Abréviations

ADN	Acide DésoxyriboNucléique
BH	Benjamini et Hochberg Correction pour test multiple
CFA	Canis FAmiliaris (chromosome)
EST	Expressed Sequence Tag
FCI	Fédération Cynologique Internationale
FDR	False Discovery Rate
GO	Gene Ontology
GOLD	Genome OnLine Database
LRT	Likelihood Ratio Test
Mb, kb, pb	Megabases, kilobases, paire de bases
OMIM	Online Mendelian Inheritance in Man
PAML	Phylogenetic Analysis by Maximum Likelihood
CSV	Comma Separated Value
SINE	Short interspersed nuclear elements
SINE_cf	SINE Canis familiaris
chr	chromosome
SNP	Single Nucleotid Polymorphism

Introduction

I. Introduction générale

Depuis la publication par Charles Darwin en 1859 de "*On the Origin of Species by means of natural selection*", le concept de l'évolution des espèces a été largement accepté par la communauté scientifique (Darwin, 1859). L'idée d'une sélection qui trie les variants au sein d'une population et sépare les 'aptes' des 'inaptes' a été exprimée bien avant le XIXe siècle. La grande avancée de la théorie de l'évolution de Darwin fut la proposition du mécanisme de sélection naturelle pour expliquer la diversité des espèces. La sélection naturelle est la clé de voûte de la doctrine de Darwin qui en fait le moteur et le mécanisme responsable de l'évolution des êtres vivants. Au sein des populations, les individus présentent des traits variables, la sélection naturelle favorise la survie et donc la reproduction des individus les mieux adaptés à leur environnement. Ces individus transmettent ces adaptations à leur descendance et permettent l'adaptation de la population complète à l'environnement. La sélection permet l'adaptation des populations, et la création des espèces qui opère à partir de la transmission héréditaire des caractères.

Les principes régissant la transmission du matériel héréditaire étaient connus depuis les travaux de Gregor Mendel (Mendel, 1865), mais c'est à la fin du XIXe siècle et au début du XXe que s'est développée la génétique avec la redécouverte des lois de Mendel par d'autres biologistes comme Hugo de Vries qui introduisit le concept de mutation biologique (De Vries, 1902). Depuis ces travaux, la génétique a joué un grand rôle dans la compréhension de l'évolution. La génétique des populations est une discipline qui émane de la génétique et est liée à la volonté d'expliquer l'évolution en termes génétiques. La génétique des populations naît avec les travaux de G.H Hardy (Hardy, 1908) et W. Weinberg (Weinberg, 1908) qui décrivirent la répartition des génotypes dans les populations panmictiques où les croisements entre individus s'effectuent au hasard. Cette discipline se développe ensuite fortement à partir des travaux du mathématicien R.A. Fisher (Fisher, 1930) et des biologistes J.B.S Haldane (Haldane J. B. S., 1927b) et S. Wright (Wright, 1921) qui ont établi des modèles statistiques pour décrire l'évolution des populations. La génétique des populations analyse l'impact des pressions évolutives sur les facteurs génétiques des populations. L'étude plus précise de l'impact de la sélection sur les génomes s'appuie désormais sur les connaissances et les ressources acquises grâce aux données de la biologie moléculaire.

Si le rôle de l'ADN dans l'hérédité est connu à partir de 1944 (Avery, *et al*., 1944), la génétique moléculaire a pris son essor à partir de 1953. En effet, le 25 avril 1953, la revue Nature publiait trois articles qui allaient révolutionner la conception de la molécule d'ADN et engendrer l'avènement de l'ère de la biologie moléculaire. Le premier article de James

Watson et Francis Crick présentait la modélisation de la structure de l'ADN en une double hélice où les quatre bases, Adénine, Thymine, Guanine et Cytosine sont liées par des liaisons hydrogènes(Watson and Crick, 1953) . Le second article écrit par Maurice Wilkins et ses collaborateurs, proposait une structure en hélice composée de deux chaînes (Wilkins, *et al*., 1953). Enfin, Rosalind Franklin arrivait à des conclusions identiques en présentant les clichés cristallographiques d'une des formes de l'ADN (la forme B) ayant servi de support à la prédiction du modèle de la double hélice (Franklin and Gosling, 1953). En 1962, quatre ans après la mort de R. Franklin, J. Watson, F. Crick et M. Wilkins se partagèrent le prix Nobel de médecine "pour leurs découvertes concernant les structures moléculaires des acides nucléiques et son importance pour le transfert des informations dans les organismes vivants".

Sur la base de ces découvertes, les progrès réalisés en biologie moléculaire, plus particulièrement, par les nouvelles techniques de biotechnologie (PCR Polymerase Chain Reaction), de séquençage automatique (robot Applied développé par L. Hood dès 1987) et dans l'automatisation des procédures expérimentales (technique du shot-gun, production d'étiquettes de séquences transcrites EST) vont aboutir au déchiffrage des séquences d'ADN de nombreux organismes dont celui de l'Homme dont la séquence brute est parue en 2001 (Lander, *et al*., 2001; Venter, *et al*., 2001). Depuis 1995 et le premier séquençage du génome de la bactérie *Haemophilus Influenzae* (Fleischmann, *et al*., 1995), le nombre d'organismes dont le génome a bénéficié d'un séquençage complet (virus, bactéries, eucaryotes multicellulaires, mammifères) a cru de façon spectaculaire. Le stockage et la mise à disposition pour les biologistes de l'énorme quantité d'informations découlant de ces projets ont été assurés par la mise en place de base de données publiques telle que GenBank (1982). En février 2010, près de 6400 projets de séquençage (complets ou partiels) sont recensés et près de 1200 organismes appartenant aux archébacteries, procaryotes et eucaryotes disposent d'un séquençage complet de leurs génomes (GOLD Database http://www.genomesonline.org/gold_statistics.htm). Cependant, bien que la séquence d'un génome puisse être considérée comme l'étape ultime de la connaissance de sa structure (i.e l'enchaînement des quatre nucléotides A, T, G et C), celle-ci ne renseigne pas sur la manière dont l'évolution a façonné ce génome. L'analyse de l'impact de l'évolution est un défi, qui s'effectue de plusieurs manières dont la comparaison des séquences des génomes de différentes espèces et l'exploration du polymorphisme génétique des populations au sein d'une espèce.

Les approches de génomique comparative reposent sur le principe que les génomes des espèces actuelles sont issus d'un génome commun ancestral qui s'est transformé sous l'effet de variations génétiques neutres et sélectionnées. Comparer les variations des régions

génomiques conservées entre espèces permet de caractériser l'impact des sélections qui ont conduit à la création de ces espèces. Au sein d'une espèce, une pression de sélection qui s'exerce sur une population n'influencera pas uniquement le gène sélectionné mais laissera une signature sur la région environnante du génome grâce à un déséquilibre de liaison lors de la transmission des caractères. Les approches de génétique des populations sont basées sur ce principe de déséquilibre de liaison : analyser les variations de polymorphisme le long du génome des populations permet de localiser et caractériser l'impact des sélections récentes et encore actives qui différencient les populations d'une même espèce.

Dès 1859, Darwin s'appuie sur la variation observée chez les espèces domestiquées pour illustrer ses théories et faire des parallèles entre les espèces domestiquées et les espèces sauvages, entre la sélection artificielle et la sélection naturelle. Que la sélection artificielle soit consciente ou non, chaque espèce domestiquée peut être considérée comme une expérience évolutive. L'espèce canine (*Canis lupus familiaris*) est la plus ancienne de ces expériences. Tous les chiens descendent du loup gris (*Canis lupus lupus*) domestiqué par l'Homme il y a environ 15.000 ans, et se seraient répandus dans toute l'Asie et l'Europe, avant d'accompagner l'Homme dans le nouveau monde. Le chien appartient à l'ordre des carnivores et à la famille des *Canidae* qui regroupe un grand nombre espèces (n=34) comme le renard, chacal, coyote, dingo, loup... toute issues d'un ancêtre commun carnivore datant de 10 millions d'années. Les différentes espèces de canidés sont classées en quatre groupes ; celui comprenant le chien domestique réunit les membres du genre *Canis* (le chacal, le coyote, le loup, le chien, le dingo), le dhôle (*Cuon alpinus*) et le lycaon (*Lycaon pictus*). Ils possèdent tous un caryotype composé de 2n=78 chromosomes acrocentriques. L'espèce canine est l'espèce *Canidae* la plus diversifiée, elle se compose aujourd'hui de plus de 350 races issues d'une sélection artificielle drastique et de croisements consanguins pratiqués durant les derniers siècles pour l'essentiel. L'espèce canine présente une double échelle de temps dans son évolution; une période dominée par la sélection naturelle et son adaptation à la vie avec l'Homme et une période récente dominée par la sélection artificielle appliquée volontairement par l'Homme.

La structuration de la population canine en races et son histoire évolutive confèrent à l'espèce canine le statut de modèle animal ayant un fort potentiel pour l'étude de données génétique et en génétique évolutive en particulier. En effet, le chien présente une extraordinaire variation phénotypique entre les races et une forte homogénéité au sein de chaque race. De par leurs origines et leurs histoires, les races canines sont des isolats génétiques à l'instar des populations humaines isolées géographiquement (islandaise par

exemple et insulaire en général) ou culturellement (Amish ....) (Galibert, *et al*., 2004). La création des races à partir d'un nombre limité de fondateurs et donc d'allèles, a engendré une diminution de la variabilité génétique, une forte consanguinité et par conséquent une perte d'hétérozygotie et un taux d'homozygotie important au sein de chaque race. Le tribut à ce mode de sélection est la forte prévalence des maladies génétiques et des traits particuliers à caractère héréditaire qui ségrègent au sein des races. Les maladies génétiques canines sont spontanées et sont pour une grande partie homologues aux maladies humaines. Le potentiel du modèle canin, la structuration en races, la présence de maladies génétiques fréquentes et d'aptitudes particulières font du chien un excellent modèle pour déterminer les relations génotypes/phénotypes. Ces atouts ont motivé plusieurs équipes à étudier la structure du génome canin et ses régions fonctionnelles. Dès 2004, le National Institute of Health (NIH) aux États-Unis lance l'initiative du séquençage profond (couverture >7X) du génome du chien. Ce projet fut mené par le BROAD Institute de Cambridge (USA) à partir du génome d'un chien femelle de race boxer, choisi sur le critère d'une faible hétérozygotie génétique qui facilite la phase d'assemblage du séquençage. Suite aux données de séquençage, deux générations de puces SNP, 22K, 50K ont été développées et commercialisées par les sociétés Affymetrix et Illumina. Grâce aux données de séquence et à une dernière génération de puces permettant de génotyper plus de 170.000 marqueurs de type SNP régulièrement espacés le long du génome, plusieurs centaines de chiens sur plusieurs dizaines de races ont pu être récemment analysées (Vaysse, *et al*., 2011). Ainsi nous disposons des outils génomiques nécessaires pour (i) étudier l'impact de la sélection naturelle sur la séquence du génome du chien (Lindblad-Toh, *et al*., 2005) et (ii) étudier l'impact de la sélection artificielle sur les différentes races de chiens.

Mon travail de thèse a permis (i) l'identification des gènes codant pour des protéines sous sélection positive dans la lignée carnivore à partir de plus de 10.000 gènes, orthologues entre 10 espèces mammifères, et des fonctions qui leur sont associées et (ii) l'établissement d'un catalogue des régions génomiques de différenciation allélique et potentiellement ciblées par la sélection artificielle lors de la création des races canines à partir de données de polymorphisme distribuées le long du génome de chaque race et permettra l'analyse et la comparaison des régions du génome détectées sous sélection positive et sous sélection artificielle.

II. Le concept d'évolution

Au XVIIIe siècle, l'idée d'évolution, la transformation d'une espèce vivante en une autre, s'oppose à celle de fixité des espèces professée par Carl Linné. Buffon, dans son ouvrage "*Histoire naturelle*" (1749 à 1789) pose les premiers concepts de changement des espèces qui donneront, au XIXe siècle, les théories de l'évolution. Parmi les naturalistes inspirés par Buffon, Jean-Baptiste Lamarck publie dans son principal ouvrage "*Philosophie zoologique*" en 1809 un développement du concept de transformation des espèces, il est un des premiers à reconnaître la nécessité théorique de l'évolution pour expliquer la complexité des êtres vivants. Enfin Darwin, à partir de la publication en 1859 de "*On the origin of species by means of natural selection",* propose un mécanisme expliquant l'évolution des espèces permettant de dériver la diversité du vivant actuelle à partir des espèces ancestrales. Ce mécanisme part d'une généralisation des principes que Malthus a développé dans son "*Essai sur le principe de population*" (1798). Ce concept repose sur 4 principes : i) un milieu donné contient des ressources limitées, ii) ces ressources ne permettent pas à tous les individus qui naissent dans une population d'atteindre l'âge de reproduction, ce qui entraîne une compétition pour la survie des individus, iii) Les individus présentent des variations héritables et iv) certaines de ces variations influent sur le succès de l'individu dans la compétition pour la survie. Par conséquent les individus les plus adaptés transmettent leurs caractères à la génération suivante, ce qui permet à une population de s'adapter à son environnement. Outre le fait d'expliquer l'histoire de la vie, le concept d'évolution et la compréhension du mécanisme de la sélection ont de nombreuses applications dans différents aspects de la biologie (Bull and Wichman, 2001) en permettant l'annotation fonctionnelle des génomes grâce à la génomique comparative (Haldane J., 1927a; Borowsky, 2008) ou en expliquant des phénomènes médicaux tel que la sélection des bactéries aptes à survivre en présence d'antibiotiques (Manten, 1963).

Aujourd'hui les principes de la théorie de l'évolution sont toujours basés sur la transmission des caractères à la descendance avec des variations d'une génération à l'autre et la sélection des caractères les plus favorables pour la survie et la reproduction des individus. Aucune autre théorie scientifique n'explique la mise en place de la diversité biologique et l'histoire de la vie.

II.1. Les forces de l'évolution

L'évolution à court terme concerne une seule espèce et même une seule population. À cette échelle, l'évolution se définit comme le changement des fréquences des allèles des

différents gènes au sein de la population étudiée. En l'absence de pressions évolutives une population "idéale", c'est à dire isolée, d'effectif infini, non soumise à la sélection ou à la mutation (telle que décrite par G.H. Hardy et W. Weinberg en 1908) conserverait indéfiniment des fréquences allèliques identiques. Dans une population réelle, quatre principales pressions évolutives influent sur la diversité génétique des populations : la mutation, la migration, la sélection et la dérive. Chacune de ces pressions évolutives a une action sur la structure allélique de la population (Figure 1).

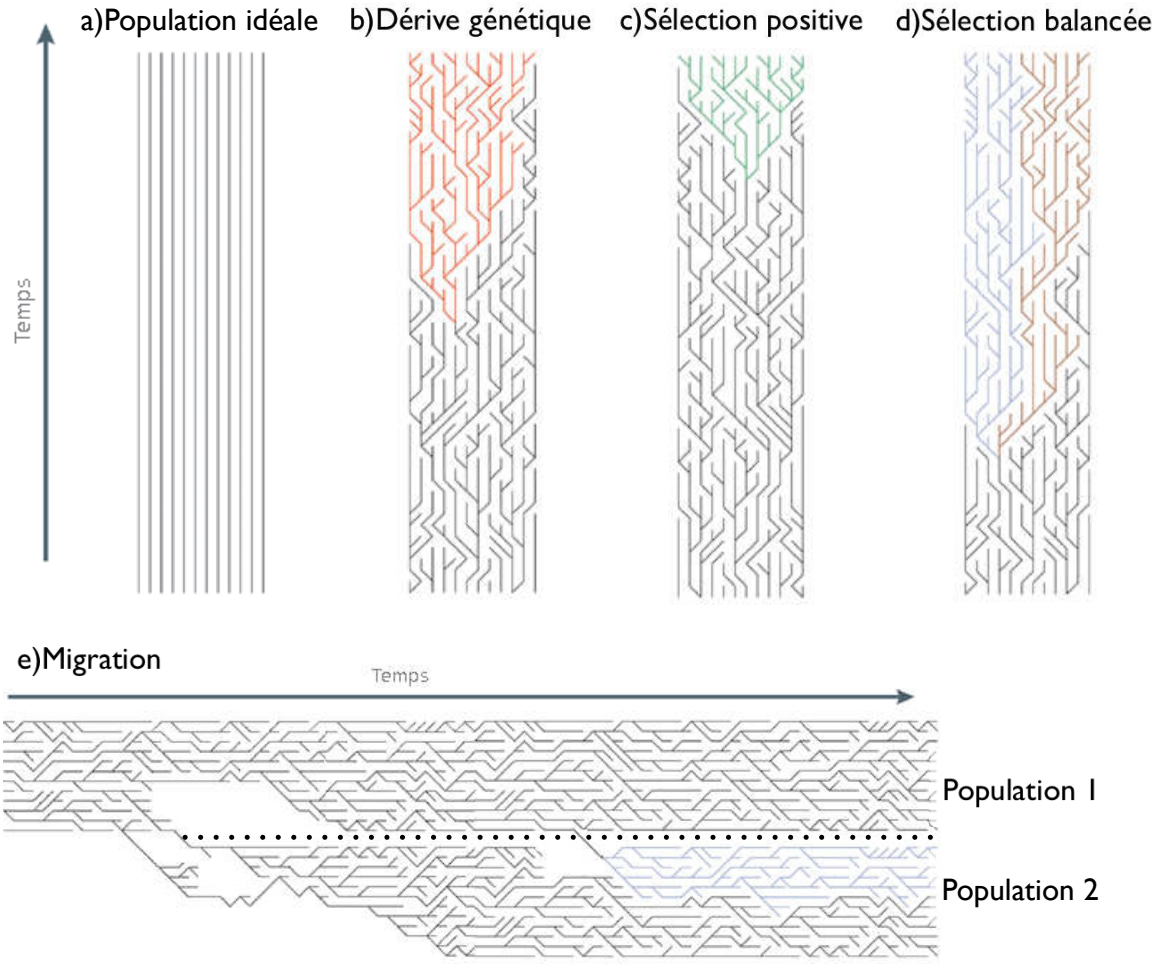


Figure 1 : ***Influence des quatre pressions évolutives****. Chaque ligne représente le devenir d'un allèle dans une population. a) En l'absence de pression évolutive, la composition en allèle de la population n'est pas modifiée au cours du temps. b) Dans une population réelle, les mutations changent les allèles, certains allèles sont transmis en plus grand nombre à la génération suivante et la dérive génétique aboutie à ce qu'un seul allèle soit fixé dans la population. c) La sélection positive permet la fixation plus rapide d'un allèle avantageux. d) La sélection balancée maintient la diversité de la population. e) La migration permet la création de populations dont le polymorphisme est une partie du polymorphisme d'origine et le transfert de polymorphisme entre population. Adapté de Bamshad et Wooding (Bamshad and Wooding, 2003).*

II.1.1. La théorie neutre de l'évolution

Les mutations et les changements de fréquences alléliques sont les variations que les gènes et les régions non-codantes subissent au cours du temps. En 1968 Motoo Kimura (Kimura, 1968) décrivit la théorie neutraliste de l'évolution. Dans la théorie neutraliste de l'évolution les variations des séquences sont neutres d'un point de vue évolutif et les fréquences des différents allèles dans la population varient de manière aléatoire par ce que l'on appelle la dérive génétique. La dérive génétique est alors la principale cause des différences entre les populations qui aboutit à la création d'espèces séparées et les pressions de sélection jouent un rôle dans l'adaptation des populations et des espèces, mais peu dans leur définition et leur différenciation.

II.1.2. La sélection directionnelle et balancée

II.1.2.1. Principe

La sélection directionnelle permet l'adaptation d'une population à son environnement ou le maintien de cette adaptation. La sélection directionnelle agit en diminuant ou augmentant les fréquences des variants génétiques responsables du phénotype sélectionné. Cette sélection varie de la sélection négative à la sélection positive. Premièrement la sélection négative ou purifiante est l'élimination des variants qui confèrent un désavantage sélectif, soit un handicap trop important, aux individus qui les portent. Ces variants délétères sont maintenus à faible fréquence dans la population et à terme sont éliminés. La sélection négative est une forme de sélection normalisante quand elle élimine les individus dont le phénotype s'écarte de l'état d'adaptation acquis. La sélection positive est la forme de sélection décrite par Darwin. La sélection positive agit sur la rétention des variants qui confèrent un avantage pour la survie et la reproduction (la fitness) de l'individu. La sélection positive va entraîner l'augmentation de la fréquence de ce variant dans la population jusqu'à aboutir à la fixation de l'allèle favorable. Ce processus permet, par exemple l'optimisation ou le changement de la fonction d'une protéine. La combinaison des sélections négative et positive peut permettre des effets de sélection disjonctive: la sélection positive favorise les individus présentant des phénotypes extrêmes (petits et grands par exemple) au détriment des individus présentant des phénotypes intermédiaires qui sont donc éliminés par la sélection négative.

La sélection balancée est un cas de sélection où les allèles vont être conservés pendant une plus grande période dans les populations en comparaison des allèles neutres qui se fixent beaucoup plus rapidement. La sélection balancée permet de conserver les variations

génétiques de la population car favorise les individus porteurs des deux allèles. Ces deux allèles vont être maintenus à une fréquence d'équilibre stable. Ceci se produit lorsque les individus hétérozygotes ont une meilleure valeur sélective que les individus homozygotes pour le gène ou la fonction responsable du caractère sélectionné ou lorsque la population vit dans un environnement dans lequel différents allèles peuvent être favorables selon les circonstances. La sélection balancée participe au maintien voire à l'augmentation de la diversité génétique et induit une augmentation de la proportion des allèles de fréquence intermédiaire.

L'évolution est dite neutre lorsque la sélection est absente ou très faible, c'est à dire lorsque les variants qui ne confèrent pas d'avantage ou de handicap important pour l'individu ont une fréquence qui varie de manière aléatoire dans la population. L'évolution neutre correspond à l'évolution en l'absence de pressions de sélection. Un variant neutre sera réparti aléatoirement dans la population. En particulier, un variant neutre peut être présent ou absent chez les individus qui vont créer la génération suivante. La fréquence des allèles de fitness équivalents va varier d'une génération à l'autre selon le principe de la dérive génétique. La variation de ces fréquences alléliques s'arrête lorsque l'un des allèles est fixé dans la population. En l'absence de pression de sélection, les variants se fixent à une vitesse et avec une probabilité qui dépendent de leur fréquence initiale dans la population.

II.1.2.2. Sélection naturelle et sélection artificielle

Au cours de l'évolution, la sélection naturelle agit sur la diversité génétique qui existe au sein des espèces, des populations, et favorise la survie et la capacité de reproduction des individus qui sont à un temps donné les mieux adaptés aux conditions et aux pressions environnementales. Les avantages sélectifs sont sélectionnés en favorisant leurs transmissions à leur descendance selon une échelle de temps plus ou moins longue de l'évolution. La sélection permet l'adaptation des populations et la création des espèces grâce à la transmission héréditaire des caractères et à la création de la diversité génétique.

Contrairement à la sélection naturelle qui agit au cours de l'évolution pendant des centaines de milliers ou millions d'années, la sélection artificielle telle qu'elle est pratiquée chez les espèces domestiquées est un processus rapide qui façonne l'architecture génétique des populations ou des races animales et végétales créées en fixant des patrons de polymorphisme. La sélection artificielle est la sélection d'animaux ou de végétaux pratiquée par l'Homme. La sélection artificielle volontaire ou involontaire se met en place dès que la co-habitation entre l'Homme et l'espèce animale ou végétale sélectionnée induit des changements morphologiques ou comportementaux liés à la domestication. La domestication

aboutit au retrait par l'Homme d'un animal ou une plante de son environnement et son maintien dans un environnement différent -Comme décrit par Darwin dans "*The variations of animals and plants under domestication*" (Darwin, 1868)- dans lequel l'Homme contrôle la reproduction des individus domestiques. Ce contrôle reproductif permet de poursuivre la sélection artificielle volontaire. La sélection artificielle volontaire se poursuit par l'élevage sélectif des animaux ou des végétaux grâce aux connaissances en génétique quantitative et en génétique des populations à des fins d'amélioration des performances agronomiques, de travail ou esthétiques. Les populations domestiques qui subissent la sélection artificielle sont moins influencées par la sélection naturelle. En effet, la prise en charge d'une espèce par l'Homme retire la compétition inter-espèce et permet ainsi le développement de caractères qui seraient "mal-adaptés" en condition sauvage comme la sélection de maladies chez le chien (Parker, *et al*., 2009) ou la baisse de reproductivité de certaines races bovine laitière (Flori, *et al*., 2009). En parallèle de cette relaxation de contrainte évolutive naturelle, la sélection artificielle pour l'amélioration des races et variétés domestiques est une pression nouvelle et forte, qui défini les règles de la compétition intra-espèce et dont les conséquences génétiques sur le polymorphisme sont visibles lorsque l'on compare des races d'une même espèce (Flori, *et al*., 2009; Akey J. M., *et al*., 2010).

II.1.3. La dérive génétique

La dérive génétique est la variation aléatoire des fréquences allèliques dans une population telle que l'a décrit Wright (Wright, 1931). En absence de pression de sélection, les variations des fréquences allèliques vont dépendre essentiellement de la taille de la population. L'impact de la dérive est alors accentuée chez les populations à petits effectifs, en raison de la forte contribution des écarts de fréquences allèliques observés d'une génération à l'autre. Inversement, les fréquences allèliques dans les populations à grands effectifs sont stables pendant plusieurs générations. La dérive génétique concerne surtout les allèles neutres, et peut entraîner la disparition totale ou la fixation d'un allèle. Le calcul de la probabilité de fixation d'un allèle en absence de sélection corrèle directement avec la fréquence de l'allèle et le nombre d'individus dans la population. Dans une petite population, un variant allélique aura plus de chance d'être fixé, la dérive génétique aura alors pour conséquence une diminution de la diversité génétique.

II.1.4. La migration

Un cas particulier de la dérive génétique est la migration d'un nombre d'individus, généralement un sous-ensemble d'une population qui va fonder une nouvelle population. Lorsqu'un nombre réduit d'individus se sépare d'une population originelle par définition plus vaste, pour aller coloniser un nouveau milieu, pour des raisons d'ordre culturel, ce petit groupe va représenter un échantillon aléatoire du pool allélique et donc du patrimoine génétique de la population originelle. Le processus de migration peut créer un véritable goulet d'étranglement de la variabilité génétique et causer un 'effet fondateur' qui entraîne une importante variabilité des fréquences alléliques à l'échelle de la population. Ainsi, il a été mis en évidence en génétique des populations humaines que la diversité génétique des populations non-africaines représentent un sous-ensemble de la diversité des populations africaines et donc un sous-ensemble de la variabilité originelle (Henn, *et al*., 2011; Javed, *et al*., 2011). La migration entre deux populations établies peut à l'inverse rétablir ou augmenter la diversité d'une population. Non seulement ces flux géniques s'opposent aux effets de baisse de diversité observé lors des événements de sélection et de dérive génétique, mais ils peuvent aboutir à l'introgression dans une population d'un phénotype présent dans une autre. Comme l'illustre l'introgression de la couleur noire du chien chez le loup (Anderson, *et al*., 2009).

II.2. Les polymorphismes génétiques

Un polymorphisme est un site de l'ADN qui présente au moins deux états dans une population qui apparaissent par mutation ou par réarrangements chromosomiques lors d'une recombinaison. Les technologies "haut débit" de recherche et de détection de polymorphismes sont basées sur la substitution d'un seul nucléotide, le SNP (Single Nucleotide Polymorphism). Un site est considéré comme polymorphe si on trouve deux allèles présents dans plus de 1% de la population. À moins de rechercher spécifiquement des variants rares, on identifiera donc des SNP ayant une fréquence d'allèle minoritaire d'au moins 1% dans les populations utilisées pour les rechercher. Les polymorphismes dus aux réarrangements chromosomiques font intervenir les quatre principaux processus de translocation, fusion, inversion et les insertions/délétions de segments chromosomiques. Le polymorphisme d'insertion correspond à des séquences d'ADN répétitif dispersées et qui possèdent une structure de transposons, ce sont des éléments d'ADN capables de s'insérer dans différentes régions du génome par rétrotransposition. Les deux principaux groupes d'insertions chez les mammifères sont les *long interspersed nuclear elements* (LINE) (Xiong

and Eickbush, 1990) et les *short interspersed nuclear elements* (SINE) (Okada, 1991), représentant plus d'un tiers du génome. Le polymorphisme de taille correspond à une variation dans le nombre de motifs répétés comme celle retrouvée dans les microsatellites (motif 1 à 6 pb), les minisatellites (motif de 6 à 10.000 pb) et les variations du nombre de copies (*copy number variation* -CNV-, motif de 10 kb à plusieurs mégabases).

II.2.1. Les polymorphismes neutres et le balayage sélectif

La sélection agit non seulement sur les gènes qui induisent directement le phénotype lié à la valeur sélective des individus (survie et reproduction) mais aussi sur les éléments régulateurs, les autres gènes et les polymorphismes qui sont proches physiquement de ces gènes. En effet les polymorphismes situés à proximité d'un gène recombineront d'autant moins avec ce gène qu'ils en seront proches. Il se crée un déséquilibre de liaison génétique qui a pour conséquence un accroissement rapide de la fréquence des polymorphismes proches d'un allèle sélectionné dans la population. Cette influence de la sélection d'un allèle avantageux sur les polymorphismes voisins se traduit par le phénomène dit d'auto-stop génétique (Maynard Smith and Haigh, 1974) et imprime une véritable signature génétique de l'évolution dans les populations actuelles.

Lorsque l'on utilise et analyse les seuls polymorphismes connus d'une population, l'ensemble des mutations responsables de la variation du phénotype sélectionné n'est pas représenté et disponible. Cependant, les fréquences des polymorphismes neutres influencés par les trois autres pressions évolutives : la migration, la dérive et la sélection sont en partie, représentées. Les effets de la migration et de la dérive s'appliquent à tout le génome, augmentant ou diminuant globalement la diversité de la population. Au contraire, la sélection affecte seulement les régions du génome responsables du phénotype sélectionné. Les allèles sélectionnés sont flanqués de locus polymorphes. Lorsque la sélection cible un locus, elle affecte aussi les locus proches qui ne recombinent pas ou peu avec le locus sélectionné de part leur proximité et conduit à un bloc d'haplotypes transmis de génération en génération.

La sélection positive aboutira à l'augmentation de la fréquence de l'haplotype constitué de polymorphismes neutres associés avec le locus sélectionné par auto-stop génétique, diminuant fortement la fréquence des autres haplotypes. Dans le cas extrême de sélection positive, on peut observer un véritable 'balayage sélectif' (selective sweep) complet de la région. Ce balayage sélectif est illustré par le cas d'une sélection forte d'un nouvel allèle d'un gène : la mutation apparue dans un haplotype confère un effet avantageux fort, la sélection positive fait alors augmenter rapidement sa fréquence dans la population sans que la

recombinaison ne rétablisse la diversité des polymorphismes à proximité. Dans cet exemple, la population sélectionnée présentera une zone de son génome fixée qui contient la sélection d'un variant nouveau, on parle alors de balayage sélectif fort (hard selective sweep). Dans le cas d'un changement de pression de sélection sur une population déjà adaptée, des allèles qui étaient neutres ou légèrement délétères peuvent devenir avantageux. Dans ce cas, les allèles nouvellement avantageux sont présents dans différents haplotypes. La sélection dans ce cas ne fixe pas une région du génome, mais va contribuer à l'enrichissement en certains haplotypes de la population. Ces changements de polymorphisme constituent un balayage sélectif modéré (soft selective sweep) (Figure 2).

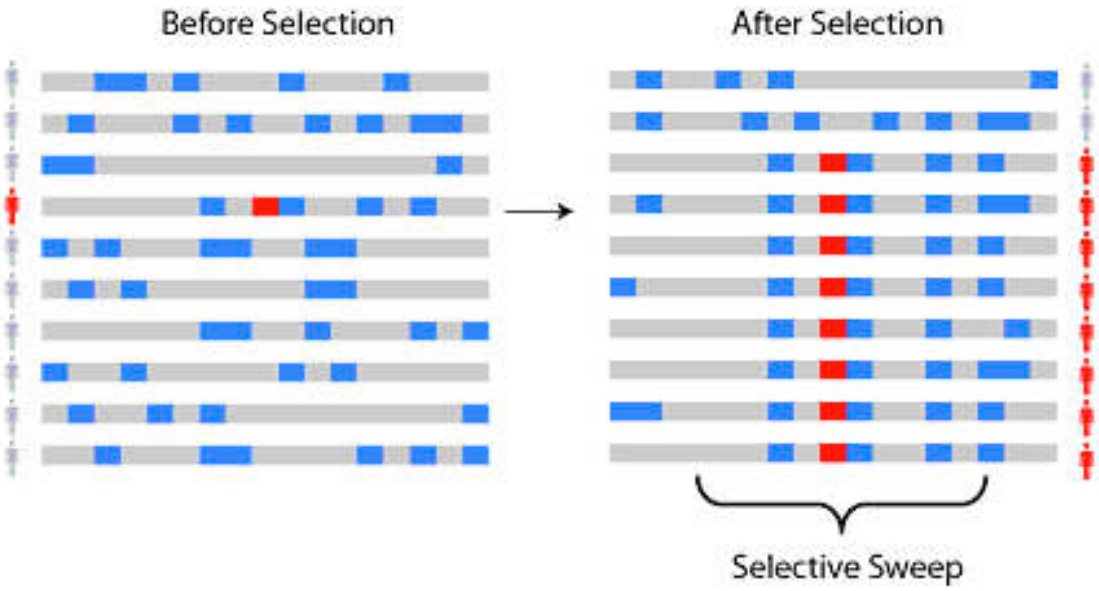


*Figure 2 : **Balayage sélectif fort**. Un nouveau variant avantageux est immédiatement sous sélection positive. Ce variant se répand dans la population et emporte les variants neutres qui étaient présents dans l'haplotype où il est apparu. La perte de polymorphisme qui en résulte représente le balayage sélectif fort. Repris de (Schaffner and Sabeti, 2008)*

II.2.2. Les mutations

Les mutations neutres ou silencieuses n'altèrent pas la séquence d'acides aminés et donc la synthèse de la protéine et n'ont pas de conséquences, en l'état des connaissances actuelles, sur l'individu. Les mutations fonctionnelles sont des polymorphismes qui ont des conséquences sur la partie codante des gènes ou sur les régions régulatrices et ont un impact direct sur le phénotype de l'individu. Ces mutations peuvent avoir des conséquences phénotypiques plus ou moins importantes. Elles correspondent aux mutations non-synonymes qui modifient la séquence protéique ainsi qu'aux mutations non-codantes mais qui influencent la régulation de l'expression du gène.

II.3. La détection de la sélection

II.3.1. Principes et tests de différenciation génétique

L'identification des événements de la sélection se divisent en deux catégories qui vont correspondre à la recherche de signatures moléculaires de la sélection naturelle entre espèces ou entre populations et à la détection de signatures de la sélection artificielle. Pour cette dernière, l'identification de signatures moléculaires repose principalement sur deux approches. La première est celle des haplotypes étendus, où un allèle sélectionné augmente sa fréquence si rapidement que son association avec les polymorphismes voisins n'est pas réarrangée par les événements de recombinaisons génétiques qui s'effectuent à chaque génération. La seconde est une approche basée sur l'identification des allèles fortement différenciés. Dans cette approche, il est considéré qu'un allèle sélectionné dans une population cause une plus grande différence de fréquence entre populations que pour les allèles sous évolution neutre.

Les tests développés pour détecter la sélection au sein des populations d'une espèce se divisent en tests intra-population et inter-populations et présentent une efficacité différente pour la détection des balayages sélectifs forts et modérés. Les méthodes d'études basées sur le polymorphisme d'une seule population ont pour objectif de détecter des changements de variabilité d'une région du génome par rapport au reste du génome. Parmi ces méthodes, les études des spectres des fréquences alléliques comparés entre régions génomiques (Simonsen, *et al*., 1995) et pouvant-être modélisés par des chaînes de Markov cachées (Kern and Haussler, 2010) s'appuient sur les distributions des polymorphismes d'une région et sont bien adaptées à la détection de balayages complets issus de mutations nouvelles , mais auront peu de puissance pour détecter les sélections de variants préexistants à la pression de sélection (Przeworski, *et al*., 2005). Les études basées sur l'hétérozygotie relative, quant à elles, recherchent une baisse de polymorphisme par la comparaison de la quantité de sites polymorphes dans la région étudiée et dans le reste du génome. Les études basées sur le déséquilibre de liaison tel que le REHH (Relative Extended Haplotype Homozygosity) (Sabeti, *et al*., 2002) ou l'iHs (Voight, *et al*., 2006) sont puissants pour détecter des balayage forts en cours d'établissement.

Les méthodes inter-populationnelles permettant la recherche de balayage sélectifs modérés sont souvent basées sur l'indice de fixation de Wrigth, le Fst, (Wright, 1951). L'indice Fst se calcule en regroupant différentes populations et représente la fraction de la variabilité de la population totale qui est due à la division en sous-populations. Cet indice varie de zéro lorsqu'il n'y a pas de structure de population à 1 quand les populations ont fixé

des allèles différents. L'indice Fst s'appuie sur les changements de fréquences alléliques des polymorphismes entre populations. En l'absence de sélection, le Fst sera influencé par la dérive génétique qui affecte tous les locus de la même manière alors que la sélection d'une population sur un locus donné augmentera le Fst obtenu en comparant cette population aux autres. La méthode originale utilisant le Fst pour détecter la sélection positive est le test de Lewontin & Krakauer (Lewontin and Krakauer, 1973) qui repose sur la comparaison de la variation de Fst entre locus par rapport à une variation théorique qui est observée si 1) les sous-populations proviennent de la sélection aléatoire d'individus d'une population générale et 2) si le Fst est identique pour chaque locus. Ce test ne tenant pas compte de la grande variabilité entre locus, différentes approches de recherche de signature de sélection ont été proposées, comme les méthodes de calcul Bayesien implémentées par l'utilisation de méthodes de Monte-Carlo sur des chaînes de Markov (Beaumont and Balding, 2004) ou l'analyse d'une valeur moyenne des Fst sur des fenêtres regroupant plusieurs marqueurs (Weir, *et al*., 2005).

D'autres méthodes inter-populationnelles ont été développées telle que l'approche du XP-EHH (Cross Population Extended Haplotype Homozygosity) (Sabeti, *et al*., 2007) qui se base sur la comparaison des haplotypes de deux populations pour détecter des balayages sélectifs forts. La comparaison de l'hétérozygotie relative entre deux populations peut également permettre de détecter une région génomique qui aurait perdu plus de diversité dans une population que dans une autre. Enfin d'autres méthodes permettent de caractériser plus en détail la structure de la population à partir des données de régions génomiques. Par exemple, l'analyse en composante principale, qui permet de déterminer des distances entres les individus sur la base des génotypes d'une région du génome (Chakraborty and Jin, 1993; Patterson Nick, *et al*., 2006). Cette approche a ainsi permis de déterminer que l'haplotype portant le gène IGF1 associé au phénotype 'petite taille' chez le chien est proche de l'haplotype des loups gris du moyen orient (Gray M. M., *et al*., 2010).

II.3.2. L'estimation de la neutralité

Il est essentiel de connaître l'évolution neutre pour estimer les variations qui vont indiquer une différence significative par rapport à la neutralité. Des tests existent pour estimer si les polymorphismes présents dans une population relèvent de la neutralité ou sont liés à de la sélection. Le test du 'D' de Tajima (Tajima, 1989) compare une région de différents individus. Lorsque la région considérée n'est pas sous sélection, le nombre de polymorphismes doit être égal à la moyenne du nombre de différences entre individus pris deux à deux. Un excès de différence entre individus pris deux à deux par rapport au nombre

de polymorphismes indique un excès d'allèles à fréquence intermédiaire et donc un goulet d'étranglement ou une sélection balancée récente. Au contraire un excès de polymorphismes indique un excès de variants rares et une sélection directionnelle négative ou positive. À la suite du test du 'D' de Tajima, d'autres statistiques basées sur le polymorphisme d'une seule population ont été développées telle que le 'F' de Fu et Li (Fu and W.H., 1993; Fu, 1997) qui permet lui aussi de détecter l'excès soit d'une sélection balancée, soit d'une sélection directionnelle. Le 'H' de Fay et Wu (Fay and Wu, 2000) permet de rechercher des mutations à forte fréquence et de détecter la sélection positive, à condition de disposer d'une estimation de l'haplotype ancestral de la région testée. Enfin le test 'E' de Zeng et al. (Zeng, *et al*., 2006) permet de détecter les balayages sélectifs récents qui commencent à rétablir leur polymorphisme.

Chacun de ces tests se focalise sur le patron de polymorphisme et n'exploite pas ou peu les informations issues d'autres populations comme les tests de différenciations génétiques ni n'utilisent les informations d'autres espèces comme le test de McDonald Kreitman. Le test de McDonald Kreitman (McDonald and Kreitman, 1991) repose sur la comparaison des séquences codantes séquencées chez plusieurs individus avec la séquence codante d'une espèce voisine. Dans le cas de la neutralité, les proportions de mutations synonymes et non-synonymes fixées entre les deux espèces doivent être les mêmes que les proportions de mutations synonymes et non-synonymes observées dans le polymorphisme de la population. Lorsque l'on a accès aux informations de séquence génomique des espèces, d'autres approches de génomique comparative permettent de détecter les signatures de la sélection naturelle moteur de l'évolution des espèces.

II.4. L'évolution des espèces

II.4.1. Évaluation de la sélection

L'évolution des populations par sélection naturelle et par dérive génétique peut conduire à la création de nouvelles espèces à partir d'une population ancestrale. La spéciation qui en résulte va se traduire par l'apparition et l'extinction d'espèces et par la stabilisation puis le changement des caractères de ces espèces. La création de nouvelles espèces dans une branche de l'arbre de la vie peut être reconstituée en réalisant des phylogénies des espèces existantes. Ces phylogénies ne permettent pas de reconstituer les extinctions des espèces mais permettent à partir des espèces actuelles, de comparer et d'évaluer les modifications génétiques liées à l'évolution des lignées considérées. Ces lignées présentent des phases pendant lesquelles leurs caractères semblent figés et des phases de changement de leurs

caractères morphologiques (Gould and Eldredge, 1977; Venditti, *et al*., 2011). Les changements d'une lignée peuvent être lents ou rapides, discrets ou importants. Les modifications génétiques peuvent être dues à des créations ou des pertes de gènes ou à des mutations dans les gènes responsables de critères déjà présents dans la lignée.

La connaissance des séquences des génomes permet de comparer et de rechercher les événements de pertes de gènes, comme notre laboratoire l'a réalisé pour le génome du chien (Derrien, *et al*., 2009) ou de rechercher la sélection positive qui a opéré sur les gènes codant pour des protéines. La comparaison des séquences des génomes permet de comparer les séquences codantes des gènes orthologues codant pour des protéines. Pour détecter la sélection positive entre espèce, il est alors possible de comparer les taux des substitutions qui changent ou non la séquence primaire de la protéine.

II.4.2. La génomique comparative

Historiquement, le terme "d'homologie" était utilisé en anatomie pour désigner le même organe chez différents animaux (Owen, 1848). Plus tard, la terminologie de gène homologue a été utilisée pour désigner des gènes ayant une origine ancestrale commune (Huxley, 1860). Dans les années 1970, Walter Fitch (Fitch, 1970) précisa la notion d'homologie en introduisant le terme de gènes orthologues, c'est-à-dire des gènes possédant une origine commune (homologues) mais ayant été séparés par un événement de spéciation et le terme de gènes paralogues faisant référence aux gènes homologues séparés par un événement de duplication.

Pour rechercher les signatures de la sélection naturelle entre espèces dans les séquences protéiques, nous nous intéressons aux gènes ayant une relation d'orthologie de type 1:1 soit un gène présent dans une espèce qui peut être mis en correspondance avec un et un seul gène présent dans une autre espèce. L'intérêt de sélectionner des gènes en relation 1:1 entre plusieurs espèces est de rechercher la sélection positive là ou elle aura un impact à priori important. En effet les gènes conservés en un seul exemplaire dans différentes espèces sont à priori sous des contraintes sélectives plus fortes que les gènes en relation d'orthologie multiple (1:n ou n:m). Pour les orthologues 1:1, il est attendu qu'une modification dans la séquence de l'un de ces gènes peut avoir un effet important susceptible de participer aux phénomènes d'adaptation à l'environnement et de spéciation au cours de l'évolution. Il est possible alors de détecter le contraste entre un gène sous sélection positive dans une espèce et ce même gène sous sélection négative dans les autres espèces.

La principale méthode pour identifier les gènes orthologues entre deux espèces est la méthode basée sur les "meilleurs alignements réciproques" des séquences développée par Fitch (Fitch, 1970, 1995). Les séquences de gènes orthologues présentent plus de similitudes entre elles que de similitudes avec une autre séquence du génome initial. Ainsi, deux gènes A et B, appartenant respectivement aux espèces E_A et E_B seront considérés orthologues si le meilleur alignement de la séquence du gène A sur le génome de E_B correspond au gène B et, réciproquement, si le meilleur alignement de la séquence du gène B sur le génome de E_A correspond au gène A.

II.4.3. Évaluation de la pression de sélection par dN/dS

La disponibilité d'alignements de séquences de gènes en relation d'orthologie 1:1 dans différentes espèces permet de mesurer les taux de substitution synonymes -dS- et non-synonymes -dN- et le ratio dN/dS -appelé aussi ω- des taux de substitution. Le taux de substitution synonyme -dS- est le nombre de mutations synonymes observées dans la séquence codante du gène dans l'espèce considérée divisé par le nombre de mutations synonymes qui peut théoriquement se produire; les mutations synonymes étant à priori neutre sélectivement, ce taux représente l'évolution de la protéine en l'absence de pression de sélection. Le taux de substitution non-synonyme -dN- est le nombre de mutations non-synonymes observé dans la séquence codante du gène dans l'espèce considérée divisé par le nombre de mutations non-synonymes théorique ; ce taux est largement influencé par les pressions de sélection négative ou positive. Le ratio dN/dS constitue une mesure des forces sélectives agissant sur les gènes. Une valeur de dN/dS ~ 1 reflète une évolution neutre, une valeur de dN/dS < 1 correspond à une sélection négative alors qu'un dN/dS > 1 met en évidence une sélection positive puisque la fixation des mutations non-synonymes est plus fréquente que par le processus de dérive génétique.

En pratique, les gènes fortement conservés au cours de l'évolution n'auront jamais une valeur de dN/dS > 1. On utilise soit des méthodes basées sur des modèles Markoviens pour évaluer la sélection des différents codons (Robinson, 2003; Ouyang and Liang, 2007), soit des méthodes de maximum de vraisemblance pour calculer le nombre de site ayant un dN/dS <<1, ~1 et >1 (Li, *et al*., 1985; Goldman and Yang, 1994) à partir des alignements multiples des séquences. Un test de rapport des vraisemblances ou "Likelihood Ratio Test" –LRT strict (Yang Z and Nielsen, 1998; Zhang Jianzhi, *et al*., 2005b; Yang Z and Dos Reis, 2011) est utilisé pour tester la présence de sélection positive dans un gène donné. L'identification des cibles de la sélection positive au sein d'une espèce par approche dN/dS a pour objectif

d'identifier les gènes et les réseaux de gènes impliqués au cours de l'évolution naturelle de l'espèce considérée.

III. Le modèle génétique canin

Le chien, *Canis lupus familiaris*, présente une formidable diversité phénotypique acquise depuis sa domestication sous l'emprise de l'Homme et constitue un modèle génétique pour élucider les causes génétiques de traits phénotypiques et un modèle pour les études de génétique des populations et de l'évolution des espèces.

III.1. Le chien, modèle de diversité phénotypique

Le chien est la première espèce domestiquée et l'espèce qui a été le plus façonnée par l'Homme. Elle est également une espèce très répandue puisque elle accompagne les populations humaines sur tous les continents et sous tous les climats. Le Livre des Origines Français (LOF), qui est le livre généalogique officiel pour l'espèce canine comporte plus de 350 races toutes différentes entre elles en terme de morphologie (Figure 3), comportement, aptitude et physiologie. Les différences morphologiques entre races canines dépassent largement celles existant au sein de tout autres espèces mammifères. Par exemple, un chihuahua de 1 kg et 20 cm de haut et un terre-neuve de 60 kg et 80 cm de haut représentent des morphotypes extrêmes. Ces deux races, appartenant pourtant à la même espèce, diffèrent d'un facteur 60 pour leur poids et d'un facteur 4 pour leur taille. Les différences de taille et de poids sont les plus évidentes, mais les races se distinguent sur de nombreux autres traits morphologiques comme la forme du crâne, des oreilles, du museau, de la queue, des variations sur la nature des poils (longs, courts, frisés ou absents) ou leurs couleurs. Certains comportements sont héritables dans les races sélectionnées à cet effet et illustrent là encore la diversité des races canines (Kukekova*, et al*., 2006) : Certaines races sont de très bons chiens de berger, d'autres des chiens de garde, chiens de chasse ou encore de compagnie. De nombreuses races présentent des aptitudes particulières, par exemple les aptitudes d'apprentissage associées au caractère calme et au besoin de plaire au maître du labrador et du golden retriever en font de bons guides pour les personnes handicapées ou déficientes visuelles, ce qui n'est pas le cas du border collie qui est en revanche particulièrement adapté pour la garde de troupeaux. Enfin la diversité canine est aussi d'ordre physiologique, comme la différence significative de pression artérielle observée entre races (Bodey and Michell, 1996; Bright and Dentino, 2002), ou bien les différences de susceptibilités au maladies et de

vieillissement qui font que l'espérance de vie moyenne des races peut varier de 6 à 14 ans (Michell, 1999; Galis, *et al*., 2007).

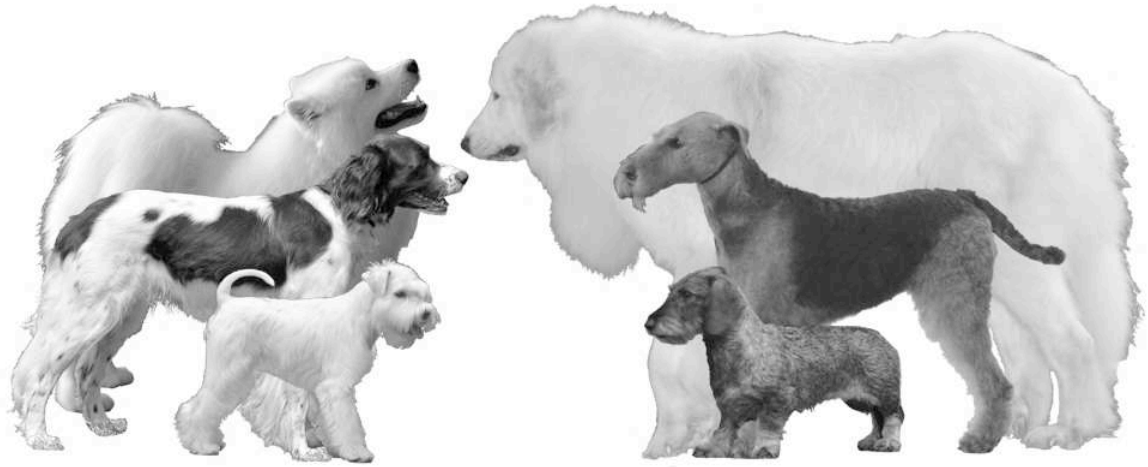


Figure 3 : ***Illustration de la diversité phénotypique de l'espèce canine.*** *qui en fait un modèle génétique particulièrement puissant.*

Les explications de l'hypervariabilité du phénotype de l'espèce canine qui s'observe par les différences de taille, de poids, de morphologie et de comportement observées entre les 350 races de chien se trouvent d'abord dans l'histoire de la domestication du loup par l'Homme et de sa transformation progressive au cours des millénaires puis de la diversification rapide par l'Homme des races canines au cours des derniers siècles.

III.1.1. La domestication

Si l'on sait aujourd'hui que tous les chiens descendent du loup, au XIXème siècle, cela n'a pas toujours été le cas: Charles Darwin pensait que le loup, le chacal et le coyote avaient été hybridés pour obtenir les différentes races de chien (Darwin, 1860). Les données actuelles sur l'origine de la domestication du chien sont en parfaite corrélation avec celles déjà énoncées par J. Clutton-Brock (Clutton-Brock, 1995) sur l'origine commune de tous les chiens à partir de loups domestiqués. Le loup est un mammifère de l'ordre des carnivores appartenant à la famille des *Canidae* qui regroupe 34 espèces (chacal, chien, coyote, dingo, loup, renard ...) issues d'un ancêtre commun carnivore il y a 10 millions d'années (figure 4). Toutes les espèces sont interfécondes entre elles (Wayne and Vilà, 2001).

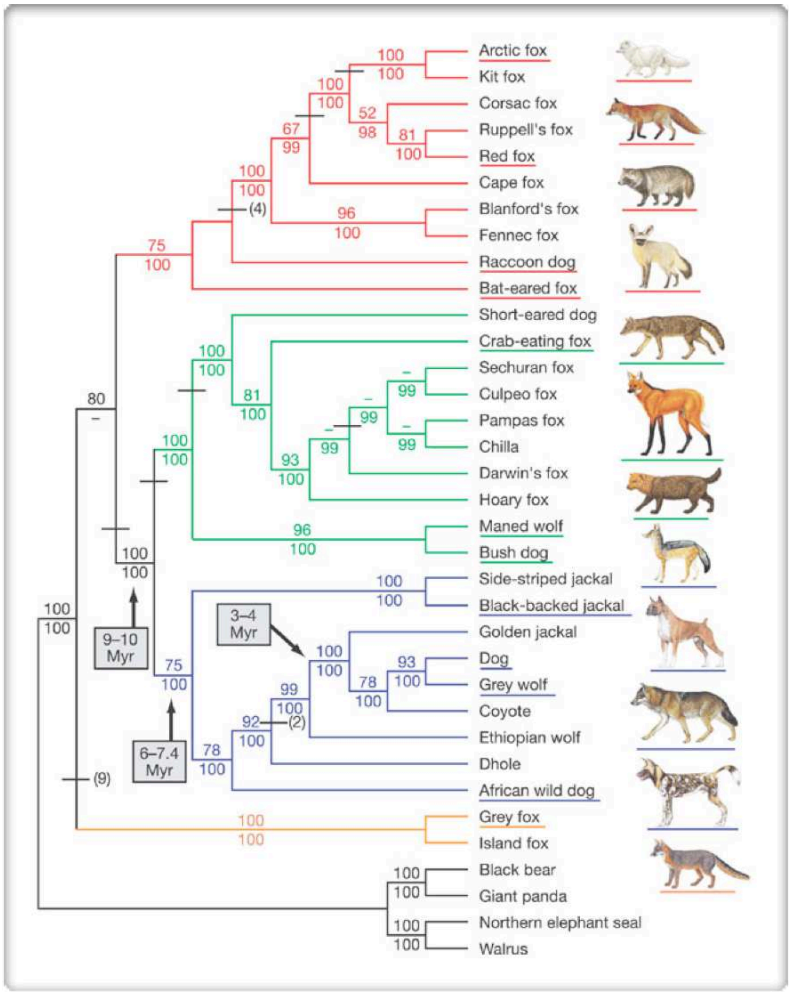


Figure 4 : ***Phylogénie des canidés.*** *Cet arbre phylogénétique est basé sur l'analyse de 15 kb de séquences introniques et exoniques. Les couleurs des 4 branches correspondent aux 4 grands taxons de canidés identifiés. Cet arbre, construit sur des critères de parcimonie, présente les valeurs de bootstrap pour chaque embranchement et le temps de divergence en millions d'années (Myr). Les espèces dont le nom est surligné dans l'arbre sont illustrées sur la droite (Lindblad-Toh, et al., 2005).*

Plusieurs travaux ont analysé la datation et la localisation du processus de domestication du chien. En comparant l'ADN mitochondrial de plusieurs espèces du genre *Canis*, il semblerait qu'un nombre limité de loups femelles d'Asie orientale soient les ancêtres directs des chiens domestiques actuels (Vila, *et al*., 1999a; Vila, *et al*., 1999b; Leonard, *et al*., 2002). Cependant, des travaux plus récents basés sur des polymorphismes nucléotidiques de l'ADN nucléaire indiqueraient une origine moyen-orientale du chien (Vonholdt, *et al*., 2010). En utilisant les fréquences d'allèles d'un locus du Complexe Majeur d'Histocompatibilité (CMH issu de l'ADN nucléaire) entre des populations de loups et des chiens, Vila et al. (Vilà, *et al*., 2005) proposent qu'un grand nombre de croisements (intentionnels ou accidentels) entre les populations sauvages et domestiquées auraient eu lieu pour toutes les espèces mammifères domestiquées, ce phénomène est d'ailleurs observé au

delà des espèces mammifères (Berthouly, *et al*., 2009). Cette hypothèse permettrait aussi d'expliquer la grande diversité génétique observée actuellement chez les espèces domestiquées. La datation du processus de domestication du chien par des recherches complémentaires en archéologie et en analyses moléculaires de séquences semblent indiquer un événement de domestication unique autour de la période de 15.000 ans avant notre ère (Leonard, *et al*., 2002; Savolainen, *et al*., 2002), bien que des données récentes de paléogénomique suggèrent l'existence de divers foyers de domestication au paléolithique (communication personnelle de Catherine Hänni). Ceci fait du chien la première espèce domestiquée (espèces animales et végétales réunies) et la seule domestiquée au paléolithique.

Pour pouvoir apprivoiser une espèce sauvage, l'Homme doit cohabiter avec cette espèce afin de pouvoir 'prélever' un ou plusieurs animaux. La domestication du chien a donc été possible dans les régions où le loup était présent. Le loup est une espèce capable de s'adapter à de grandes variations de milieu et est donc présent dans de nombreux écosystèmes de l'hémisphère nord (Mech and Boitani, 2003). Ainsi le loup était, lors de sa domestication, une espèce présente dans les écosystèmes dans lesquels les sociétés humaines évoluaient. Les deux espèces étaient des prédateurs aguerris, privilégiant les grands herbivores et pouvaient chasser en groupes familiaux (Mech, 1970). La répartition démographique et les comportements sociaux et alimentaires ont impliqué une cohabitation entre le loup et l'Homme. En Asie du sud-est, cette cohabitation a entraîné un rapprochement entre les deux espèces et a amené les sociétés humaines à percevoir l'utilité de posséder des loups, que ce soit pour la chasse, la garde ou autres et ainsi, au fil des générations, à obtenir une espèce domestique distincte de son espèce ancestrale. La domestication a eu un impact considérable sur le mode de vie des chiens comparé à leurs ancêtres les loups gris. Une étude de Björnerfeldt et al. a mis en évidence un mécanisme de relaxation des contraintes sélectives suivant la période de domestication du chien (Bjornerfeldt, *et al*., 2006). En effet, les auteurs ont montré que le taux d'accumulation des substitutions non synonymes des gènes mitochondriaux était significativement plus rapide chez le chien comparé à celui des loups. Si le relâchement des forces sélectives a aussi agi sur le génome nucléaire, cela pourrait aussi, en partie, corréler avec la fascinante diversité phénotypique observée chez le chien.

Le processus de domestication est le premier goulet d'étranglement de l'histoire évolutive du chien et a eu pour conséquence une dérive génétique à partir d'un pool relativement restreint d'allèles (Lindblad-Toh, *et al*., 2005). Ceci a engendré, dans la plupart des races, des changements morphologiques et comportementaux éloignant peu à peu le chien du loup.

III.1.2. La création des races canines.

Au fil de l'histoire, les morphologies canines se sont diversifiées avec l'apparition de différences de taille, de forme du crâne, de carrure, de forme des oreilles, etc. qui représentent une diversité phénotypique à partir de laquelle les races actuelles ont été créées. Le second goulet d'étranglement que l'espèce canine a connu est récent. Il s'agit des 200-300 dernières années de sélection artificielle (soit 100 à 150 générations) qui ont abouti à la création des races actuelles (figure 5) (Galibert and André, 2008). En effet le XIXème siècle a vu la naissance de la cynophilie avec, en France, la création de la Société Centrale Canine (SCC) en 1882 et l'ouverture du premier Livre des Origines Français (L.O.F) en 1885. Ceci a permis de définir des standards pour chaque race et de définir la règle générale de séparation des races qui stipule que pour enregistrer un chien au LOF, les deux parents de ce chien doivent appartenir à cette race et être enregistrés au livre des origines, la saillie doit être déclarée, et que le chien doit être examiné et certifié conforme au standard de la race à partir de 10 à15 mois selon les races. Ainsi les races canines se sont multipliées en Europe à la fin du XIXème (Parker and Ostrander, 2005) avec une standardisation qui a entraîné une sélection intensive. Cette sélection artificielle a favorisé les individus correspondant le mieux au standard de race avec notamment l'utilisation "d'étalon champion", mâles correspondants parfaitement aux critères de sélection et fortement utilisés en reproduction.

Ces pratiques intenses d'élevage et de sélection ont eu pour première conséquence l'homogénéisation des races, leur consanguinité et la fixation des différences de comportement, de morphologie et d'aptitude entre elles. Chaque race présente donc une variabilité génétique beaucoup plus faible que la population canine générale. La seconde conséquence est que chaque race présente une très forte isolation génétique (Parker*, et al*., 2004). Par ailleurs, dans certaines races, on observe un goulet d'étranglement supplémentaire lié à un nombre limité d'individus ayant permis de créer, de maintenir ou de reconstituer une race. Par exemple, la race léonberg est issue de seulement cinq mâles rescapés de la première guerre mondiale (Ostrander and Kruglyak, 2000), et les épagneuls tibétains actuels proviennent de trois individus mâles ayant survécu à la seconde guerre mondiale (Vila*, et al*., 1999b). Certaines races canines «éteintes» ont également été reconstituées à partir de deux ou trois chiens et de croisements avec des chiens au phénotype semblable. C'est notamment le cas des braques du bourbonnais (communication personnelle du président du club). En parallèle de la fixation des traits d'intérêt, l'augmentation de la consanguinité au sein de chaque race et la réduction des flux géniques a fixé des variants de gènes responsables de

maladies génétiques de telle sorte que toutes les races sont atteintes de maladies génétiques qui affectent tout type de tissus ou de fonctions (Galibert and André, 2008).

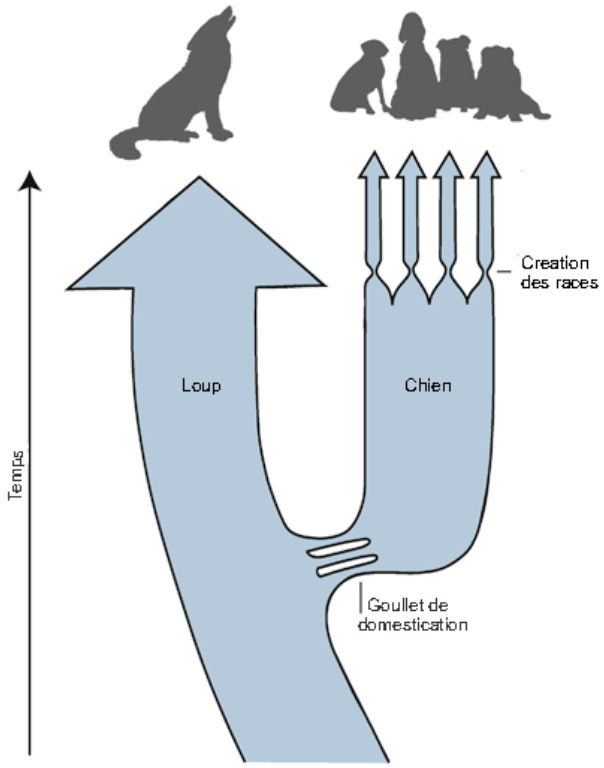


Figure 5 : ***Les deux goulets d'étranglement de l'histoire du chien*** *-adapté de (Karlsson and Lindblad-Toh, 2008)-. Le premier goulet correspond à la domestication du loup par l'Homme puis, plus récemment, la création des races modernes.*

III.1.3. Structure actuelle de la population canine

Aujourd'hui l'espèce canine est unique tant par l'étendue de sa diversité phénotypique, que par l'homogénéité observée au sein de chaque race (Galibert, *et al*., 2004). Les 350 races reconnues par la Fédération Cynologique Internationale (FCI) sont classées en 10 groupes selon leurs morphologies et aptitudes; ce qui implique que des races très éloignées telles que le chihuahua, le caniche ou le bouledogue français se soient retrouvées dans le groupe IX, rassemblant les chiens de compagnie. Le corollaire de la perte de diversité de chaque race canine, est qu'elles sont autant d'isolats génétiques (Figure 6). Dans une moindre mesure, les populations humaines insulaires, islandaise par exemple, ou isolées par l'histoire ou la tradition (Amish, Ashkénaze) constituent des formes d'isolats génétiques. En tant que population isolée et consanguine, chaque race de chien présente une perte de l'équilibre Hardy-Weinberg d'autant plus prononcé que la race est ancienne et la taille de la population

effective est petite (Irion, *et al*., 2003). Les races canines présentent un déséquilibre de liaison allant jusqu'à 1000 kb en moyenne 50 fois supérieur à celui de l'espèce humaine (Lindblad-Toh, *et al*., 2005). Une conséquence très intéressante de la réduction de la variabilité au sein d'une race et de l'extrême diversité entre races est la possibilité d'établir des corrélations entre le phénotype et le génotype plus facilement chez le chien que chez l'Homme. Ces corrélations pouvant être établies sur des traits qui ségrègent au sein d'une race comme la couleur des labradors, ou la taille des caniches, pour lesquelles les études profitent d'une variabilité phénotypique qui opère au sein de la structure génétique homogène de la race. Les corrélations peuvent être recherchées entre le génotype et des traits fixés différenciellement entre plusieurs races comme le travail réalisé par Cadieu et al (Cadieu, *et al*., 2009) dans leur étude sur l'identification des gènes responsables de la texture et la longueur du pelage. Il s'agit de tirer partie de la diversité et de la structure de la population canine en général.

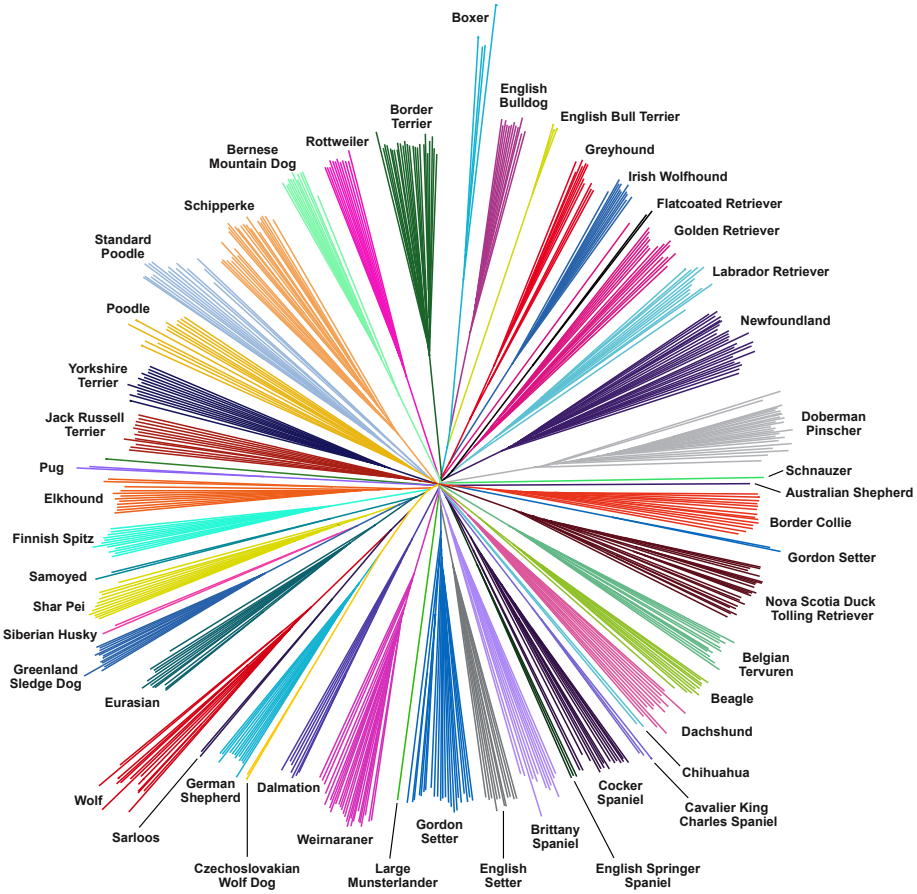


Figure 6 : ***Classification par proximité génétique de 46 races de chien plus le loup****. réalisée à partir du génotypage de plus de 170.000 SNP. Chaque race est bien génétiquement séparée des autres (Vaysse, et al., 2011)*

III.2. Le génome canin

C'est la forte prévalence des maladies génétiques chez le chien dont la plupart sont homologues aux maladies humaines et la structuration particulière de sa population qui ont été déterminantes dans la motivation de plusieurs équipes de recherche dont la notre à développer des outils d'analyse du génome canin. Ces outils sont un préalable nécessaire à l'analyse exhaustive du génome tel que l'identification des signatures génétiques de différenciation de races. Le génome canin est composé de 78 chromosomes (figure 7) : 38 paires d'autosomes acrocentriques et une paire de chromosomes sexuels, le chromosome X étant le plus grand (128 Mb) et le chromosome Y le plus petit du caryotype (27 Mb), la plupart (de 22 à 38) de ces chromosomes sont de petite taille et ont un profil de bandes chromosomiques très similaire (Switonski, *et al*., 1996). Les techniques d'hybridation fluorescentes *in situ* (FISH) ont permis d'orienter de façon univoque et de ranger par paires tous les chromosomes du caryotype (Breen, *et al*., 1999). Les chromosomes canins sont décrits par l'acronyme CFA dérivé de '*Canis familiaris*'. Le développement des outils moléculaires d'étude du génome canin ont commencé dès 1995, bien avant l'avènement des techniques de séquençage massif. Les premiers outils développés ont été les cartes du génome, c'est à dire le balisage de tous les chromosomes par des marqueurs génomiques.

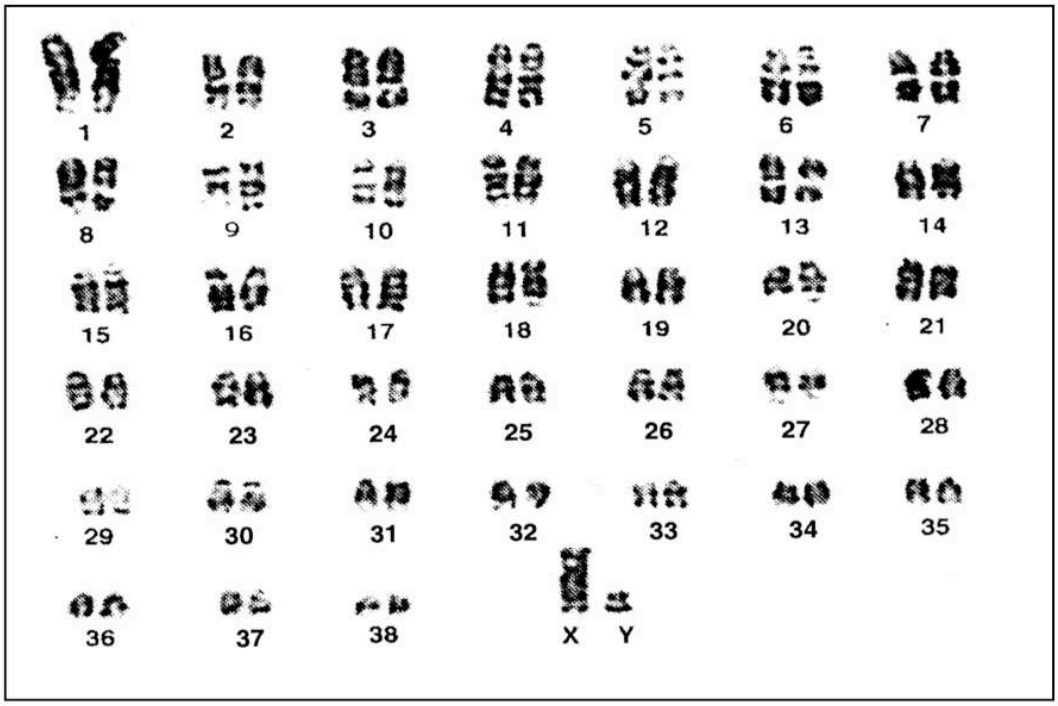


Figure 7 : ***Caryotype canin.*** *Le génome diploïde du chien se compose de 78 chromosomes dont 38 paires d'autosomes et une paire de chromosomes sexuels X et Y. Tous les autosomes sont acrocentriques et certains d'entre eux présentent des tailles et des profils de bandes similaires (Langford, et al., 1996).*

III.2.1. La cartographie du génome canin

Les premiers projets de cartographie du génome canin ont débuté en 1995 lorsque notre laboratoire intéressé par l'originalité et l'intérêt du modèle canin comme modèle génétique pour la médecine, a développé les outils moléculaires nécessaires à la construction d'une carte du génome par la méthode des hybrides irradiés (RH). Le positionnement et l'ordonnancement des marqueurs sur les 39 paires de chromosomes canins consiste à estimer les distances les séparant par l'intermédiaire de la fréquence de présence sur un même fragment obtenu par cassure après irradiation. En pratique, le génome de fibroblastes canins est fragmenté par irradiation, ces cellules sont fusionnées avec des cellules CHO issues de fibroblastes de hamster et dépourvues du gène de la thymidine kinase (TK) qui est indispensable à la survie des cellules en milieu sélectif HAT. Les cellules qui ont incorporé une partie du génome canin contenant l'orthologue du gène TK forment des lignées hybrides cellulaires (Vignaux, *et al*., 1999). Des marqueurs proches sur le génome se retrouveront fréquemment sur le même fragment chromosomique après irradiation et seront donc plus fréquemment retenus dans les mêmes lignées que des marqueurs distants. Cette technique a permis de réaliser la première carte RH du génome canin dès 1998 qui positionne 400 marqueurs, assurant une couverture d'environ 80% du génome (Priat, *et al*., 1998). À cette époque, le chien est, après la souris, le deuxième "modèle animal" à disposer d'une carte d'hybrides irradiés.

En 1998, deux projets, menés en parallèle, ont eu pour but de développer des cartes génétiques aboutissant à la localisation d'une centaine de marqueurs sur les chromosomes canins. La cartographie génétique évalue les distances entre marqueurs en suivant la transmission de marqueurs polymorphes au sein d'une famille sur plusieurs générations. En effet, plus la distance qui sépare deux marqueurs sur un même chromosome est grande, plus la probabilité de recombinaison au cours de la méiose entre ces deux marqueurs est proche de 0,5. Les distances entre marqueurs sont exprimées en centimorgans (cM), un cM indique que l'on observe une recombinaison entre les marqueurs lors d'une méiose sur 100. La première carte génétique, issue de l'initiative européenne « DogMap », comprenait 94 microsatellites (Lingaas, *et al*., 1997). De son côté, le groupe du Dr. Elaine Ostrander aux Etats-Unis, publia la même année une carte génétique de 150 marqueurs (Mellersh, *et al*., 1997). Ces deux équipes ont ultérieurement densifié leur carte. En 1999, la carte de l'équipe d'E. Ostrander contenait 341 marqueurs distants en moyenne de 9 cM (Werner, *et al*., 1999). L'existence de ces deux types de cartes génomiques a poussé les différents laboratoires "cartographes" du génome du chien à collaborer afin de créer, en 2000, la première carte d'intégration entre les

cartes génétiques et la carte issue d'hybrides irradiés qui a pu être établie suite à une étroite collaboration entre notre groupe et celui d'Elaine Ostrander (Mellersh, *et al*., 2000). De la même façon, Breen et al ont rassemblé conjointement un ensemble de marqueurs afin de disposer d'une carte intégrée du génome du chien contenant 302 marqueurs positionnés par la technique d'hybridation fluorescente *in situ* (FISH), 354 marqueurs de liaison génétique et 1500 marqueurs positionnés par la méthode RH (Breen, *et al*., 2001). Par la suite, les techniques de cartographie étant maîtrisées, les efforts se sont poursuivis afin de densifier les cartes. En 2003, une nouvelle carte RH a été produite au laboratoire et a permis de positionner 3270 marqueurs, comprenant 1596 microsatellites, 900 marqueurs de gènes, 668 marqueurs de BAC et 106 STS, avec une moyenne de 1 Mb entre deux marqueurs, assurant une couverture de 95% du génome canin (Guyon, *et al*., 2003b). En 2004, un sous ensemble de 804 BACs (Bacterial Artificial Chromosome) a aussi été cartographié par cytogénétique par l'équipe de Matthew Breen renforçant la complémentarité et l'apport d'information des cartes intégrées RH/FISH pour aboutir à une carte comportant 4249 marqueurs d'une densité d'un marqueur tous les 900 kb (Breen, *et al*., 2004). Depuis 2005, l'équipe de Mark Neff aux Etats-Unis développe une carte génétique plus dense. En septembre 2008, celle-ci contenait 3075 marqueurs espacés en moyenne de 0,7 cM (http://www.vgl.ucdavis.edu/research/canine/projects/linkage_map/data/) (Wong, *et al*., 2010).

Enfin en 2005, l'utilisation des données de séquençage léger (1,5X) du génome d'un caniche (Kirkness, *et al*., 2003) dans la réalisation d'une carte RH de haute densité a permis au laboratoire de répondre à trois objectifs: (i) le positionnement de 10.000 marqueurs de gènes sur l'ensemble des chromosomes canins, (ii) l'analyse comparée résolutive de la structure des génomes canin et humain et (iii) l'optimisation de l'assemblage du séquençage profond (7.5x) du génome canin (Hitte, *et al*., 2005; Hitte, *et al*., 2008).

III.2.2. La séquence du génome canin

Grâce à son intérêt en tant que modèle génétique et en tant que représentant du clade mammifère "Laurasiatheria", le chien a été choisi pour être la troisième espèce mammifère séquencée après le clade des Euarchontoglires auquel appartiennent l'Homme et la souris. Le premier séquençage a été initié en 2001 par la société américaine Celera et analysé par le TIGR (The Institute for Genomic Research). Le chien sélectionné était un caniche 'moyen' et la séquence établie à une profondeur de 1,5X ce qui signifie que chaque nucléotide est séquencé en moyenne 1,5 fois (Kirkness, *et al*., 2003). Ce projet a permis de confirmer que la séparation des carnivores a été plus précoce que celle des rongeurs (~90 MA vs 80 MA) et

d'identifier via les cartes RH, les blocs de synténie conservés par la construction des cartes comparées Homme/chien très denses. Ce séquençage a révélé qu'au niveau nucléotidique, le génome canin présente un taux de substitution comparable à celui de l'Homme alors qu'il est 1,6 fois supérieur chez la souris que chez le chien. De plus, environ 25% de séquences non-répétées du génome canin s'alignent à cette profondeur avec le génome humain (cette valeur est à mettre en correspondance avec les 40% de séquence murine issue d'un séquençage profond s'alignant avec le génome humain). De plus ce projet a permis d'identifier près de 18.500 fragments de gènes orthologues avec les gènes humains.

Cependant, ce type de séquençage dit 'léger' présente des limites car par nature discontinu, peu de gènes sont entièrement séquencés et plus de 20% de ces gènes sont totalement absents de la séquence. Ces limites ont motivé la communauté scientifique canine à rédiger en 2002 un «Livre blanc» à l'intention du National Institute of Health (NIH) en faveur d'une analyse approfondie de la séquence du génome canin. En 2004, le NIH, lance l'initiative d'un séquençage complet du génome du chien. Ce projet a été mené par le BROAD Institute de Cambridge (USA) à partir d'un chien, 'Tasha', une femelle Boxer. Des travaux préalables (Parker*, et al*., 2004) avaient démontré le faible taux d'hétérozygotie de cette race. La faible variabilité génétique est, en effet, considérée comme avantageuse pour la phase d'assemblage. La possibilité de disposer de deux copies du chromosome X renforce l'intérêt de travailler à partir d'un organisme femelle, au détriment du séquençage du chromosome Y. Au final, 36 millions de séquences, réparties en 20.000 contigs et 87 super-contig couvrant les 39 chromosomes, ont été produites afin d'obtenir un séquençage complet avec une couverture de 7,5X. Les cartes RH développées au laboratoire (Breen*, et al*., 2004; Hitte*, et al*., 2005) ont, d'une part, servi d'ossature à l'assemblage des milliers de contigs générés, et d'autre part, ont permis de guider l'orientation des super-contigs le long des chromosomes. La taille totale du génome canin est déterminée à 2,41 Gb dont 2,38 Gb de séquences nucléotidiques, les 1% restant correspondant à des 'trous' de séquence ("gaps"). L'assemblage réalisé en mai 2005 (CanFam2.0) du génome canin représente une première ressource formidable pour exploiter le potentiel du modèle génétique canin. Il permet de lister les gènes et de déterminer leurs structures exoniques/introniques à partir de logiciels de prédiction et des alignements d'ADN complémentaires canins ou d'autres espèces. Il permet de manière plus générale de décrire la structure du génome canin. C'est la connaissance de cette séquence qui permet les études de génomique comparative entre le chien et les autres espèces.

III.2.3. La structure du génome canin

La taille totale du génome canin, estimée à environ 2,41 Gb, est sensiblement inférieure à celle des génomes de l'Homme et de la souris (Wade, *et al*., 2006). Ceci s'explique en partie par la plus faible présence de séquences répétées chez le chien par rapport à ces deux autres espèces. En effet, les éléments répétés constituent respectivement 31% du génome canin, 38% du génome murin et 46% du génome humain (Kirkness, *et al*., 2003).

III.2.3.1. Eléments répétés

Parmi les séquences répétées du génome canin, on observe la présence d'une seule famille de SINE (Short INterspersed Element) qui a été mise en évidence dans certaines maladies et de nombreux microsatellites qui sont utiles dans les analyses de liaison génétique. Les SINE sont des éléments rétrotransposables d'environ 200 paires de bases (pb). Une famille de SINE, SINE_Cf, est particulièrement présente dans le génome canin. En effet, 50% des gènes canins annotés contiennent au moins un SINE_Cf dans leurs séquences introniques et 14% en contiennent un polymorphique, c'est-à-dire absent sur certains chromosomes de la population mais présent dans d'autres (Wang W. and Kirkness, 2005). En effet la comparaison avec la séquence 1.5X du caniche a mis en évidence plus de 10.000 locus polymorphes. À titre de comparaison, le nombre d'insertions bimorphiques de SINE chez l'Homme est inférieur à 1000. L'insertion d'un SINE dans un exon, dans une séquence régulatrice ou dans un intron peut entraîner de nombreuses répercutions sur la nature et la régulation du transcrit. L'effet des SINE_Cf a déjà été mis en évidence dans plusieurs maladies génétiques comme la narcolepsie (Lin L., *et al*., 1999), la myopathie centronucléaire du labrador retriever (Pele, *et al*., 2005) ou la couleur de robe merle (Clark, *et al*., 2006; Hédan, *et al*., 2006). Il est fortement envisageable que cette famille de séquences répétées ait un rôle important dans la plasticité et la dynamique du génome canin (Kirkness, *et al*., 2003).

Les microsatellites sont de courtes séquences composées de 2 à 6 nucléotides, répétées en tandem et flanquées de séquences uniques. Le polymorphisme des microsatellites est un polymorphisme de taille dû à la variation du nombre de répétitions du motif de base. Ces séquences sont dispersées tout le long du génome à raison d'un microsatellite tous les 47 kb en moyenne chez le chien (Ostrander, *et al*., 1993; Jouquand, *et al*., 2000). Les microsatellites ont permis la caractérisation génétique de la structure de la population canine avec comme unité la race et une structuration très faible au niveau des groupes de races (Irion, *et al*., 2003; Parker, *et al*., 2004; Leroy, *et al*., 2009). En 2007 un ensemble composé de 507 marqueurs

microsatellites a été établi de façon à proposer un jeu de marqueurs disposés tous les 5 Mb en moyenne (Sargan, *et al.*, 2007).

III.2.3.2. Génomique comparative

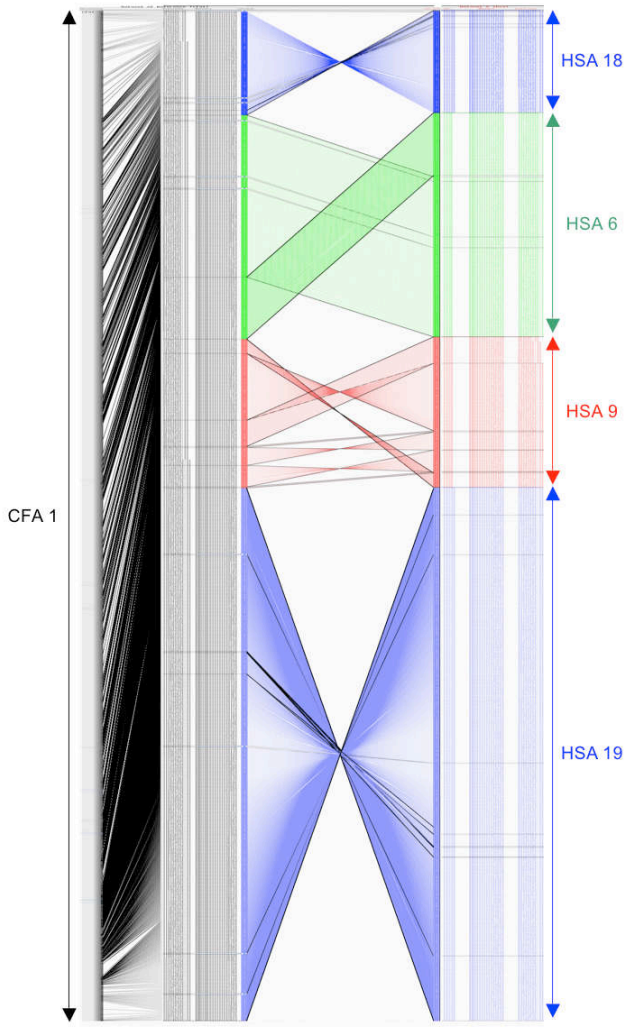


*Figure 8 : **Carte de synténie construite par le programme AutoGRAPH** entre le chromosome 1 du chien et les chromosomes orthologues humains Le chromosome 1 du chien (CFA 1) se situe à gauche et les relations d'orthologies avec les séquences humaines, à droite, sont représentées par des traits colorés en fonction de l'appartenance aux chromosomes humains. Ainsi, le CFA 1 correspond à quatre régions chromosomiques humaines appartenant aux HSA 18, HSA 6, HSA 9 et HSA 19 (Derrien, et al., 2007).*

La comparaison des génomes du chien et de l'Homme, réalisée à partir de la carte de 3000 marqueurs (Guyon, *et al.*, 2003a), a permis de définir les régions chromosomiques humaines homologues des 40 chromosomes canins. Par ailleurs, au laboratoire, Thomas Derrien a développé au cours de sa thèse le serveur web « AutoGRAPH » qui permet de formaliser et de visualiser l'organisation comparée de plusieurs génomes (Derrien, *et al.*,

2007). Ce serveur est un outil de construction de cartes comparées entre plusieurs espèces qui identifie les blocs de conservation de synténie entre génomes, évalue la colinéarité des gènes au sein de ces blocs conservés et identifie les zones de rupture de synténie. Ainsi, à partir d'un génome de référence comme celui du chien par exemple, il est possible de naviguer entre plusieurs génomes, de connaître pour un gène canin donné son orthologue humain ou murin, le bloc de synténie auquel il appartient et de déterminer l'ensemble des gènes adjacents (figure 8).

III.2.3.3. L'annotation du génome canin

L'annotation en gènes canins a été, principalement, produite à partir du système standard développé par Ensembl (Flicek, *et al*., 2011). Le nombre de gènes codant pour des protéines chez le chien est inférieur à celui de l'Homme et de la souris (19.305 *vs* 21.823, et 22.667 respectivement dans la version v.62 d'Ensembl). De plus, le nombre d'ESTs canins (~380.000) recensés dans la base de données DB_EST est très inférieur au nombre disponible pour l'Homme (8.000.000) et la souris (5.000.000) et peut expliquer, pour partie, la moins bonne efficacité de la première phase du programme Ensembl basée sur un alignement de séquences issues de données expérimentales propres à l'organisme (ESTs, protéines, ADNc...). À partir de cette observation, une approche de génomique comparative multi-espèces a été entreprise au laboratoire afin d'analyser des gènes annotés chez l'Homme, le chimpanzé, la souris et le rat mais absents chez le chien. L'approche est basée sur la construction de cartes de synténie à haute résolution (Derrien, *et al*., 2009; Derrien, *et al*., 2011). Enfin, les résultats des recherches de déterminant génétiques de traits ou de maladies permettent l'annotation fonctionnelle des génomes, par la mise en évidence de gènes impliqués dans des processus physiologiques (Fondon and Garner, 2004; Sutter, *et al*., 2007) ou pathologiques tel que la narcolepsie (Mignot, *et al*., 1994; Lin L., *et al*., 1999), l'atrophie progressive de la rétine (Guyon, *et al*., 2007)ou une forme de lipofuscinose (Abitbol, *et al*., 2010).

III.2.3.4. Le polymorphisme nucléotidique

En parallèle du séquençage intégral du génome du boxer, un programme de séquençage de 100.000 lectures a été réalisé sur l'ADN de neuf races de chiens (beagle, bedlington terrier, berger allemand, chien d'eau portugais, labrador retriever, lévrier italien, malamute, rottweiller et setter anglais), ainsi que quatre loups et un coyote. L'ensemble de ces données de séquences a constitué une source essentielle pour déterminer des variants de type SNP (Single Nucleotide Polymorphism). En effet, à l'issue du premier séquençage léger, 974.000

nucléotides de la séquence du génome du caniche présentaient un polymorphisme allélique. À ceux-ci s'ajoutent les 768.948 SNP identifiés sur la séquence du boxer séquencé. Enfin, la comparaison des 100.000 séquences aléatoires obtenues pour les neuf autres races de chiens, les quatre loups et le coyote ont permis de définir près de 373.000 SNP. Ainsi, au total, à partir de l'analyse comparée de toutes ces séquences, plus de deux millions de SNP ont été identifiés (Lindblad-Toh, *et al*., 2005).

Cette connaissance à permis la conception et la commercialisation des puces Affymetrix comportant 50.000 SNP et Illumina comportant 27.000 SNP dédiées au génome canin qui ont été utilisées dans différentes études d'association pour découvrir des causes génétiques de traits fixés dans certaines races ou de certaines maladies (Awano, *et al*., 2009; Cadieu, *et al*., 2009; Akey J. M., *et al*., 2010). Enfin le consortium européen de génétique canine LUPA, auquel l'équipe génétique du chien prend part dans le cadre du 7e PCRD (Plan Cadre de Recherche et Développement), a permis la conception d'une puce haute densité Illumina de 170.000 SNP canins qui a été utilisée pour génotyper plusieurs centaines de chiens répartis en différentes races. C'est la connaissance de ces génotypes qui permet d'étudier les populations canines du point de vue génétique afin de caractériser les différentes traces laissées par la sélection artificielle après la domestication.

III.3. Le chien, modèle d'étude des maladies génétiques

III.3.1. Les maladies génétiques

Outre la découverte des bases génétiques de l'extrême diversité morphologique et comportementale présente entre les races de chiens, le chien présente un intérêt en tant que modèle d'étude des déterminants de maladies génétiques qui ségrègent au sein d'une race donnée. La création des races a concentré, de façon involontaire, des allèles morbides ou des combinaisons non adéquates d'allèles prédisposant à des maladies génétiques (Ostrander and Giniger, 1997; Galibert, *et al*., 2004; Sutter and Ostrander, 2004). L'utilisation de mâles qualifiés d'étalons champions a aussi contribué à la fixation d'allèles de prédisposition et à l'augmentation de la consanguinité. Aujourd'hui, la quasi-totalité des races canines est atteinte de diverses maladies génétiques naturelles et spontanées. S'intéresser aux causes génétiques des maladies canines est important à la fois en médecine vétérinaire et en tant que modèles pour les maladies humaines. L'intérêt en médecine vétérinaire est de diminuer la prévalence de nombreuses maladies génétiques par le développement de tests de dépistage et par l'amélioration des thérapies afin de contribuer ainsi à améliorer la santé des près de 8 millions de chiens de compagnie recensés en France (*FACCO/SOFRES 2008 http://*

www.facco.fr/-Population-animale). Le plus grand intérêt des études des causes génétiques des traits et maladies canines est l'utilisation du chien en tant que modèle pour l'étude des maladies humaines. Chez l'Homme, le mode de transmission de nombreuses maladies génétiques est complexe, rendant difficile les analyses génétiques. Pour faciliter l'étude de ces maladies, la souris est souvent utilisée en laboratoire en induisant des mutations qui vont mimer les phénotypes, les maladies étudiées. Le chien est un modèle génétique d'intérêt pour l'étude des maladies homologues aux maladies humaines de par l'environnent dans lequel les chiens vivent, leur physiologie, la survenue des maladies de manière spontanée et les possibilités d'études familiales et populationnelles liées à la disponibilité de larges fratries et de vaste cohorte. Le chien partage depuis toujours la vie et donc le même environnement que l'Homme et est exposé aux mêmes facteurs environnementaux. De ce fait, il représente un modèle très adapté pour la plupart des études de recherche des causes génétiques des cancers ou des maladies multifactorielles impliquant le milieu environnemental, par rapport à la souris par exemple. De plus, compte tenu de sa taille et de son métabolisme plus proche de l'Homme que celui de la souris, le chien est déjà utilisé en routine pour des tests pharmacologiques (Starkey, *et al*., 2005). Il présente une clinique et des réponses aux médicaments semblables à celles de l'Homme. Enfin un chien vieillit 5 à 8 fois plus vite qu'un humain et est soigné tout au long de sa vie; ainsi le suivi médical de l'espèce canine est le mieux documenté après celui de l'Homme. De plus, les études génétiques sont menées sur des chiens de races pures inscrits au LOF, ce qui permet un accès aux informations généalogiques pour la construction de pedigrees. Il est relativement aisé d'avoir accès à de larges familles dans lesquelles les traits étudiés ségrègent avec de fortes incidences. Il est donc possible de construire des pedigrees particuliers pour réaliser des analyses de liaisons génétiques avec des familles nombreuses (une portée de chien comporte souvent 8 à 9 petits) sur plusieurs générations. C'est ainsi que des colonies de dobermans atteints de narcolepsie ont permis l'identification de deux mutations différentes dans un même gène (Lin L., *et al*., 1999), le gène du récepteur de l'hypocrétine (HCTR). Ces résultats ont identifié que le gène codant le ligand de ce même récepteur était muté dans certains cas familiaux de narcolepsie chez l'Homme (Peyron, *et al*., 2000).

Comme nous l'avons décrit, chaque race représente un isolat génétique et l'utilisation de mâles reproducteurs favorisent les "effets fondateurs" qui impliquent qu'une mutation délétère unique peut être transmise à un grand nombre d'individus de la race. De ce fait, plus de la moitié des maladies sont spécifiques de races et ségrègent dans une ou un petit nombre de races (Switonski, *et al*., 2004). De plus, la consanguinité tend à augmenter le nombre de

locus homozygotes et favorise ainsi l'apparition de maladies héréditaires récessives et contrairement à ce qui est observé chez l'Homme, les maladies se transmettent préférentiellement selon un mode autosomique récessif chez le chien (Galibert, *et al*., 2004). Cette transmission naturelle des maladies est importante car l'atout majeur offert par le modèle canin pour la recherche de causes génétiques de différents phénotypes réside dans le caractère spontané de ces maladies, aucune manipulation génétique n'est nécessaire, et dans la limitation du phénomène de phénocopie. En effet, lorsque deux populations ou individus humains présentent les mêmes symptômes, les gènes responsables peuvent être différents. C'est pourquoi les généticiens ont recours à des populations humaines isolées présentant un effet fondateur.

L'éventail des anomalies héréditaires canines est large et aucun métabolisme n'est épargné. À ce jour, 552 anomalies héréditaires ségrègent au sein des races canines dont 261 sont considérés comme des modèles potentiels pour des maladies humaines (OMIA : http://omia.angis.org.au/) tel que les atrophies progressives rétiniennes, équivalentes des rétinites pigmentaires humaines (Lin C. T., *et al*., 2002). D'autre part, les maladies génétiques ont des prévalences bien supérieures à celles observées chez l'Homme. Chez ce dernier, une affection avec une fréquence de 1/500 est considérée comme très fréquente, tandis que chez le chien, beaucoup de désordres héréditaires ont une prévalence de 1 à 25% dans certaines races comme le sarcome histiocytaire du bouvier bernois (Giger, *et al*., 2006; Abadie, *et al*., 2009). Nous disposons maintenant de nombreux exemples confirmant que souvent les mêmes gènes sont à l'origine des mêmes maladies chez l'Homme et le chien (Ostrander, *et al*., 2000; Patterson D. F., 2000; Switonski, *et al*., 2004; Tsai, *et al*., 2007; Grall, *et al*., 2011). Par exemple, des mutations dans le gène RPE65 sont responsables d'une même forme d'atrophie rétinienne conduisant à la cécité dans les deux espèces : l'amaurose congénitale de Leber chez l'Homme (Cremers, *et al*., 2002) et le CSNB (congenital stationary night blindness) chez le chien (Lin C. T., *et al*., 2002). Par l'existence de modèles canins spontanés présentant des défauts génétiques similaires à ceux de l'Homme, le chien se révèle être un modèle très prometteur pour engager des essais de thérapie génique et de thérapie cellulaire (Herzog, *et al*., 1999; De Meyer, *et al*., 2006; Sampaolesi, *et al*., 2006; Le Meur, *et al*., 2007).

III.3.2. La recherche des bases génétiques des traits et maladies

Rechercher les causes génétiques d'un trait héréditaire consiste la plupart du temps à rechercher un ou plusieurs gènes à l'origine de ce trait. La réalisation de ces études implique

la collecte d'un grand nombre d'échantillons d'ADN d'individus présentant le trait d'intérêt et d'individus ne présentant pas ce trait. Le trait étudié doit donc être suffisamment répandu et bien décrit (cliniquement, histologiquement) afin de pouvoir collecter un nombre suffisant d'individus des deux catégories. A partir de ces échantillons d'intérêt, il est possible de rechercher un gène impliqué dans le trait d'intérêt. Dans ce cadre, l'approche gène candidat consiste à rechercher des différences dans la séquence d'un gène sélectionné dans le génome entre les individus qui présentent le trait étudié et les individus qui ne le présentent pas. Ce gène peut-être sélectionné par différentes méthodes : (i) l'étude de la physiologie d'une maladie peut aider à sélectionner un gène candidat impliqué dans les voies métaboliques concernées, (ii) le gène testé peut être homologue à un gène responsable du trait étudié dans une autre espèce, (iii) enfin le gène d'intérêt peut être situé dans un locus génomique identifié par les résultats d'une étude de cartographie génétique. La cartographie génétique consiste à exploiter le fait que lors de la méiose, des allèles de marqueurs de deux locus situés sur un même chromosome à proximité l'un de l'autre sont plus fréquemment transmis ensemble à la génération suivante que deux marqueurs éloignés ou sur deux chromosomes différents. Ceci permet d'utiliser des marqueurs polymorphes, dont la position sur le génome est connue, comme balises pour signaler la présence d'un facteur génétique potentiellement impliqué dans le trait étudié. Lorsque les individus dont on possède l'ADN font partie d'une même famille, on procède à une étude de liaison génétique ; si les individus dont on possède l'ADN ne sont pas apparentés, c'est alors une étude d'association qui est menée.

III.3.2.1. Analyse de liaisons génétiques

Avec la disponibilité de données familiales, la cartographie génétique consiste à suivre la ségrégation des marqueurs disposés le long du génome et la ségrégation de la maladie au sein du pedigree. L'objectif est de déterminer un locus dans lequel les marqueurs ségrègent avec le phénotype d'intérêt. L'analyse de liaisons génétiques se base sur le taux de recombinaison lors de la méiose entre le phénotype et chacun des marqueurs. Le taux de recombinaison est compris entre 0 (aucune recombinaison) et 0,5 (50% de recombinaison). Deux marqueurs situés sur des chromosomes différents sont transmis indépendamment au cours de la méiose. Leur taux de recombinaison est alors de 50%. Deux locus situés sur un même chromosome et dont le taux de recombinaison est inférieur à 50% sont dits liés. Les gènes candidats sont sélectionnés parmi les gènes dont la position physique sur le génome est entre les bornes du locus délimité par les marqueurs liés avec le phénotype d'intérêt. Le modèle canin se prête bien aux analyses de liaisons car il est possible de collecter, avec l'aide des éleveurs, des vétérinaires et des propriétaires, des prélèvements de chiens issus de

plusieurs générations familles dans lesquelles ségrègent une ou des maladies spontanées d'intérêt. Cette méthode a permis de localiser les gènes impliqués dans un certain nombre de maladies monogéniques autosomiques récessives comme l'épilepsie qui affecte le teckel miniature à poil dur (Lohi, *et al*., 2005) ou du collapsus induit par l'exercice physique qui affecte le labrador retriever (Patterson E. E., *et al*., 2008). Malgré les avantages des analyses de liaison dans l'espèce canine, notamment dans le cas de certaines races à petits effectifs qui contiennent principalement des individus apparentés, cette méthode présente des limites. Elle nécessite des familles complètes dans lesquelles les relations de parenté entre individus sont bien connues, si la déclaration de la maladie est tardive, les parents d'un sujet atteint sont le plus souvent déjà décédés et il peut être difficile de recruter plusieurs générations d'une même famille. De plus les locus découverts sont de grande taille et les traits les plus facilement mis en évidence sont des traits récessifs à forte pénétrance. Enfin les projets collaboratifs facilitent le recrutement d'un grand nombre d'individus non apparentés.

III.3.2.2. Études d'associations

Une étude d'association est réalisée sur une cohorte d'individus non-apparentés appartenant à une même population parmi lesquels certains présentent le trait intérêt, ce sont les individus cas, et les autres qui ne le présentent pas, sont les individus contrôles. L'analyse d'association se base sur la corrélation entre la présence d'un allèle d'un marqueur donné chez un individu et le statut cas ou contrôle de cet individu. Lorsqu'un marqueur est proche du variant impliqué dans le phénotype, c'est l'allèle de ce marqueur qui est sur le même segment chromosomique que le variant qui sera transmis avec et se retrouvera enrichi dans la population des individus cas. Les études d'associations sont plus sensibles pour détecter les allèles à faible pénétrance des maladies complexes. Les gènes candidats sont sélectionnés parmi les gènes dont la position physique sur le génome est située entre les bornes du locus délimité par les marqueurs pour lesquels les fréquences alléliques sont significativement différentes entre les cas et les contrôles.

Le modèle canin se prête bien aux analyses d'association car les déséquilibres de liaison observés au sein des races canines sont 50 fois supérieur à ceux observés en génétique humaine (Sutter, *et al*., 2004; Lindblad-Toh, *et al*., 2005; Parker, *et al*., 2006). Ainsi, en théorie, le nombre de marqueurs suffisant pour couvrir l'ensemble du génome canin lors d'études d'associations est beaucoup plus faible que celui nécessaire pour les études menées chez l'Homme. De plus, il a été montré que si le caractère recherché est de transmission autosomique récessif, un faible nombre de chiens, seulement 10 cas et 10 contrôles, suffit pour identifier le locus génomique impliqué (Karlsson, *et al*., 2007; Merveille, *et al*., 2011).

Un autre avantage majeur des analyses d'associations est l'exploitation du fait que chaque race présente un certain nombre de traits fixés. Il est ainsi possible, sous l'hypothèse d'une mutation commune, de rassembler les individus de plusieurs races possédant ce trait dans le groupe des cas et des individus de plusieurs races ne présentant jamais ce trait dans le groupe des contrôles. Les analyses d'associations ont permis de localiser les gènes impliqués dans les maladies qui ségrègent au sein d'une race donnée. Le groupe de Awano a identifié le gène responsable de la myélopathie dégénérative, une maladie de la moelle épinière, dans la race pembroke welsh corgi (Awano, *et al*., 2009). Notre équipe a identifié la mutation du gène responsable de l'ichtyose -une maladie dermatologique- qui affecte jusqu'a 30% des golden retrievers (Grall, *et al*., 2011), ainsi que des gènes responsables de traits fixés comme la texture et la longueur du pelage (Cadieu, *et al*., 2009) ou la chondrodysplasie -une maladie qui provoque la déformation et le raccourcissement des os notamment du cubitus observé chez le teckel par exemple (Parker, *et al*., 2009).

Les limites des études d'associations sont liées à la structure des causes génétiques du trait étudié et à la structure de la population dans laquelle se fait l'étude. Dans le cas d'une étude réalisée au sein d'une race, si un gène présente plusieurs variants différents causant un même trait qui de surcroît seront associés à différents allèles des marqueurs ou si le trait étudié est polygénique, l'analyse sera complexe et la puissance statistique restreinte ; Cette situation est un cas moins fréquent chez le chien que chez l'Homme car la structure de la population canine implique le plus souvent un fort effet fondateur. La présence de stratification génétique dans les populations étudiées peut baisser la puissance des études d'associations ou induire des associations statistiques erronées, c'est le cas, par exemple lorsque les cas sont plus apparentés que les contrôles. Au sein d'une race de chien qui présente une faible diversité, la stratification ne posera à priori pas de problème, alors que pour une race présentant une plus forte diversité telle que le golden retriever, le processus expérimental doit respecter un échantillonnage ayant la même proportion de cas dans chaque sous-population (Quignon, *et al*., 2007).

III.3.2.3. Les approches mixtes

D'autres méthodes ont été mises au point pour tenter de tirer parti des avantages des deux méthodes. Par exemple l'approche "d'homozygosity mapping" consiste à rechercher des régions systématiquement homozygotes chez des individus atteints d'une maladie récessive et issus de parents apparentés dans différentes familles (Lander and Botstein, 1987) ; Le test de déséquilibre de transmission consiste à tester si des individus atteints dans

différentes familles nucléaires ont reçu de leurs parents un allèle particulier plus fréquemment que ne le voudrait le hasard (Spielman, *et al*., 1993).

Quelle que soit l'approche utilisée, ces études portent sur des traits identifiés, dont nous possédons la connaissance à priori pour séparer les individus présentant le phénotype d'intérêt et les individus qui ne le présentent pas. Dans la réalité, chaque race possède une combinaison de traits fixés qui lui est propre, dont une partie ne s'exprime pas par un phénotype 'visible', les traits physiologiques ou métaboliques en sont un exemple. La possibilité d'obtenir des génotypes d'individus de plusieurs races permet de rechercher et d'identifier le catalogue des locus qui présentent des fortes différenciations génétiques entre les races. Pour identifier les locus qui différencient une (ou quelque) race de toutes les autres de la population canine, cette approche permet de dresser le répertoire des locus et des gènes qui présenteraient des variants spécifiques de race correspondants à des traits sans avoir la connaissance à priori du phénotype.

Objectifs du travail de thèse

La structure de la population canine, l'existence d'une puce SNP haute résolution et la disponibilité de la séquence et de son annotation font du chien un très bon modèle pour rechercher les signatures génétiques de la sélection. Au cours de mon projet de thèse, j'ai recherché les signatures de la sélection correspondant aux deux principales périodes de temps de l'évolution de l'espèce canine. Dans la première partie de mon travail de thèse, nous avons analysé les gènes codants afin de détecter la sélection positive chez le chien et neuf autres mammifères. Cette étude s'est réalisé à partir de l'utilisation de données de séquences des gènes avec pour principaux objectifs d'établir le catalogue complet des gènes canins sous sélection positive dans le contexte phylogénétique de 10 mammifères qui disposent d'une séquence bien annotée de leur génome, de déterminer si l'impact de la sélection positive sur le génome canin est similaire à l'impact de la sélection positive sur le génome d'autres mammifères euthériens non-domestiqués, et de déterminer si ces gènes appartiennent préférentiellement à des réseaux métaboliques.

L'objectif du second volet de ma thèse est de détecter les régions de forte différenciation allélique entre races canines qui vont constituer les locus candidats de la sélection artificielle qui ont été influencés par la création des différentes races canines actuelles. La stratégie mise en oeuvre a constitué à identifier les locus génétiquement les plus différenciés à partir de données de génotypes de SNP sur l'ensemble du génome de 30 races comportant au moins 10 individus. Ce second volet se déroule dans le cadre du consortium européen de génétique du chien LUPA (www.eurolupa.org) qui regroupe 22 instituts européens dont le but est de découvrir les bases génétiques de maladies héréditaires chez le chien (Lequarré, *et al*., 2011). Cette étude s'est déroulé en étroite collaboration avec le groupe du Dr. M. Webster (Université d'Uppsala, Suède). Pour réaliser cette étude, une méthodologie basée sur l'indice Fst a tout d'abord été automatisée par la confection d'un pipeline bioinformatique. Les locus identifiés ont été analysés par une approche statistique de recherche "d'outliers", puis filtrés pour la recherche de faux-positifs (p valeur marginale et FDR). L'établissement du catalogue des régions potentiellement ciblées par la sélection artificielle dans différentes races de chiens, a pour perspectives de faciliter l'identification des gènes impliqués dans la différenciation en races et permet d'inférer leurs rôles et d'éventuelles nouvelles fonctions ou l'émergence de fonctions spécifiques. De plus, ces régions constituent des locus candidats pour comprendre les bases génétiques de maladies génétiques associées à la sélection de certains caractères dans les races de chien. Nous présentons dans la partie résultat les locus de différenciation allélique entre races canines.

L'identification des locus du génome canin qui sont la cible à la fois de la sélection naturelle et de la sélection artificielle (figure 9) constitue un dernier objectif et une perspective du projet de thèse. Nous présentons dans la partie discussion les résultats préliminaires de cette analyse et notre interprétation de la co-détection ou à l'absence de co-détection des locus impliqués dans les événements de sélection naturelle et artificielle.

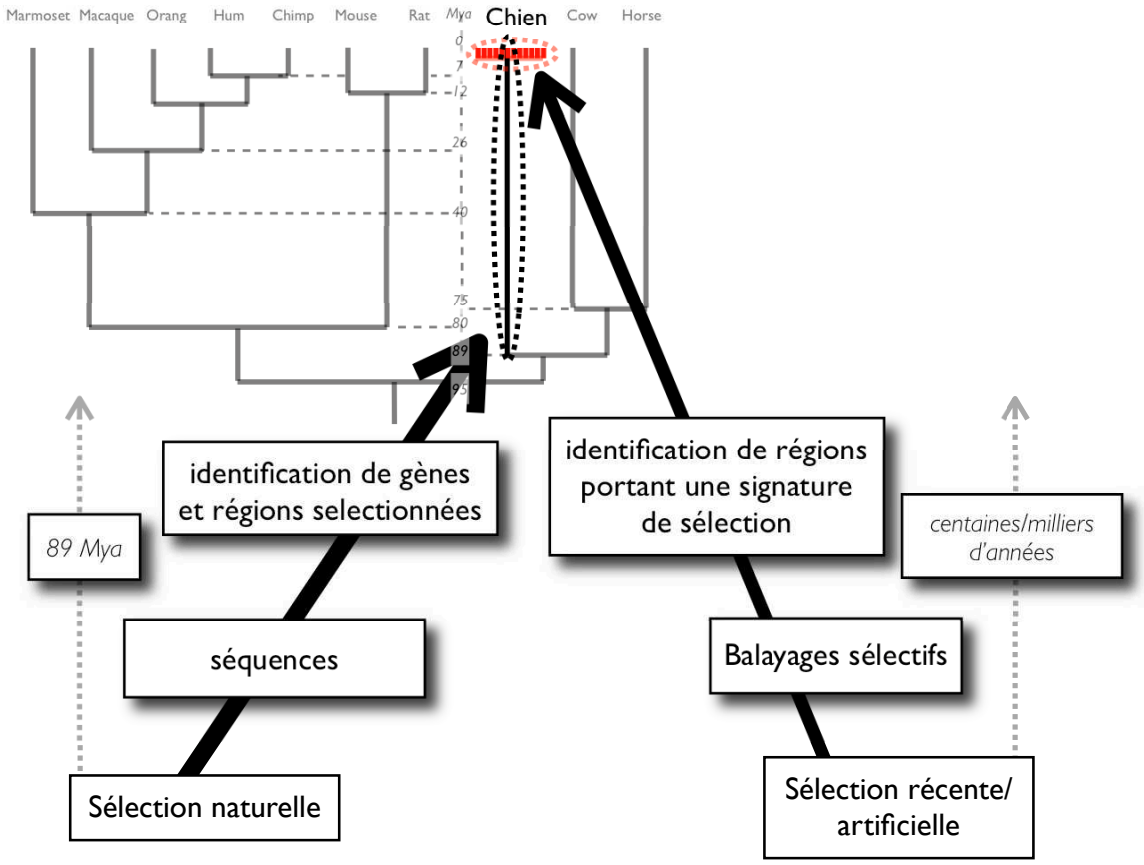


*Figure 9: **L'objectif du projet** est de caractériser les sélections naturelles et artificielles qui ont agi ou agissent encore sur le génome canin et de les comparer pour déterminer si certaines cibles de ces sélections sont récurrentes.*

Méthodologie et Résultats

I. Sélection positive naturelle chez le chien

L'étude menée dans la première partie du projet de thèse est l'analyse de la sélection positive au sein de la lignée qui mène au chien et qui s'exerce sur 89 millions d'années. La sélection positive est le processus par lequel la sélection d'allèles avantageux au cours de l'évolution va contribuer à la modification de la fonction d'une protéine et conférer une adaptation fonctionnelle. Elle se traduit par la sélection, puis la fixation de mutations dites non-synonymes plus fréquente que par le processus de dérive génétique.

Les gènes en relation d'orthologie de type 1:1 (un seul gène orthologue identifié dans chaque espèce) sont fortement conservés au cours de l'évolution et c'est parmi ces gènes que la détection de modifications adaptatives est la plus aisée, en raison de la capacité à comparer leurs séquences codantes. Il est alors attendu que tout changement parmi ces gènes conservés a un fort impact et participe aux phénomènes d'adaptation et de spéciation à l'environnement au cours de l'évolution. L'identification des cibles de la sélection positive au sein d'une espèce par approche dN/dS a pour objectif d'identifier les gènes et les réseaux de gènes impliqués au cours de l'évolution de l'espèce considérée.

I.1. Le contexte phylogénétique

Pour établir le catalogue complet des gènes sous sélection positive dans le génome canin à l'aide de gènes orthologues, il faut disposer des séquences des gènes de plusieurs espèces qui sont phylogénétiquement proches.

I.1.1. Le chien comparé à neuf autres espèces

Nous avons utilisé les génomes du chien et de neuf autres mammifères pour lesquels nous disposons du séquençage complet et d'une annotation à priori complète du génome. Les génomes utilisés sont (i) quatre génomes de primates qui ont été séquencés en raison notamment de leur proximité phylogénétique avec l'Homme, ce sont les espèces ouistiti, macaque, orang-outan et chimpanzé; (ii) deux génomes de rongeurs, souris et rat; (iii) et deux espèces plus proches du chien, le cheval et la vache.

I.1.2. L'annotation 'Ensembl'

Les génomes de ces espèces sont annotés et disponibles sur le site de la base Ensembl (http://www.ensembl.org) (Flicek, *et al*., 2011) qui utilise des méthodes automatiques pour annoter les génomes c'est à dire établir le catalogue des gènes présents dans le génome

étudié. L'annotation du serveur Ensembl se base sur la connaissance de séquence protéique de l'espèce considérée et des espèces proches phylogénétiquement ainsi que des séquences d'ADN complémentaire pour l'espèce d'intéret (Potter, *et al*., 2004). La version 'Ensembl Genes 63' contient 19.305 gènes canins codant pour des protéines, plus de 20.000 gènes en moyenne pour les autres génomes de notre étude (21.494 humain, 20.993 ouistiti, 21.905 macaque, 20.068 orang outan, 19.829 chimpanzé, 22.667 souris, 22.938 rat, 20.436 cheval et 21.048 vache). Le serveur Ensembl contient aussi les informations d'orthologie qui permettent de déterminer des listes de gènes orthologues ainsi que leur type de relation d'orthologie. À partir de leurs identifiants, les gènes orthologues peuvent être utilisés pour extraire leurs séquences codantes.

I.1.3. Extraction des 10.000 orthologues

L'extraction et l'alignement des séquences orthologues est une étape déterminante dans la recherche de gènes sous sélection positive. Pour obtenir les orthologies et alignements les plus fiables possibles, nous avons établi une collaboration avec le Dr. Hugues Roest Crollius (équipe DYOGEN ENS Paris) qui a réalisé 10.730 alignements protéiques issus de la redéfinition de 10.730 orthologues de type 1:1 entre les 10 espèces mammifères considérées et l'alignement des codons de ces séquences en cumulant un alignement nucléotidique de chaque exon et un alignement protéique. La longueur des alignements de séquences entre les 10 espèces varie de 150 à 15.150 nucléotides (moyenne 1353 ; écart type 1254).

I.1.4. Le Likelihood-Ratio Test -LRT-

Nous avons utilisé les alignements de séquences des 10.730 gènes orthologues pour calculer le test de sélection positive par branche et par site proposé par Yang et Neilsen (Yang Z and Nielsen, 1998; Zhang Jianzhi, *et al*., 2005b; Yang Z and Dos Reis, 2011). Ce test détecte la sélection positive dans la séquence codante des gènes. Cette approche est basée sur l'évaluation des taux de mutations non-synonymes (dN) et les taux de mutations neutres (dS). Ces taux permettent de calculer le ratio dN/dS (appelé aussi ω). Les gènes orthologues étant conservés au cours de l'évolution, la valeur ω d'un gène sur l'ensemble de sa séquence est très faible. La valeur ω est donc évaluée par site afin d'être mesurée de manière très résolutive. Pour réaliser cette analyse, nous utilisons le programme codeML du package PAML (Phylogenetic Analysis by Maximum Likekihood) (Yang Z., 1997) pour calculer les proportions de sites sous sélection positive, absente ou négative. Ces calculs nécessitent de préciser une branche phylogénétique d'intérêt au logiciel CodeML. Le programme CodeML

calcule les proportions de sites qui sont sous sélection positive, absente ou négative pour l'espèce d'intérêt selon un modèle qui implique que ces mêmes sites soient sous évolution neutre ou sélection négative dans les autres branches de l'arbre. Le programme détermine alors la vraisemblance de ce modèle. Nous effectuons le calcul de ces proportions et leurs vraisemblances pour chacun des 10.730 gènes de chacune des 10 espèces.

En parallèle, nous avons calculé les mêmes proportions et vraisemblances pour un second modèle dans lequel la valeur ω qui était calculée pour la sélection positive est fixé à 1 pour représenter une hypothèse nulle d'évolution neutre. Les hypothèses nulle d'absence de sélection positive (H_0) et alternative de présence de sélection positive (H_1) sont incluses l'une dans l'autre. Par conséquent nous pouvons comparer les logarithmes des vraisemblances de ces deux modèles ($lnLH_0$ et $lnLH_1$). En effet, en absence de sélection positive, le résultat du calcul $2*(lnLH_1-lnLH_0)$ suit une distribution du chi-deux (χ^2) à un degré de liberté. Ce test devient plus fiable à mesure que l'on augmente le nombre d'espèces considérées pour les gènes présentant des relations d'orthologie 1:1. Enfin les p valeurs du test de χ^2 sont corrigées par la méthode de Benjamini-Hochberg -BH- (Benjamini and Hochberg, 1995) qui permet de contrôler le taux de faux positifs inhérents à une série de tests multiples. Le seuil de la p valeur de 0,05 est utilisé pour identifier les gènes candidats sous sélection positive dans chaque espèce. C'est dans ce contexte que nous avons détecté 633 gènes canins (138 gènes après correction BH) sous sélection positive dans la branche phylogénétique menant au chien, à comparer aux valeurs de la branche du ouistiti (n=855), du macaque (n=367), de l'orang-outan (n=459), de l'Homme (n=169), du chimpanzé (n=360), de la souris (n= 673), du rat (n=790), de la vache (n=711) et du cheval (n=677), comme illustré dans la figure 10.

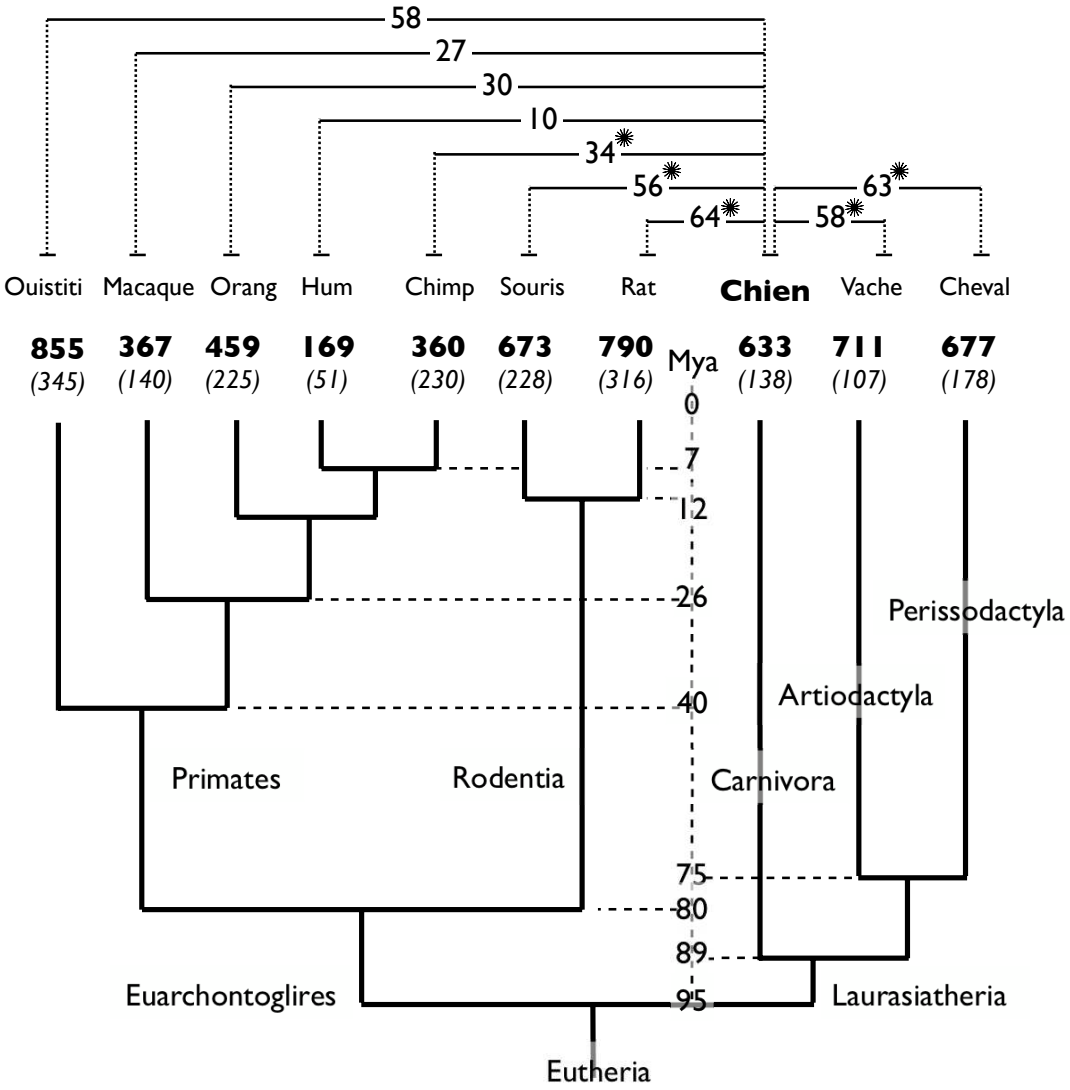


*Figure 10 : **Arbre des 10 espèces utilisées**. Cette figure représente l'arbre phylogénétique du chien et des 9 autres espèces utilisées pour établir le contexte phylogénétique nécessaire au calcul des vraisemblances des modèles. Les chiffres en gras au dessus du nom des espèces représentent le nombre de gènes détectés sous sélection positive par le test LRT parmi les 10.730 testés. Les chiffres entre parenthèses représentent le nombre de gènes toujours détectés après avoir corrigé les p valeurs par la méthode de Benjamini-Hochberg (Benjamini and Hochberg, 1995). Les chiffres indiqués en dessus du nom des espèces représentent le nombre de gènes sous sélection positive conjointement dans chaque espèce et l'espèce canine. Les étoiles indiquent les co-occurrences significatives (p<0.01).*

I.2. Les gènes canins sous sélection positive dans les autres espèces.

Le catalogue des gènes sous sélection positive dans le génome canin et celui des autres espèces permet 1) de comparer l'impact fonctionnel de la sélection positive sur le génome canin sur le génome avec les autres mammifères euthériens ; 2) de déterminer si les gènes canins détectés sous sélection positive appartiennent préférentiellement à des réseaux

métaboliques, soit par la sélection d'une fonction générée par le réseau, soit par adaptation aux changements consécutifs à l'évolution d'un gène responsable de la fonction d'intérêt.

Le nombre de gènes sous sélection positive communs entre le chien et une autre espèce varie de 10 à 64 comme indiqué sur la figure 10 et récapitulé dans la table 1. La significativité de la co-occurrence observée par rapport à une co-occurrence aléatoire a été testée par paire. Nous considérons que chacun des effectifs de gènes sous sélection positive est un tirage aléatoire parmi les 10.730 gènes, la probabilité qu'un gène donné de la liste des gènes sous sélection positive canine soit sélectionné dans la liste des gènes sous sélection dans la lignée bovine, par exemple, est de 711/10730. Le nombre de gènes attendus en commun entre la liste canine et celles des autres espèces varie de 10 à 50 et est récapitulé dans la table 1.

	# gènes sous sélection positive	# gènes communs avec le chien	# attendu	Probabilité d'écart au moins aussi extrême
cheval	677	63	39,94	0,0002
vache	711	58	41,94	0,007
Homme	169	10	9,97	0,54
ouistiti	855	58	50,44	0,14
macaque	367	27	21,7	0,14
orang-outan	459	30	27,08	0,3
chimpanzé	360	34	21,24	0,004
souris	673	56	39,7	0,005
rat	790	64	46,6	0,005
Total	3927	282	NA	NA

*Table 1 : **Nombre de gènes sous sélection positive** en commun entre le chien et une autre espèce.*

Classiquement, pour tester si une valeur dévie significativement de l'attendu, on calcule la probabilité d'obtenir une valeur au moins aussi extrême que dans l'hypothèse où cette déviation soit due au hasard. Par exemple, la probabilité qu'exactement 58 gènes bovins de la liste des 711 gènes sous sélection positive soient aussi sous sélection positive dans la lignée menant au chien correspond au nombre de possibilités de sélectionner exactement 58 gènes parmi les 633 canins et 653 gènes parmi les 10.097 gènes restants, divisé par le nombre de possibilités de prendre 711 gènes parmi les 10.730 de départ. Nous pouvons ainsi calculer les probabilités exactes de tous les niveaux de co-occurrence et en déduire la probabilité d'un écart au moins aussi important. Ainsi la probabilité d'obtenir un écart à la co-occurence attendue au moins aussi important sera considéré significatif lorsqu'il est <0.01 (cf Table 1).

La lignée menant au chien partage un plus grand nombre nombre de gènes sous sélection positive avec les lignées des autres Laurasathéria et des rongeurs que l'attendu dans le cas de recoupements aléatoires. En revanche nous identifions dans les lignées primates, à l'exception de la lignée chimpanzé, des gènes sous sélection positive commun avec le chien qui correspondent à ce que l'on attend d'un recoupement obtenu par le hasard.

I.3. Développement d'un serveur d'analyse des contraintes sélectives des séquences codantes : OMEGA

Au cours de ce travail, nous avons développé un outil d'analyse, OMEGA, dédié à l'automatisation des calculs de dN/dS, et au calcul du test LRT à partir de séquences de gènes codant pour des protéines au format fasta non alignées. OMEGA a été conçu et mis en ligne pour permettre son utilisation *via* une interface web.

I.3.1. Principe du serveur OMEGA

Ce serveur web OMEGA (http://dogs.genouest.org/OMEGA) utilise en entrée des séries de jeux de gènes dont chaque séquence est copiée/collée dans un formulaire HTML. Un jeu de données est constitué des séquences nucléotidiques sans code d'ambiguïté d'un même gène pour les différentes espèces analysées. Les jeux de données sont traités selon la procédure décrite dans la figure 11 :

I.3.1.1. Alignement avec le programme T-Coffee

Les séquences de chaque jeu de données sont soumises à un premier traitement qui va les traduire en codon d'acides aminés puis les aligner entre eux à l'aide du méta-aligneur T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation). T-Coffee (Notredame, *et al*., 2000) est un outil d'alignement multiple qui fonctionne en combinant les algorithmes d'alignements global de ClustalW (Thompson, *et al*., 1994) et d'alignement local du package FASTA (Pearson and Lipman, 1988). À partir des alignements de chaque paire de séquences par les deux méthodes, T-Coffee génère une librairie indiquant le poids de chaque association possible entre deux nucléotides issus de séquences différentes. T-Coffee utilise ensuite une approche progressive pour générer l'alignement multiple : un premier alignement de deux séquences est complété par l'ajout des autres séquences pour obtenir l'alignement multiple. Si l'utilisateur a fourni un arbre phylogénétique des espèces correspondant aux séquences de son jeu de données, le serveur OMEGA fournit cet arbre à T-Coffee pour guider l'étape de génération progressive de l'alignement multiple. Enfin OMEGA rétro-traduit

chaque séquence de l'alignement pour obtenir un alignement nucléotidique basé sur les codons des séquences d'origine.

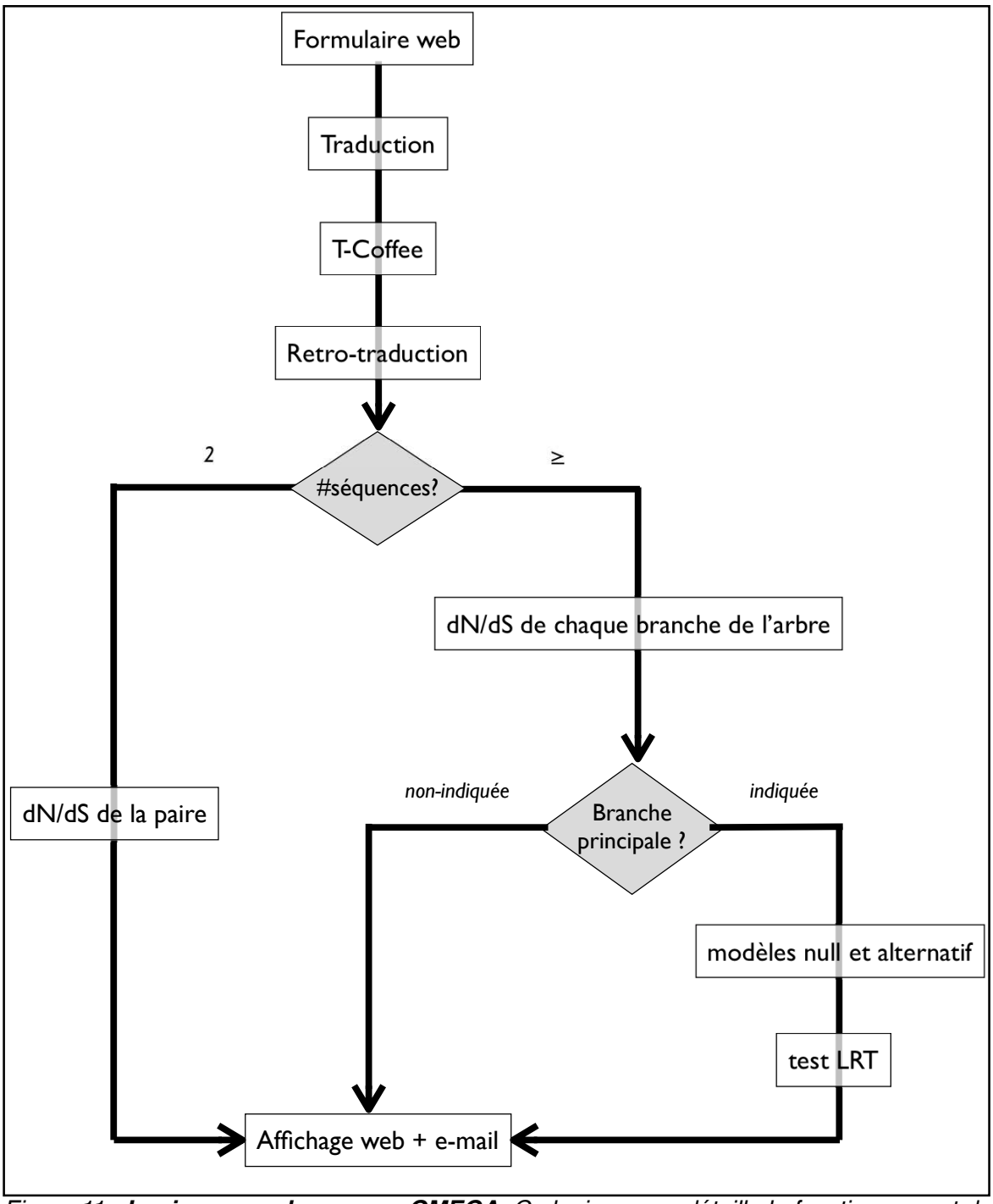


Figure 11 : ***Logigramme du serveur OMEGA.*** *Ce logigramme détaille le fonctionnement du serveur OMEGA qui pour chaque jeu de séquences entré simultanément par l'interface web réalise un alignement par codons puis calcule les ratios dN, dS et omega selon le nombre de séquences du jeu de données et de l'indication d'une branche d'intérêt. Les résultats sont affichés sur le navigateur et envoyés par mail.*

I.3.1.2. Calcul des ratios par paire et par branche

À partir des séquences alignées, le serveur calcule les ratios dN, dS et ω à l'aide de l'outil codeML du package PAML. Lorsque les jeux de données sont constitués des séquences de seulement deux espèces, le serveur retourne l'alignement et les valeurs des ratios calculées avec l'option "pairwise" de codeML (model = 0 NSsites = 0). La valeur ω indique alors l'accumulation de mutations synonymes comparée à l'accumulation de mutations non synonymes sans indiquer quelle séquence évolue plus vite que l'autre. Lorsque les jeux de données sont constitués d'un nombre de séquences supérieur à deux espèces, le serveur calcule les trois ratios pour chaque espèce et pour chaque branche interne de l'arbre (model = 1 NSsites = 0). La valeur ω indique alors un taux d'évolution de chaque branche de l'arbre fourni par l'utilisateur.

I.3.1.3. Calcul des modèles "Branch site"

Les calculs par branche de l'arbre ne donnent qu'exceptionnellement des ω supérieurs à 1. En effet, des gènes conservés entre espèces sont globalement sous sélection négative. Pour les gènes conservés entre espèces, la sélection positive peut cependant agir très ponctuellement sur certains sites constitués d'un seul codon/nucléotide par exemple, alors que les autres sites sont sous sélection négative. Pour évaluer la sélection positive au sein d'un gène d'une espèce, le serveur calcule le modèle par site dit 'branch site' (model = 2, NSsites = 2). Pour que ce modèle soit calculé, l'utilisateur doit indiquer la branche d'intérêt. Le serveur détermine les proportions de sites qui sont :

- 1. sous évolution neutre à la fois dans la branche d'intérêt et dans les autres branches
- 2. sous sélection négative à la fois dans la branche d'intérêt et dans les autres branches
- 3. sous sélection positive dans la branche d'intérêt et sous évolution neutre dans les autres branches
- 4. sous sélection positive dans la branche d'intérêt et sous sélection négative dans les autres branches

Ces proportions sont calculées à l'aide du programme codeML. En plus de ces proportions, les valeurs d'ω moyen pour les sites en sélection négative ou positive et le logarithme naturel de la vraisemblance de ce modèle évolutif sont calculées.

I.3.1.4. Likelihood-Ratio Test

Pour tester la présence de la sélection positive le serveur calcule le logarithme naturel de la vraisemblance pour un autre modèle dans lequel la valeur d'ω pour les sites potentiellement sous sélection positive est fixée à 1 (évolution neutre: model = 2, NSsites =

2, fix_omega = 1, omega = 1). C'est à partir des deux valeurs de vraisemblance que le serveur calcule le test LRT du modèle 'branch-site strict'. En effet le modèle H_0 d'absence de sélection positive est un cas particulier du modèle H_1 dont le seul paramètre libre supplémentaire est la possibilité pour certains sites d'avoir un $\omega \geq 1$. Le double de la différence entre le logarithme naturel de la vraisemblance du modèle H_1 et le logarithme naturel de la vraisemblance du modèle H_0 est alors une valeur de χ^2 à un degré de liberté. Le serveur calcule donc le test en utilisant l'outil de χ^2 présent dans le package PAML modifié pour être utilisé au sein d'un script.

I.3.2. Interface Web du serveur OMEGA

L'interface web du serveur Omega est disponible à l'adresse http://dogs.genouest.org/OMEGA. La page d'accueil permet de choisir entre la recherche parmi des calculs de test LRT pré-insérés (OMEGAbase : partie en cours de développement) et la possibilité de rentrer ses propres données pour calculer les ratios et le test pour un jeu de données d'intérêt (OMEGAtool).

Lorsque le module "OMEGAtool" est utilisé, un premier formulaire apparaît (figure 12) permettant à l'utilisateur d'indiquer le nombre d'espèces qu'il veut comparer et le nombre de gènes pour lesquels il veut faire cette comparaison. Ce formulaire permet la création d'un second formulaire (figure 13) permettant de rentrer dans un premier cadre son e-mail, les noms des espèces d'intérêts et l'arbre de ces espèces au format Newick avec la branche d'intérêt suivie de "#1". Le formulaire se poursuit par une série de cadres qui permettent de rentrer le nom de chaque gène d'intérêt et les séquences pour chaque espèce.

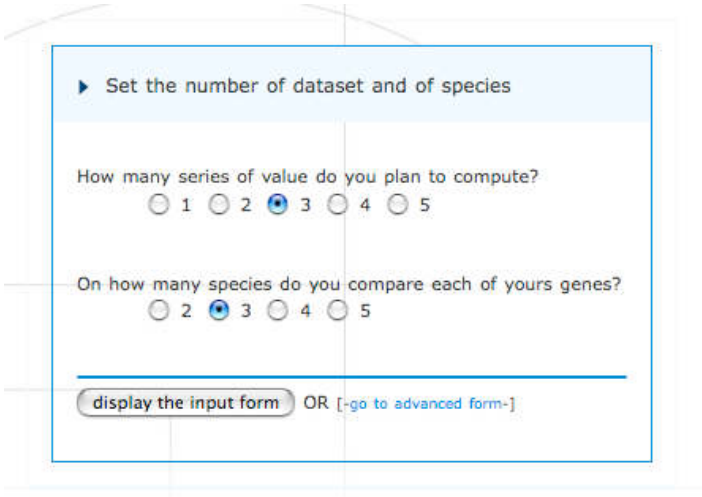


*Figure 12 : **Formulaire de préparation**. Formulaire permettant d'indiquer le nombre de jeux de données et le nombre de séquences par jeu*

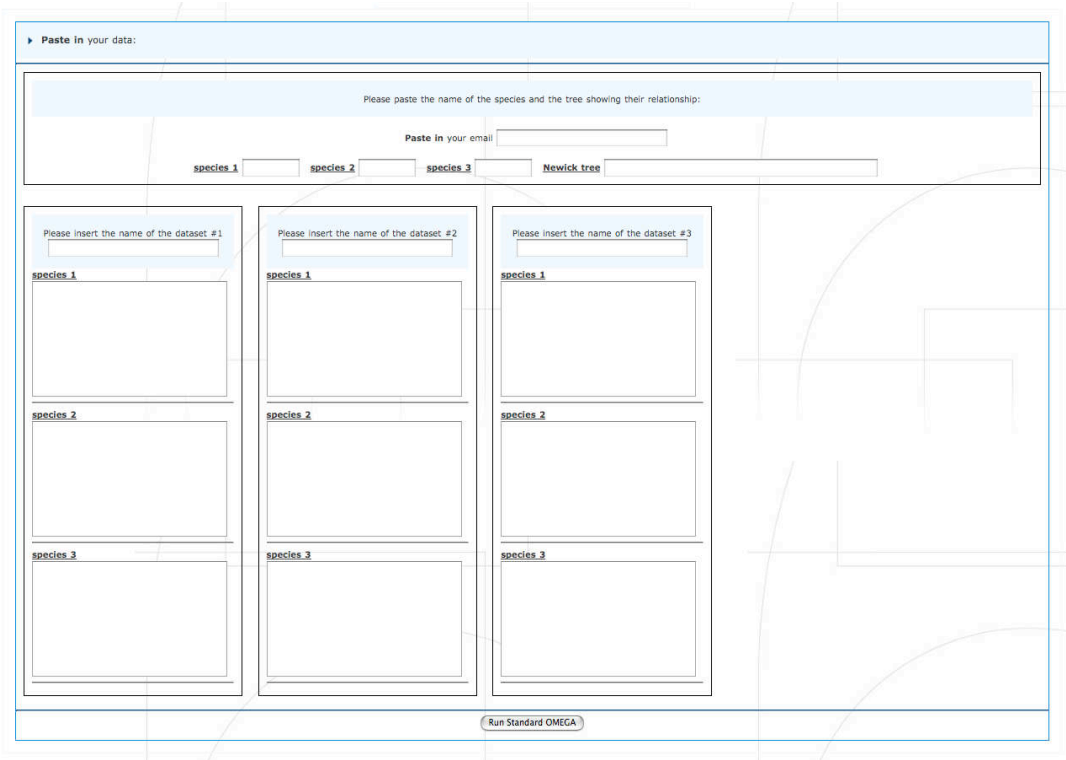
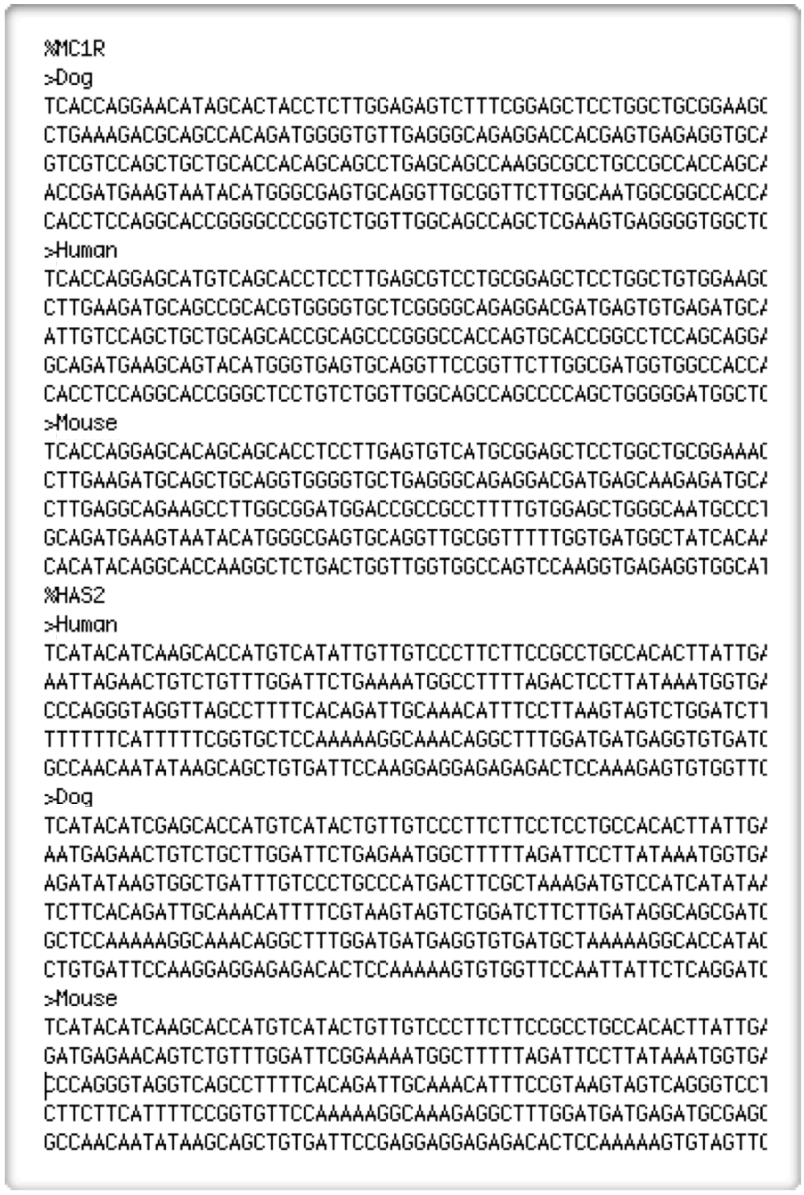


*Figure 13 : **Formulaire d'insertion**. Formulaire permettant de rentrer son e-mail, les noms des espèces d'intérêt, l'arbre de ces espèces au format Newick avec la branche d'intérêt suivie de "#1" et les différentes séquences.*

Une autre possibilité pour rentrer des données personnelles est d'utiliser le formulaire avancé en utilisant le lien "[-go to advanced form-]" qui permet d'insérer dans le cadre central les différentes séries de données. Dans cette entrée l'utilisateur doit insérer pour chaque gène une ligne commençant par le symbole "%" et contenant le nom du gène. Chaque ligne "%" est alors suivie des différentes séquences à comparer pour ce gène au format fasta; OMEGA considère que le nom de l'espèce de chaque séquence est la ligne de description de cette séquence (figure 14).



```
%MC1R
>Dog
TCACCAGGAACATAGCACTACCTCTTGGAGAGTCTTTCGGAGCTCCTGGCTGCGGAAGC
CTGAAAGACGCAGCCACAGATGGGGTGTTGAGGGCAGAGGACCACGAGTGAGAGGTGCA
GTCGTCCAGCTGCTGCACCACAGCAGCCTGAGCAGCCAAGGCGCCTGCCGCCACCAGCA
ACCGATGAAGTAATACATGGGCGAGTGCAGGTTGCGGTTCTTGGCAATGGCGGCCACCA
CACCTCCAGGCACCGGGGCCCGGTCTGGTTGGCAGCCAGCTCGAAGTGAGGGGTGGCTC
>Human
TCACCAGGAGCATGTCAGCACCTCCTTGAGCGTCCTGCGGAGCTCCTGGCTGTGGAAGC
CTTGAAGATGCAGCCGCACGTGGGGTGCTCGGGGCAGAGGACGATGAGTGTGAGATGCA
ATTGTCCAGCTGCTGCAGCACCGCAGCCCGGGCCACCAGTGCACCGGCCTCCAGCAGGA
GCAGATGAAGCAGTACATGGGTGAGTGCAGGTTCCGGTTCTTGGCGATGGTGGCCACCA
CACCTCCAGGCACCGGGCTCCTGTCTGGTTGGCAGCCAGCCCCAGCTGGGGGATGGCTC
>Mouse
TCACCAGGAGCACAGCAGCACCTCCTTGAGTGTCATGCGGAGCTCCTGGCTGCGGAAAC
CTTGAAGATGCAGCTGCAGGTGGGGTGCTGAGGGCAGAGGACGATGAGCAAGAGATGCA
CTTGAGGCAGAAGCCTTGGCGGATGGACCGCCGCCTTTTGTGGAGCTGGGCAATGCCCT
GCAGATGAAGTAATACATGGGCGAGTGCAGGTTGCGGTTTTTGGTGATGGCTATCACAA
CACATACAGGCACCAAGGCTCTGACTGGTTGGTGGCCAGTCCAAGGTGAGAGGTGGCAT
%HAS2
>Human
TCATACATCAAGCACCATGTCATATTGTTGTCCCTTCTTCCGCCTGCCACACTTATTGA
AATTAGAACTGTCTGTTTGGATTCTGAAAATGGCCTTTTAGACTCCTTATAAATGGTGA
CCCAGGGTAGGTTAGCCTTTTCACAGATTGCAAACATTTCCTTAAGTAGTCTGGATCTT
TTTTTTCATTTTTCGGTGCTCCAAAAAGGCAAACAGGCTTTGGATGATGAGGTGTGATC
GCCAACAATATAAGCAGCTGTGATTCCAAGGAGGAGAGAGACTCCAAAGAGTGTGGTTC
>Dog
TCATACATCGAGCACCATGTCATACTGTTGTCCCTTCTTCCTCCTGCCACACTTATTGA
AATGAGAACTGTCTGCTTGGATTCTGAGAATGGCTTTTTAGATTCCTTATAAATGGTGA
AGATATAAGTGGCTGATTTGTCCCTGCCCATGACTTCGCTAAAGATGTCCATCATATAA
TCTTCACAGATTGCAAACATTTTCGTAAGTAGTCTGGATCTTCTTGATAGGCAGCGATC
GCTCCAAAAAGGCAAACAGGCTTTGGATGATGAGGTGTGATGCTAAAAAGGCACCATAC
CTGTGATTCCAAGGAGGAGAGACACTCCAAAAAGTGTGGTTCCAATTATTCTCAGGATC
>Mouse
TCATACATCAAGCACCATGTCATACTGTTGTCCCTTCTTCCGCCTGCCACACTTATTGA
GATGAGAACAGTCTGTTTGGATTCGGAAAATGGCTTTTTAGATTCCTTATAAATGGTGA
CCCAGGGTAGGTCAGCCTTTTCACAGATTGCAAACATTTCCGTAAGTAGTCAGGGTCCT
CTTCTTCATTTTCCGGTGTTCCAAAAAGGCAAAGAGGCTTTGGATGATGAGATGCGAGC
GCCAACAATATAAGCAGCTGTGATTCCGAGGAGGAGAGACACTCCAAAAAGTGTAGTTC
```

Figure 14 : ***Exemple d'entrée dans le formulaire avancé****. L'utilisateur demande le calcul du ratio dN/dS pour les gènes MC1R et HAS2 chez trois espèces.*

Quelle que soit la manière utilisée pour soumettre les données, OMEGA réalise l'alignement multiple des séquences, détermine les ratios dN, dS et ω puis le test LRT si une branche principale est indiquée. Ces informations sont alors retournées à l'utilisateur sous la forme d'un affichage HTML sur le navigateur et *via* un lien vers une archive de type "tarball" des résultats qui est envoyée par e-mail. Cette archive a un nom de la forme omega_année_mois_jour_identifiant.tar. L'identifiant permet de conserver des noms de fichier unique et est composé de l'heure Posix et du nombre de nanosecondes correspondant au moment où la requête est arrivée au serveur. L'archive de résultats contient : un fichier HTML appelé UserDataset.htm.main qui correspond à l'affichage des résultats tel que sur le site web ; un dossier UserDataset.resultFolder qui contient une archive "tarball" compressée

en bzip2 pour chaque gène. L'archive d'un gène contient jusqu'à cinq fichiers : les fichiers de sortie brute de codeML pour les différents modèles (deux séquences, un calcul par branche, modèles H_1 et H_0) qui sont au nombre de 1 à 3 selon les requêtes; le fichier d'alignement au format Phylip; enfin, un fichier résultat qui regroupe les mêmes informations que dans le fichier UserDataset.htm.main mais avec un seul gène par fichier et sans formatage HTML.

Le module OMEGAbase inclut l'ensemble des valeurs calculées des ratio dN, dS et d'ω pour chacun des ~10.000 gènes orthologues de chacune des 10 espèces utilisées. Un ensemble de requêtes peut-être utilisé pour extraire ces informations de la base de données OMEGA.

II. Différenciation génétique entre races canines

Le second volet de ma thèse a consisté à rechercher les signatures génétiques qui différencient le génome des races canines actuelles, selon l'hypothèse qu'elles sont issues d'un processus de sélection artificielle. Cette sélection très récente n'a pas encore nécessairement impliquée les séquences codant pour les protéines, mais doit avoir influencé les patrons de polymorphismes génétiques des génomes des différentes races. La sélection des caractères désirés dans une race conduit à l'enrichissement du ou des variants alléliques qui confèrent ou contribuent à ce caractère. Ceci se traduit par un changement des fréquences alléliques des marqueurs proches sur le génome. Les patrons de variations génétiques qui reflètent ces événements récents de sélection et les gènes sous-jacents et leurs mutations, sont encore largement inconnus. L'étude du polymorphisme d'une population permet d'analyser le génome au delà des cadres ouverts de lecture des gènes codant pour des protéines et de réaliser des analyses qui criblent l'ensemble du génome.

II.1. Données expérimentales

Pour établir le catalogue des régions qui différencient le génome entre races de chiens, il faut disposer de données de séquences ou de génotypes de plusieurs individus au sein des différentes races étudiées. Cette partie de mon travail est effectuée dans le cadre du consortium européen de génétique du chien LUPA (7ème programme cadre européen) qui réunit 22 instituts européens dont l'objectif est d'identifier les bases génétiques de maladies génétiques similaires aux maladies de l'Homme chez le chien. La compréhension des bases génétiques de ces maladies chez le chien a pour but de prédire les bases génétiques

responsables de ces mêmes maladies génétiques chez l'Homme et de réduire la prévalence des maladies chez le chien.

Le consortium LUPA a tout d'abord développé une nouvelle génération de puce SNPs haute résolution capable de cribler le génome canin avec plus de 170.000 SNPs (1 SNP/13 kb). Deux générations de puces SNPs ont été développées ces dernières années, une puce de basse résolution 22K SNP (1 SNP/110 kb) par la technologie Illumina et une puce de moyenne résolution de 48K (1 SNP/ 50 kb) par la technologie Affymetix. Le groupe de notre collaborateur Matthew Webster du consortium LUPA (Université d'Uppsala en Suède) a réalisé la sélection des SNP de cette puce à partir des deux millions SNP identifiés au cours du séquençage du génome canin (Lindblad-Toh, *et al*., 2005) et grâce à l'apport de nouvelles données de séquences qui ont ciblé les zones de trous (gap) du séquençage initial.

II.1.1. La puce CanineHD 170K

Cette puce "CanineHD" contient 172.115 SNPs validés pour obtenir des génotypes régulièrement espacés en moyenne de 13 kb le long du génome. Les SNPs de la puce proviennent:

-à 65,1% de la comparaison entre la séquence du génome de la race boxer (séquence canine de référence) et la séquence partielle du génome de caniche (Kirkness, *et al*., 2003).

-à 25,4% des sites hétérozygotes dans la séquence du génome boxer.

-à 21,7% de la comparaison entre le boxer et des données de séquençages légers (0.5X) de neuf autres races canines.

-à 1,2% de la comparaison entre des séquences du génome du loup et/ou de coyote et du génome de référence.

-à 0,9% de données de reséquençages ciblés d'un individu de quatre races (lévrier irlandais, west highland white terrier, berger belge et shar-pei) et d'un loup pour disposer de SNP dans des intervalles non couverts par les autres méthodes de détection de SNPs.

race	Gentrain	LUPA complet	LUPA réduit
beagle	11	10	10
berger allemand	12	12	12
berger australien	0	1	0
berger belge tervueren	16	12	12
border collie	25	16	16
border terrier	0	25	25
bouvier bernois	12	12	12
boxer	0	8	0
braque de weimar	28	26	26
bull terrier anglais	0	8	0
bulldog anglais	13	13	13
caniche standard	12	12	12
cavalier king charles spaniel	0	5	0
chien d'élan suédois	12	12	12
chien de traîneau groenlandais	12	12	12
chien-loup tchécoslovaque	0	3	0
chihuahua	0	2	0
cocker anglais	0	2	0
cocker spaniel	15	14	14
dalmatien	0	7	0
doberman	0	25	25
english springer spaniel	0	3	0
epagneul breton	12	12	12

race	Gentrain	LUPA complet	LUPA réduit
eurasier	12	12	12
gordon setter	0	25	25
grand epagneul de münster	0	1	0
greyhound	11	11	11
lévrier irlandais	0	11	11
carlin	0	2	0
retriever de la nouvelle écosse	41	23	23
retriever du labrador	15	14	14
retriever golden	14	14	14
rottweiler	12	12	12
saarloos	0	2	0
samoyède	0	2	0
schipperke	0	25	25
schnauzer	0	3	0
setter anglais	12	12	12
shar pei	12	11	11
siberian husky	0	2	0
spitz finlandais	12	12	12
teckels	12	12	12
terre-neuve	0	25	25
terrier jack russell	13	12	12
terrier yorkshire	12	12	12
tretiever à poil plat	0	2	0
effectif total	358	509	456

*Table 2 : **Effectifs des races** des 3 jeux de données. Les races présentes uniquement dans le jeu "Gentrain" ne sont pas représentées ici.*

II.1.2. Test et validation de la puce CanineHD 170K

II.1.2.1. Génotypage: les jeux de donnees

Trois jeux de données ont été générés par le génotypage des chiens sur la puce CanineHD:

- Le jeu de données appelé "Gentrain" est constitué de 450 chiens de 26 races et a été conçu et utilisé pour valider la puce d'un point de vue technique, et pour la qualité de la détection et fiabilité de l'allèle détecté.

- Le jeu "LUPA complet" est constitué des chiens non apparentés du jeu 'Gentrain' et de chiens de plusieurs autres races. Ces chiens sont issus de cohortes contrôles provenant des

études génétiques d'associations analysées par les différents groupes constituant le consortium LUPA. Au total ce jeu est constitué des génotypes de 509 chiens de 46 races et de 15 loups. Ce jeu de données a servi à vérifier l'informativité des SNPs choisis et à réaliser des études d'associations de traits spécifiques de races.

- Le jeu "LUPA réduit" est constitué des populations pour lesquelles nous disposons de races représentées par au moins 10 individus. Ce jeu contient 456 chiens et constitue le jeu de données utilisé pour la recherche de régions de différenciation génétiques entre 30 races canines.

II.1.2.2. 'Call rate' et arbre phylogénétique

Un des deux critères majeurs d'évaluation de qualité d'une puce SNP est le taux de mesure effectif (call rate), qui est le taux de SNP dont les génotypes sont effectivement déterminés à l'effectif complet de 170K. Le jeu de données "Gentrain" a permis d'établir que le 'call rate' moyen est de 99,8% lorsque l'ADN est issu d'une extraction de sang et n'a pas été amplifié à posteriori (Rincon, *et al*., 2011). Ainsi cette mesure de contrôle qualité indique un niveau de détection pour chaque individu de la quasi-totalité des génotypes des SNPs de la puce.

L'autre critère majeur de qualité d'une puce est la fiabilité des génotypes obtenus. Cette fiabilité est évaluée par la reproductibilité des génotypes que l'on évalue par deux approches : (1) l'utilisation de l'ADN d'un même individu séparé en deux échantillons pour lesquels on teste la reproductibilité du génotypage de chaque SNP, et (2) l'analyse de la cohérence mendélienne à partir d'individus apparentés. La puce CanineHD présente une reproductibilité et une cohérence mendélienne supérieure à 99,9%. L'informativité des SNPs et surtout la possibilité de comparer les races grâce aux génotypes de ces marqueurs est essentielle pour la recherche des bases génétiques des traits fixés des races et des régions génomiques potentiellement soumises à la sélection artificielle. Cette informativité à été vérifiée en construisant un arbre par l'approche de neighbor-joining basé sur la distance génétique entre les individus du jeu "LUPA complet" (Figure 6). On observe trois éléments importants : les chiens d'une même race sont systématiquement regroupés ensemble ; la structure interne de l'arbre est faible ; les branches internes du loup et de la race boxer sont plus longues que les autres. La structuration des races en branches indiquent que les données de génotypages des 509 chiens par la puce Illumina CanineHD sont bien cohérentes avec la structure connue de la population canine à savoir une véritable collection d'isolats génétiques globalement issus d'une population commune puis séparés en une courte période de temps. La longueur des branches 'loup' et 'boxer' reflètent respectivement la plus grande distance génétique de la

population loup par rapport aux races canines et un plus fort nombre de SNPs polymorphes chez le boxer dû à l'utilisation de la séquence d'un boxer comme principale source d'identification des SNPs (1/4 des SNPs proviennent exclusivement du boxer et les autres proviennent d'une comparaison avec le boxer). Le faible niveau de regroupement des races en groupe de races indique que la création des races a aboutit à une entité génétique qui a un impact fort sur le génome de ces races en comparaison des relations historiques entre elles. La cohérence de cet arbre avec la structure connue de la population canine a permis de valider la précision et la fiabilité de la puce CanineHD et de garantir son utilisation pour les études d'associations et de recherche de différenciation génétiques.

II.1.3. Études d'association avec la puce 170k

Les premières analyses des données du jeux "LUPA complet" ont été une série d'études d'associations génétiques pan-génomique. Cette série d'études consiste à rechercher les causes génétiques de traits qui sont spécifiques de race. Pour cela les races ont été classées en différentes catégories selon le phénotype standard de la race pour des critères morphologiques (taille, poids, courbures des oreilles et de la queue, etc.) ou comportementaux (hardiesse, sociabilité). Pour corriger les biais potentiels liés à la structure des races, la significativité des associations obtenues entre un SNP et le phénotype analysé, est corrigée par une méthode de permutations dans laquelle ce sont les races qui sont soumises à la permutation. Chaque analyse est répétée 1000 fois en changeant aléatoirement le statut cas ou contrôle des races. La p valeur du meilleur SNP est retenue et la p valeur corrigée de chaque SNP correspond à la proportion de valeurs plus significative parmi les 1000 valeurs permutées par rapport à la valeur observée.

Ces études d'associations ont tout d'abord permis de reproduire les résultats d'associations antérieures basées sur d'autres puces SNPs. Par exemple pour l'étude génétique du phénotype poil fourni ('furnishing' qui correspond à la présence des moustaches, barbe et sourcils prononcés), le meilleur SNP (p valeur après permutation < 0.001) de l'analyse d'association se trouve à proximité (44 kb) de la mutation identifiée comme responsable de ce caractère (Cadieu, *et al*., 2009). Les données issues de la puce CanineHD permettent de reproduire l'association entre la courbure de l'oreille et un locus du chromosome 10 à 11,17 Mb (Jones, *et al*., 2008; Boyko, *et al*., 2010) ainsi que l'association de la taille moyenne des races avec la région du gène IGF1 (chromosome 15 à 44 Mb) précédemment identifié par Sutter (Sutter, *et al*., 2007). Notre étude indique une nouvelle association entre la variation de taille et la région du gène HMGA2 (Chromosome 10 à 11,07

Mb) qui est connu pour être associé au phénotype taille chez l'Homme (Gudbjartsson, *et al*., 2008; Lettre, *et al*., 2008; Weedon, *et al*., 2008). Au delà de la validation d'associations déjà connues, cette étude définit de nouvelles associations pour les phénotypes "courbure de la queue" et "boldness". Le phénotype de courbure de la queue est un phénotype très variable entre races de chien mais qui est fixé au sein de la plupart des races avec par exemple les races yorkshire terrier et chien-loup de saarloos qui présentent une forte courbure de la queue alors que les races retriever du labrador et dalmatien ont une queue non courbée. La queue, chez le chien, est un organe de communication, ainsi sa forme et son développement peuvent participer aux messages véhiculés entre animaux. Les variations de la morphologie de la queue sont multiples allant de l'anourie de certain schipperke, à la queue longue du barzoï en passant par la queue courte du bull terrier ou moyenne du dalmatien. Parmi les queues longues, on distingue un fouet tombant, horizontal ou relevé, qui se subdivise à nouveau en chandelle, en faucille ou encore enroulée. Les morphologies de la queue sont donc diverses et peuvent avoir un rôle important dans la perception qu'ont les congénères au cours d'une interaction. Les muscles sacro-coccygiens (SC) dorsaux sont des releveurs de la queue et peuvent être considérés comme un des paramètres permettant d'expliquer la courbure de la queue. Par ailleurs, la première vertèbre de la queue appartient à la colonne vertébrale. Par conséquent, une hypothèse est que la courbure de la queue peut être liée à des déficiences de la vertèbre de la queue qui font partie des anomalies squelettiques, ce qui renforce l'intérêt d'étudier ce phénotype. L'analyse de la courbure de la queue est réalisée par la comparaison de races possédant la queue courbée avec les races sans courbure de la queue présentes dans le jeu de données (Vaysse, *et al*., 2011). Cette analyse a permis de définir une association entre la courbure de la queue et un locus du chromosome 1 entre 96,26 et 96,96 Mb. Ce locus n'était pas présent dans une précédente étude de ce phénotype (Jones, *et al*., 2008) et suggère un locus candidat entre les gènes RCL1 et JAK2.

Une seconde analyse d'association a porté sur le phénotype 'boldness' ou 'hardiesse/ intrépidité' des races de chiens. Ce trait 'hardiesse/intrépidité' aurait été déterminant dans la spéciation du chien, dans la mesure où le tempérament individuel des loups ait pu être un élément important dans le succès de la mise en place d'une relation de domestication entre l'Homme et les futurs chiens. Un ensemble de 18 races appartenant à la catégorie 'bold' documenté par une étude précédente de l'association canine suédoise (Swedish Kennel Club), a été comparé à un ensemble de 19 races caractérisées par leur absence du phénotype 'bold' par la même étude. L'étude d'association a identifié un locus (p valeur après permutation

< 0.001) dans une région du chromosome 10 entre 10 et 12 Mb, un locus qui est commun au locus retrouvé associé au phénotype de la courbure de l'oreille et à la taille des races.

II.2. Recherche de régions de différenciation génétique : la méthode Fst-di

L'identification des régions de différenciation génétique repose principalement sur deux méthodes que sont l'analyse des haplotypes étendus et la recherche des allèles dont la fréquence est fortement différenciée. Les analyses d'haplotypes étendus caractérisent une région où un allèle rare sélectionné augmente sa fréquence si rapidement que son association avec les polymorphismes voisins n'est pas réarrangée par la recombinaison. Dans les approches de recherche des allèles fortement différenciés, un allèle sélectionné dans une population cause une plus grande différence de fréquence entre populations que pour des allèles sous évolution neutre. Mon travail dans la recherche de signatures de la différenciation génétique a constitué à identifier de manière statistique les régions de forte différenciation sur l'ensemble du génome entre races canines. Ces régions génomiques possèdent à priori des patrons de polymorphismes fortement différenciés entre races qui peuvent correspondre aux signatures liées à la création des races canines. La méthode que nous avons développée est basée sur une statistique appelée d_i introduite par J. Akey (Akey J. M., *et al*., 2010). Cette statistique d_i est dérivée de l'indice de fixation de wright (Fst) qui mesure le niveau de différenciation allélique entre populations.

II.2.1. Principe du Fst et de son indice dérivé le 'd_i'

L'indice Fst calculé sur une population constituée d'un ensemble de races permet de déterminer la part de la structure en race et la part de la variabilité individuelle dans les différences observées au sein de la population. Cette mesure de la divergence entre sous-populations d'une espèce est déjà utilisée comme indicateur de sélection positive (Akey Joshua M, *et al*., 2002; Oleksyk, *et al*., 2008), sans en être une preuve (Gardner, *et al*., 2007). Afin de pouvoir obtenir un calcul de Fst le plus fiable possible pour le plus grand nombre possible de races, l'analyse s'est effectuée à partir du jeu issu des 30 races canines représentées par au moins 10 individus (jeu "LUPA réduit"). Ces races sont représentées par 10 à 26 individus (pour le beagle et le braque de weimar respectivement; cf. table 2) avec une moyenne de 15 individus par race soit la prise en compte de 30 chromosomes en moyenne. Un total de 90% des SNPs de la puce est polymorphe dans ce jeu de données avec, en moyenne 40 SNPs spécifiquement polymorphes dans une seule race et en moyenne, 69% des

SNPs polymorphes dans une race donnée. Les SNPs ayant été découverts par la comparaison de séquences entre races, les SNPs non polymorphes au sein d'une race donnée peuvent être différents (polymorphes ou fixés différemment) dans une autre race. Par conséquent nous avons utilisé les SNPs polymorphes dans ce jeu de données présentant au moins 99% de génotypes déterminés. Le jeu de données utilisé pour notre analyse est ainsi constitué de 150.670 SNPs génotypés sur 456 chiens de 30 races.

Le principe du calcul du d_i consiste à intégrer des valeurs de Fst entre paires de races : $d_i = \sum_j \frac{F_{st_{ij}} - E(F_{st_{ij}})}{sd(F_{st_{ij}})}$ où $E(F_{st_{ij}})$ et $sd(F_{st_{ij}})$ représentent respectivement la valeur attendue et l'écart-type du Fst entre les races i et j calculée à partir de la totalité des SNPs pour déterminer des régions génomiques présentant une différenciation extrême.

II.2.1.1. Fst par paire de race

Le calcul de la statistique d_i est basée sur le calcul des Fst entre paire de races pour chaque SNP. Le Fst est une mesure de différenciation entre populations (Wright, 1943, 1951). L'indice Fst se calcule au niveau d'un marqueur ou d'un groupe de marqueurs et sur deux ou plusieurs populations. Le Fst représente la comparaison entre le nombre moyen de différences entre paires d'individus d'une même population et le nombre moyen de différences entre paires d'individus de populations différentes. La valeur de l'indice Fst varie entre 0 et 1. Un Fst de 0 indique que les variations de polymorphisme entre individus issus de deux populations et entre individus issus de la même population sont semblables. Un Fst de 1 indique que deux individus issus d'une même population présentent systématiquement le même polymorphisme et que toutes les variations entre les individus considérés sont liées à leur appartenance à des populations différentes. Lorsque la sélection affecte deux populations de manière différente, la différence entre les populations va être indiquée par des valeurs de Fst forte, (Fst>0,30 chez le chien). Particulièrement, les valeurs de Fst calculées à partir des marqueurs proches du site sélectionné seront plus élevées que les valeurs calculées sur les sites non-sélectionnés. L'indice Fst calculé à partir de deux races indique que ces deux races ont eu des pressions de sélection différentes au niveau des régions génomiques qui présentent les plus fortes valeurs. En comparant une race à chacune des autres alternativement, il est possible de pointer vers une région du génome sur laquelle la race considérée a subi une pression de sélection différente du reste de la population canine. Nous avons ainsi calculé le Fst de chacun des 150.670 SNPs pour chacune des 435 paires de races. Par exemple le SNP 'BICF2P563564' qui se trouve à la position 3.307.151 du chromosome 1 présente un Fst de 0.64 entre les deux populations constituées des races braque de weimar et caniche.

II.2.1.2. Normalisation et calcul du d_i

Bien que les différences de proximité génétique entre races canines soient faiblement prononcées, le phénomène de dérive génétique implique que les distributions de Fst pour chaque paire de race ne sont pas équivalentes, en effet le Fst gobal entre chacune des 435 combinaisons de paires de races du jeu "LUPA réduit" varie entre 0,12 et 0,48. Pour intégrer les différentes distributions de comparaisons de paires, chaque distribution est normalisée en retranchant la moyenne et en divisant par l'écart type des valeurs de Fst obtenues pour chaque SNP en comparant cette paire de race. Ainsi, lors de l'intégration des valeurs de Fst normalisées par race, la valeur attribuée à un SNP pour une race représentera une caractéristique de la race en question sans donner de poids plus important à une comparaison avec une race plus distante de la race d'intérêt qu'une autre race. Par exemple, la valeur normalisée issue de la comparaison des races braque de weimar et caniche au niveau du SNP 'BICF2P563564' est de 2. Pour une comparaison donnée, une valeur normalisée faible (proche de 0, voire négative) indique que les deux races concernées se différencient moins au niveau de ce marqueur qu'au niveau des autres marqueurs pris dans leur ensemble. Plus la valeur normalisée est forte plus le SNP représente une différence entre les deux races.

L'intégration des données de Fst normalisées s'effectue par sommation. Pour chacune des 30 races 150.670 valeurs de d_i sont calculées, une par SNP. La valeur de d_i d'un SNP pour une race donnée est la somme des valeurs de Fst normalisées calculées à partir de ce SNP pour les 29 comparaisons impliquant la race d'intérêt. Par exemple les valeurs de Fst normalisées des comparaisons de races pour le SNP 'BICF2P563564' vont de -1,15 à 3,4 (pour les comparaisons bouledogue anglais-chien d'élan suédois et terreneuve-border terrier respectivement). Pour les comparaisons de races impliquant la race caniche pour le SNP 'BICF2P563564', les valeurs de Fst normalisées s'étendent de -0,79 à 2,6 (comparaisons caniche-doberman et caniche-terreneuve). La somme de ces valeurs est de 38. La valeur de l'indice d_i du SNP 'BICF2P563564' pour le caniche est donc de 38. L'indice d_i calculé par race et par marqueur représente l'importance d'un marqueur donné dans la différenciation génétique de la race considérée par rapport à la population globale. Plus le d_i d'un SNP pour une race est élevé, plus la différence de polymorphisme de ce SNP est importante entre la race d'intérêt et chacune des autres races par rapport aux différences des autres SNPs.

II.2.1.3. Utilisation de fenêtres génomiques

Comme toute mesure réalisée sur des SNPs individuels, la valeur de d_i varie aléatoirement d'un SNP à l'autre. Par exemple, pour la race shar-pei, les valeurs de d_i varient de -24 à 114 avec des différences allant de 0 jusqu'à 130 entre deux marqueurs consécutifs (moyenne=14, écart type=15). Pour limiter cette variation aléatoire entre les valeurs de différenciation des marqueurs consécutifs nous utilisons les valeurs de plusieurs marqueurs consécutifs. Nous avons calculé la moyenne des valeurs des d_i pour chaque race sur des fenêtres de 150 kb, qui permettent de considérer les valeurs de d_i de 10 SNPs en moyenne. Le début de chaque fenêtre est décalé du début de la fenêtre précédente par un pas de 25 kb. Dans les régions les moins couvertes du génome certaines fenêtres ne contiennent pas de SNP alors que dans d'autres régions certaines fenêtres contiennent 23 SNPs. De plus les bornes des fenêtres dans lesquelles nous calculons les moyennes ne correspondent pas à des coordonnées de SNPs l'étendue réelle de la fenêtre entre le premier et le dernier SNP est inférieur à 150 kb. Nous avons par conséquent filtré le jeu de fenêtres pour ne conserver que celles qui contiennent au minimum 5 SNPs et d'une taille d'au moins 100 kb. Ces filtres aboutissent à la considération d'un jeu de 78.672 fenêtres contenant entre 5 et 23 SNPs (moyenne 10,6, écart-type 2,48) et couvrant chacune entre 100 kb et 150 kb (moyenne: 130,9 kb ; écart-type:10,8 kb). Pour chaque fenêtre ainsi déterminée, les moyennes de d_i sont calculées pour chaque race.

II.2.1.4. Extraction des valeurs extrêmes de la distribution des valeurs des fenêtres

Pour chaque race, les fenêtres dont le d_i moyen est le plus fort sont les intervalles du génome pour lesquels la race d'intérêt se distingue le plus de l'ensemble de la population des 29 autres races. Pour sélectionner les régions les plus différenciées et potentiellement cibles de la sélection artificielle, nous avons sélectionné 1% des fenêtres qui présentent la plus forte valeur de d_i (figure 15). Au total plus de 780 fenêtres d'environ 130 kb sont sélectionnées pour chaque race, soit plus de 23.000 fenêtres lorsque l'on cumule l'ensemble des données pour toutes les races.

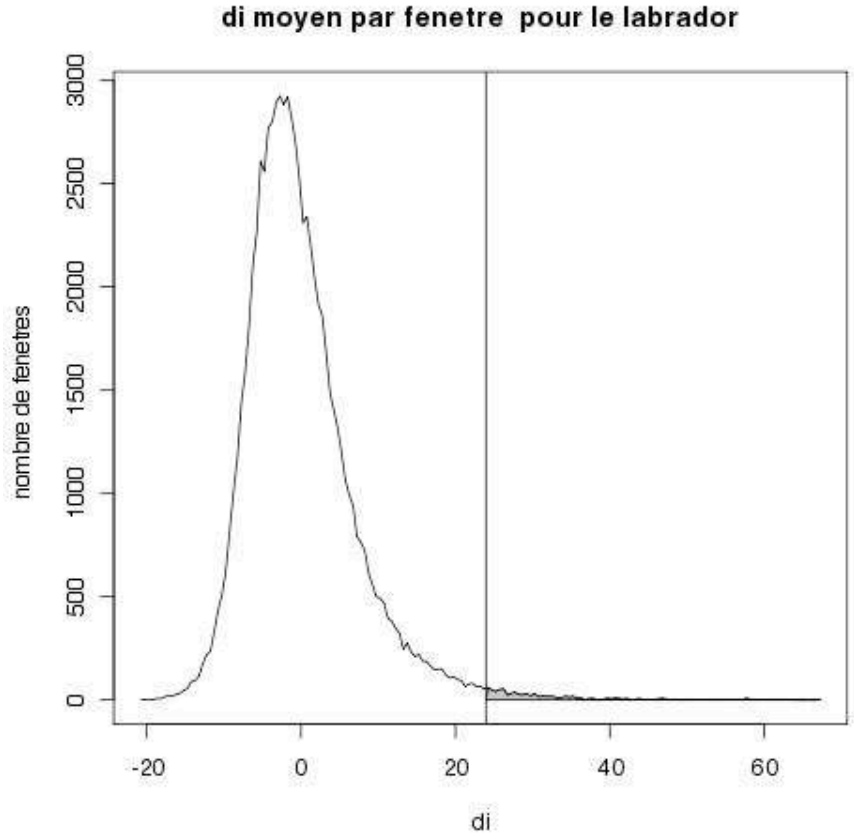


*Figure 15 : **Distribution des valeurs de di par fenêtres.** Pour chaque race -ici le labrador- nous considérons 1% des fenêtres de plus fort di.*

II.2.1.5. Projection des fenêtres en régions

Le chevauchement de certaines fenêtres sur plus des 3/4 de leur longueur implique qu'une grande région du génome différenciée dans une race sera détectée par plusieurs fenêtres de différenciation. Pour définir ces régions de différenciations, les coordonnées des fenêtres sont projetées sur le génome. Ainsi les fenêtres chevauchantes vont définir des régions distinctes de différenciation (Figure 16). Le catalogue des régions de différenciation des 30 races canines analysées est obtenu par projection des régions de différenciation de chaque race.

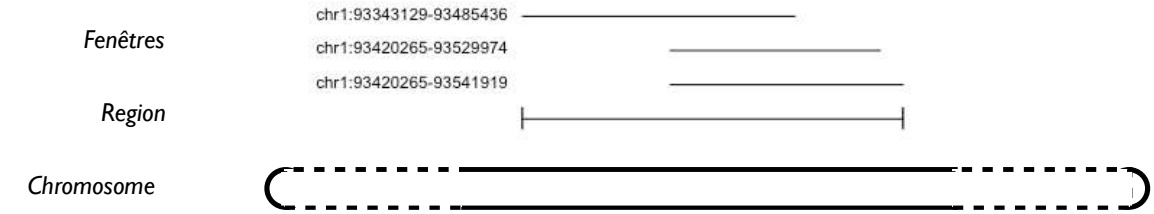


*Figure 16 : **Projection des fenêtres chevauchantes** différenciant les races en régions de différenciation. Ici les fenêtres chr1:93343129-93485436, chr1:93420265-93529974 et chr1:93420265-93541919 du rotweiller sont réunies dans la région de différenciation chr1:93343129-93541919.*

II.2.2. Développement d'un pipeline bioinformatique d'automatisation du calcul du Fst et du d_i

Au cours de la conception de cette étude, les principales étapes de la méthode ont été automatisées par la réalisation de différents scripts, ceci permettant de modifier facilement les traitements possibles après chaque étape. Ces informations d'intérêt sont les résultats des calculs suivants :

-les Fst entre chaque paire de races et pour chaque SNP,

-les valeurs de d_i de chaque SNP pour chaque race,

-les listes de fenêtres avec leur valeurs de d_i, afin de pouvoir sélectionner les fenêtres dont les valeurs sont extrêmes dans la distribution globale,

-les listes des régions de différenciation pour chaque race, qui constituent le résultat final de ce pipeline.

L'ensemble des scripts a été regroupé au sein d'un pipeline bioinformatique qui comprend cinq principales étapes : a) Le formatage et la répartition des données pour permettre de répartir le calcul pour chaque chromosome sur une machine d'une ferme de calcul dédiée ; b) le calcul du Fst par paire pour chaque sous ensemble au moyen d'un script Perl utilisant le module PopGen de la librairie BioPerl (Bioperl est une librairie qui permet la création de code en langage perl pour des applications en biologie) ; c) le calcul de l'indice d_i ; d) l'obtention de la liste des fenêtres avec leur moyenne de d_i et e) la définition des régions de différenciation génétique par races (voir le diagramme du pipeline Figure *17*).

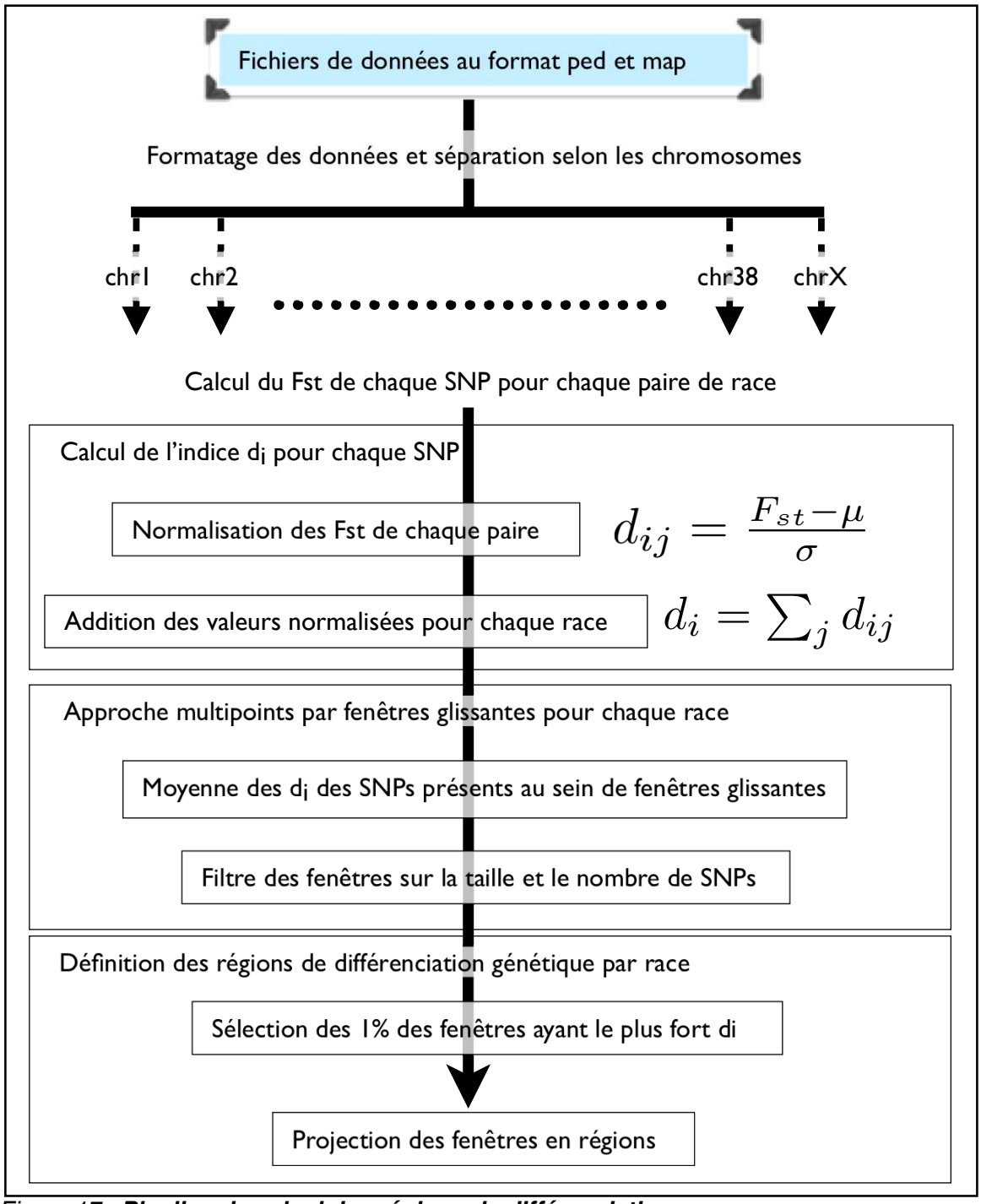


Figure 17 : ***Pipeline de calcul des régions de différenciation***

II.2.2.1. Format des données

Les données se présentent sous la forme d'un fichier PED et d'un fichier MAP tel qu'ils sont utilisés dans le logiciel PLINK (Purcell, *et al*., 2007) couramment utilisé dans les études d'associations. Les formats de ces fichiers sont présentés dans les figures 18 et 19. Un fichier d'entrée pour une population dans le module PopGen (http://www.bioperl.org/wiki/HOWTO:PopGen) est un fichier CSV (coma separated value) qui est constitué d'une ligne d'entête contenant le mot 'SAMPLE' et les noms des marqueurs séparés par des virgules. Les lignes suivantes représentent chacune un individu. Une ligne d'individu consiste en

l'identifiant de l'individu suivi du génotype de chacun des marqueurs. Le nom de l'individu et le génotype de chaque marqueur sont séparés par des virgules et les deux allèles d'un même génotype sont séparés par un espace. Le premier script du pipeline est un script 'bash' qui utilise les fichiers MAP et PED, sépare chaque chromosome dans un dossier puis converti les fichiers PED et MAP de chaque chromosome en un fichier CSV et le sépare en autant de fichiers CSV que de races.

Chromosome	identifiant	position(cM)	position(pb)
1	BICF2G630707759	0	3014448
1	BICF2S2358127	0	3068620
1	BICF2P1173580	0	3079928
1	BICF2G630707846	0	3082514
1	BICF2G630707893	0	3176980
1	TIGRP2P175	0	3198886
1	BICF2P1383091	0	3212349
1	TIGRP2P259	0	3249189
1	BICF2P186608	0	3265742
1	BICF2G630707908	0	3273096

*Figure 18 : **Format du fichier MAP**. Le fichier MAP décrit chaque marqueur par quatre colonnes : le numéro du chromosome sur lequel se trouve le marqueur, l'identifiant du marqueur, la position génétique du marqueur en centiMorgan et la position physique du marqueur sur le chromosome.*

id_famille	id_individu	père	mère	sexe	statut	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5
BoT_LU45	BoT_LU45	0	0	1	0	G G	A A	C C	A A	T T
BoT_LU48	BoT_LU48	0	0	2	0	G G	A A	C C	A A	T T
BoT_LU50	BoT_LU50	0	0	1	0	A A	C C	T T	T T	C C
BoT_LU60	BoT_LU60	0	0	2	0	G G	A A	C C	A A	T T
BoT_LU62	BoT_LU62	0	0	2	0	G A	A C	C T	A T	T C
BoT_LU63	BoT_LU63	0	0	2	0	G G	A A	C C	A A	T T

*Figure 19 : **Format du fichier PED**. Le fichier PED décrit les individus d'un pedigree. Chaque ligne représente un individu et est composée de six colonnes d'information sur l'individu : identifiants de la famille, de l'individu, des parents ; indications du sexe et du statut cas/contrôle pour un phénotype. Ces six colonnes sont suivies de paires de colonnes indiquant les génotypes de chaque marqueur.*

II.2.2.2. Calcul du Fst réparti sur un cluster de calcul

Chaque dossier de données d'un chromosome comporte 30 fichiers CSV, un par race. Le second script du pipeline est codé en langage perl en utilisant le module PopGen de bioperl (Stajich, *et al*., 2002; Stajich and Hahn, 2005). Ce module permet d'effectuer des calculs de génétique des populations tel que le calcul du Fst par la méthode décrite par Weir

(Weir, 1996). Le script utilise les fichiers CSV d'un dossier pour analyser les différentes races et écrire dans un fichier de sortie une matrice de Fst avec une ligne par marqueur et une colonne par paire de race. La division en chromosome des données permet de répartir le calcul des Fst sur 39 processeurs sur un cluster de calcul 'genocluster2' de la plate-forme bioinformatique GenOuest (http://www.genouest.org/). Enfin les différentes matrices de Fst sont réunies dans le premier fichier intermédiaire conservé.

II.2.2.3. Calcul de l'indice d_i

La troisième étape du pipeline consiste à normaliser les valeurs de Fst des différentes comparaisons. Pour cela deux scripts en language R ont été développés (R Development Core Team, 2009). Le premier script normalise les valeurs de Fst des comparaisons de chaque paire de races et crée un fichier par race qui contient les valeurs normalisées des comparaisons de cette race avec chacune des 29 autres. Le second script traite un fichier de race pour calculer la somme des valeurs normalisées par race. Ce second script est utilisé en parallèle sur chacun des 30 fichiers de races. Enfin les résultats des 30 calculs sont regroupés dans le second fichier intermédiaire conservé qui comporte une ligne par marqueur et une colonne par race.

II.2.2.4. Approche multipoints par fenêtres glissantes

La quatrième étape de la méthode développée consiste à former des fenêtres issues du regroupement des valeurs individuelles et consécutives de d_i calculées par SNPs. Le fichier des d_i par marqueur et par race est lu par un script R qui définit les bornes de fenêtres de 150 kb glissantes de 25 kb sur chacun des chromosomes à partir de la position du premier SNP du chromosome. Au sein de chaque fenêtre, le script va récupérer les informations de chaque SNP pour chaque race et enregistrer dans le fichier des fenêtres une ligne comportant 8 colonnes d'informations sur la fenêtre: 1- le numéro de la fenêtre, 2- la liste des SNPs présents dans cette fenêtre avec 3- le nombre de SNPs présents; 4- le chromosome sur lequel cette fenêtre est définie avec 5- la position du premier SNP de la fenêtre et 6- la position du dernier SNP, ces positions définissent les bornes exactes de la fenêtre d'intérêt; enfin 7- la position centrale et 8- la taille réelle de la fenêtre. Les fenêtres contenant moins de 5 SNPs et/ou ayant une taille inférieure à 100 kb sont éliminées. Les colonnes 9 à 38 correspondent aux valeurs de d_i moyen pour chacune des races dans la fenêtre.

II.2.2.5. Définition des régions de différenciation génétique par race

Dans une race donnée, la fenêtre qui différencie le plus cette race de l'ensemble de la population canine est la fenêtre qui présente le plus fort d_i. La cinquième étape consiste à sélectionner au sein de chaque race le 1% des fenêtres dont la valeur de d_i est la plus forte. Le script projette les coordonnées des fenêtres qui se chevauchent en une région au sein d'une race.

II.3. Description des régions identifiées

Ce pipeline nous a permis d'identifier 165 à 257 régions de différenciation allélique par race (moyenne:213, écart type: 22). La taille des régions varie de de 100 kb à 1404 kb (moyenne 205 kb, écart type 105 kb) pour une couverture de 32,16% du génome. La moyenne de la taille des régions au sein d'une race varie entre 186 kb et 236 kb (moyenne : 206 kb; écart type 12 kb). La taille minimale des régions d'une race varie entre 100 et 109 kb (moyenne : 102 kb; écart type 2 kb). La taille maximale des régions d'une race varie entre 569 kb et 1404 kb (moyenne : 880 kb; écart type 257 kb). Les distributions des tailles des régions de différenciation allélique ont une moyenne semblable entre races mais les tailles des plus grandes régions varient d'une race à l'autre. La moyenne des tailles des régions d'une race varient peu, et la proportion du génome canin couvert par les régions de sélection est similaire d'une race à l'autre (entre 1,54 et 1,89 % avec une moyenne de 1,73 et un écart type de 0,09)

Deux races différentes peuvent se différencier du reste de la population canine sur une même région. Pour décrire les régions de différenciation partagées entre races et les régions retrouvées uniquement dans une race donnée, nous avons projeté les coordonnées des régions des différentes races en un catalogue unique de régions de différenciation canine. Cette projection permet de définir 2999 régions de différenciation alléliques entre race canine d'une taille de 271 kb en moyenne (cf. table 3 et table 4). Parmi ces régions 1346 (45%) différencient plusieurs races du reste de la population canine. Les 1653 autres régions sont spécifiques d'une seule race.

	#	min(kb)	max(kb)	moyenne(kb)	écart type(kb)	couverture(%)
Communes	1346	101	2835	384	229	20,40
Spécifiques	1653	100	896	180	72	11,76
Toute régions	2999	100	2835	271	192	32,16

*Table 3 : **Statistiques des régions de différenciation** : effectifs et tailles en kb des régions qui différencient plusieurs races (Communes) et des régions spécifiques de races (Spécifiques).*

	min	max	moyenne	écart type
# régions par races	165	257	213	22
Moyenne de taille au sein de la race (kb)	186	236	206	12
couverture (%)	1,54	1,89	1,73	0,09

*Table 4 : **Statistiques des régions de différenciation par race***

Parmi les 1346 régions partagées de 384 kb en moyenne entre plusieurs races, 1164 contiennent entre 1 et 48 gènes annotés (moyenne : 5,1). Parmi ces 1164 régions, 1065 contiennent des gènes codant pour des protéines dont 181 contiennent un unique gène codant.

Chaque race compte entre 32 et 76 régions propres de différenciation pour un total de 1653 régions spécifiques de race. La taille moyenne des régions de différenciation propre à chaque race est globalement inférieure à la taille moyenne de l'ensemble des régions de différenciation de chaque race (p valeur du test Mann Withney = $1e^{-9}$; table 5). De même, au sein de chaque race, les régions spécifiques sont plus petites que l'ensemble des régions de la race. La moyenne des tailles des régions d'une race varie encore peu, la proportion du génome canin couverte par les régions de sélection spécifiques de race est similaire d'une race à l'autre. Parmi les 1653 régions spécifiques d'une race, 1220 contiennent entre 1 et 27 gènes annotés (moyenne:2,7). Parmi ces 1220 régions, 1093 contiennent des gènes codant pour des protéines dont 362 contiennent un unique gène codant.

d_j	min	max	moyenne	écart type
#régions spécifiques par race	32	76	55	11
moyenne de taille au sein de la race (kb)	158	211	180	13
couverture (%)	0,23	0,55	0,39	0,08

*Table 5 : **Statistiques des régions de différenciation spécifiques par race***

Plusieurs contrôles suggèrent la validité des régions de différenciation allélique identifiées. Premièrement, nous comparons les régions avec les régions de différenciation décrites par Akey et al. (Akey J. M., *et al*., 2010). Les auteurs décrivent 155 vastes loci génomiques de 1 Mb contenant 1630 gènes (~11 gènes par locus) candidats à la sélection

artificielle canine. Nous retrouvons 150 de ces régions (96%) par 311 régions 'd_i'. Par ailleurs, plusieurs locus de différenciation de races canines ont été préalablement caractérisés et publiés. Nous avons recherché si les régions de différenciation identifiées dans cette étude recoupaient les régions connues. Nous nous sommes intéressés à trois types de différenciation :

1- Les régions de différenciation qui permettent de distinguer une seule race des autres de la population. Par exemple la race Shar-Pei: Akey et al. (Akey J. M., *et al*., 2010) et Olsson et al. (Olsson, *et al*., 2011) ont mis en évidence que la région du gène HAS2 (chr13:23348773-23364912) ainsi qu'une duplication spécifique en aval du gène sont associés au caractère "peau plissée" caractéristique et spécifique de la race.

2- Les différenciations qui permettent de distinguer un petit groupe de race du reste de la population comme pour le phénotype "poils fournis" caractérisé par la présence de moustaches et sourcils fournis chez plusieurs races dus à une mutation du gène RSPO2 (chr13: 11685070-11706046) (Cadieu, *et al*., 2009).

3- Les différenciations sur un caractère à variation continue tel que la différence de taille : Le phénotype 'petite taille' est associé à un haplotype incluant le gène IGF1 (chr15:44213546-44285854) qui explique la majeure partie des différences de taille entre races de chiens (Sutter, *et al*., 2007).

Les régions de différenciation identifiées dans ce travail incluent ces trois catégories de signatures connues. Au total, les régions de différenciation identifiées dans cette étude englobent 19 des 20 loci répertoriés dans la littérature qui expliquent ou sont associés à des phénotypes connus de différenciation de race. La détection des signaux connus de différenciation chez le chien illustre la sensibilité de notre méthode basée sur l'indice d_i pour détecter des régions cibles de la sélection artificielle. Cette méthode se base sur la sélection de 1% des fenêtres, ce qui, comme toute méthode de sélection de données extrêmes (oultiers), implique l'inclusion de faux positifs. Pour filtrer la présence de faux positifs dans cette liste de régions, nous avons généré des données issues de simulations de génotypes afin d'établir une distribution théorique à comparer aux données observées.

II.4. Sélection statistique des régions les plus différenciées

Les régions théoriques utilisées sont issues de données de génotypage simulées par le Dr. Erik Axelson de l'université d'Uppsala qui ont été traitées avec le même pipeline bioinformatique que les données réelles. Ceci a permis de comparer les caractéristiques des

régions observées et théoriques afin de retenir les régions qui présentent une signature de différenciation la plus robuste possible.

II.4.1. Données simulées

La simulation du jeu de données est basée sur le principe de coalescence en prenant en compte les paramètres de taille de population et de taux de recombinaison du jeu de données réel. Ces données simulées représentent un patron de diversité génétique que l'on s'attend à observer au sein et entre les races canines en l'absence de sélection. La simulation du jeu de données a été faite en trois étapes: Tout d'abord l'inférence du taux de recombinaisons a été réalisée à l'aide du package LDhat (Auton and McVean, 2007). Les données de chaque chromosome ont été divisées en blocks de 2000 SNPs se chevauchant sur 200 SNPs. Les taux de recombinaison ont été estimés pour chaque bloc séparément et les estimations de chaque bloc sont concaténées pour obtenir des estimations sur les chromosomes entiers. Pour vérifier la qualité de cette carte, les taux de recombinaison sont moyennés sur 5 Mb et comparés avec la carte de liaison génétique canine précédemment publiée (Wong, *et al*., 2010; Axelsson, *et al*., 2011).

Dans une seconde étape, la modélisation des goulets d'étranglements de chaque race est basée sur des paramètres de tailles effectives de population, nombre de générations depuis la domestication et taux de mutations estimés dans des études antérieures (Lindblad-Toh, *et al*., 2005; Gray M, *et al*., 2009), des taux de recombinaisons estimés à partir des données réelles et les distributions de fréquences alléliques. Plusieurs simulations de jeux de SNPs ont été réalisées avec MaCS (Chen, *et al*., 2009) pour chaque taille d'échantillon du jeu de données expérimentales pour une gamme de goulet d'étranglement allant de 1‰ à 3% de la taille effective estimée de la population loup. La décroissance du déséquilibre de liaison de chaque simulation est comparé à la décroissance du déséquilibre de liaison des jeux réels pour sélectionner l'estimation des goulets d'étranglement les plus adaptés pour chaque race.

La simulation principale utilise les tailles des goulets d'étranglement modélisés pour chaque race, les taux de recombinaisons et les fréquences alléliques issues des jeux de données expérimentales pour créer un jeu de données unique des génotypes simulés pour chaque race. Le jeu de données simulé final est constitué des génotypes de plus de 170.000 SNPs de 30 races.

II.4.2. Test de p valeur marginale

Nous avons utilisé le pipeline développé dans ce travail de thèse pour analyser les données de SNPs simulées dans les mêmes conditions que pour les données de SNPs réels. Ces analyses ont permis de définir 3441 régions simulées qui couvrent 32.39% du génome à comparer avec les 2999 régions issues des données expérimentales et qui couvrent 32,16% du génome.

II.4.2.1. Taille

La notion de balayage sélectif implique que plus une sélection est forte et récente, plus l'effet sera détectable à des positions éloignées du variant sélectionné sur le chromosome. En effet, les régions simulées sans sélection se différencient des régions réelles principalement sur le critère de la taille (Figure 20). Les régions simulées d'une race ont une taille minimale de 100 kb, moyenne de 179 kb et une taille maximale de 921 kb. La distribution des tailles de régions détectées dans le jeu de données expérimentales est constitué de valeurs plus fortes que la distribution des tailles de régions détectées dans le jeu de données simulées sans sélection (Mann-Withney P.val=1.7e^{-58}). Cette tendance relate une différence significative des distributions de tailles entre les régions observées et théoriques.

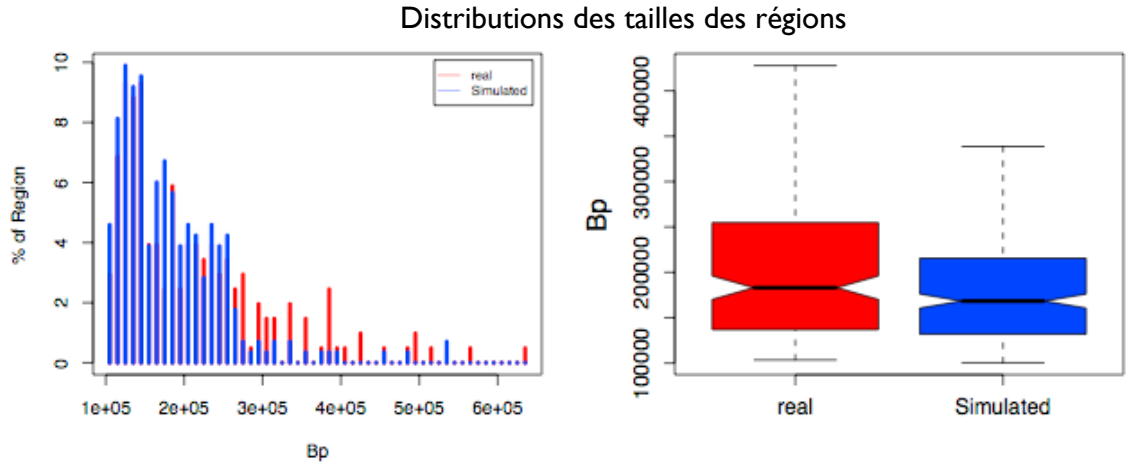


*Figure 20 : **Comparaison des tailles des régions** issues des données expérimentales et simulées. Les distributions des tailles des régions issues des génotypes expérimentaux (rouge) et simulés (bleu) de la race caniche sont représentées sous forme d'histogramme et de boites à moustache.*

II.4.2.2. Calcul des p valeurs

Nous considérons que, au sein d'un race, les régions simulées sans sélection définissent la distribution théorique des tailles des régions observées sous l'hypothèse H_0 d'absence de sélection. La détection de grandes régions est moins probable sous l'hypothèse H_0 que sous

l'hypothèse de sélection. La proportion de régions détectées au dessus d'une taille donnée dans une race est la probabilité de détecter des régions d'au moins cette taille sous l'hypothèse de non-sélection. Par conséquent, nous associons à chaque région issue du jeu expérimental de chaque race une p valeur qui correspond à la proportion de régions issues de données simulées qui sont plus grandes que cette région. Nous faisons une exception pour les régions qui sont plus grandes que toutes les régions simulées pour une race donnée nous n'attribuons pas une p valeur de 0, mais nous les traitons comme si une et une seule région simulée était plus grande.

Nous identifions 11 à 33 régions par race avec une p valeur <0.05. En projetant ensemble les coordonnées des régions identifiées dans les différentes races, nous obtenons un jeu final de 530 régions de différenciation allélique. Ces régions couvrent 9,38% du génome canin (cf Table 6 et Table 7). Parmi ces régions, 115 (22%) différencient plusieurs races du reste de la population canine. Les 415 autres régions sont spécifiques de race. Parmi les 115 régions partagées entre plusieurs races, 107 contiennent entre 1 et 48 gènes annotés (moyenne: 7,4). Parmi ces 107 régions, 97 contiennent des gènes codant pour des protéines dont trois contiennent un unique gène codant.

d_i Pval<0.05	#	min(kb)	max(kb)	moyenne(kb)	écart type(kb)	couverture(%)
Communes	115	280	2116	616	297	2,8
Spécifiques	415	273	1211	401	117	6,58
Toute régions	530	273	2116	448	194	9,38

*Table 6 : **Statistiques des régions de différenciation de p valeur <0.05**: effectifs et tailles en kb des régions qui différencient plusieurs races (Communes) et des régions spécifiques de races (Spécifiques).*

d_i Pval<0.05	min	max	moyenne	écart type
#régions par race	11	33	25	7
moyenne de taille au sein de la race (kb)	363	532	430	42
couverture (%)	0,2	0,59	0,4	0,1

*Table 7 : **Statistiques des régions de différenciation de p valeur <0.05 par race***

Chaque race contient 4 à 24 régions qui contribuent à sa caractérisation spécifique (cf Table 8). La somme des régions spécifiques des 30 races aboutit à un total de 415 régions spécifiques. La taille moyenne des régions de différenciation propre à chaque race est toujours inférieure à la taille moyenne de l'ensemble des régions de différenciation de chaque race (valeur p du test Mann Withney = 0.027) (cf TABLE II.4-1). Parmi les 415 régions spécifiques d'une race, 375 contiennent entre 1 et 47 gènes annotés (moyenne: 4,9). Parmi

ces 375 régions, 338 contiennent des gènes codant pour des protéines dont 55 contiennent un unique gène codant.

d_i Pval<0.05	min	max	moyenne	écart type
#régions spécifiques par race	4	24	14	6
moyenne de taille au sein de la race (kb)	353	500	406	41
couverture (%)	0,6	0,36	0,22	0,09

Table 8 : ***Statistiques des régions de différenciation de p valeur <0.05 spécifiques par race***

II.5. Recherche de perte d'hétérozygotie

En parallèle de notre travail sur l'indice d_i, Abhirami Ratnakumar de l'université d'Uppsala a utilisé un autre indice, appelé Si, pour rechercher des signatures de différenciation de races basé sur la perte d'hétérozygotie. L'indice Si s'initie à partir du calcul de l'hétérozygotie relative au sein d'une race. L'hétérozygotie relative d'une fenêtre génomique dans une race donnée est la proportion de SNPs polymorphes au niveau de cette fenêtre, dans cette race particulière. Les hétérozygoties relatives de chaque race, ont été calculées sur les fenêtres que nous avons défini pour le calcul du d_i et comparées aux hétérozygoties relatives de chacune des autres races pour la même fenêtre. Par exemple la comparaison entre les races beagle et labrador revient à calculer l'hétérozygotie relative de la race beagle sur la fenêtre considérée divisée par la somme des hétérozygoties relatives des deux races. À l'inverse, la comparaison entre les races labrador et beagle sur une fenêtre revient à calculer l'hétérozygotie relative de la race labrador sur la fenêtre divisée par la somme des hétérozygoties relatives des deux races. Les étapes suivantes sont semblables aux étapes du calcul du d_i des SNPs. Les 870 distributions des comparaisons sont normalisées, les valeurs normalisées des différentes comparaisons d'une race avec chacune des 29 autres sont additionnées pour obtenir une valeur de Si par race. Plus le Si d'une fenêtre est faible pour une race et plus la race d'intérêt présente une perte de polymorphisme importante qui implique que la sélection du meilleur pourcentage des fenêtres se réalise sur les fenêtres de faible Si, soit la zone proche de zéro de la distribution des valeurs de Si. Après regroupement par régions, les p valeurs sont déterminées en comparant la taille des régions Si aux tailles des régions Si calculées à partir des données simulées.

L'indice Si permet d'identifier 44 à 76 régions d'hétérozygotie réduite par race, soit 1804 régions au total qui après projection constituent un jeu final de 990 régions Si de perte

de diversité génétique (cf. Table 9 et Table 10). Parmi ces régions 359 (36%) différencient plusieurs races du reste de la population canine. Les 631 autres régions sont spécifiques de race.

Si	#	min(kb)	max(kb)	moyenne(kb)	écart type(kb)	couverture(%)
Communes	359	212	3769	530	385	7,52
Spécifiques	631	206	2511	298	143	7,43
Toute régions	990	206	3769	382	281	14,94

*Table 9 : **Statistiques des régions de perte d'hétérozygotie**: effectifs et tailles en kb des régions qui différencient plusieurs races (Communes) et des régions spécifiques de races (Spécifiques).*

Si	min	max	moyenne	écart type
#régions par race	44	76	60	9
moyenne de taille au sein de la race (kb)	265	448	322	44
couverture (%)	0,58	0,95	0,75	0,08

*Table 10 : **Statistiques des régions de perte d'hétérozygotie par race***

II.6. Intégration des approches d_i et Si

Les régions basées sur la différenciation (d_i) et sur la perte de diversité (Si) des races représentent des signaux complémentaires. L'analyse d_i est basée sur des niveaux de différenciation entre populations alors que l'analyse Si est basée sur la comparaison des niveaux de diversité interne aux populations. En effet, la sélection forte d'une nouvelle mutation sur une population va simultanément générer une baisse de la diversité au sein de la population sélectionnée par balayage sélectif et une augmentation des différences avec les populations qui ne subissent pas de sélection. Dans ce cas précis, les deux indices devraient permettre de détecter des régions communes. Cependant les analyses de différenciation et de perte de diversité peuvent capturer des événements différents (Storz, 2005; Oleksyk, *et al*., 2010). Par exemple si plusieurs races subissent une pression de sélection sur un même locus ou deux locus proches qui aboutissent à la fixation d'allèles différents, ces différentes races perdront en diversité et aucune ne se distinguera des autres en hétérozygotie relative, alors que les comparaisons des Fst pourront être significatives. Par ailleurs, si au sein d'une race la sélection agit sur les variations préexistantes, la différence d'hétérozygotie relative sera maximale juste après le balayage sélectif, mais le niveau de diversité de la population sélectionnée va être restauré plus rapidement que dans le cas d'une sélection sur une nouvelle

mutation (Wiehe, 1998) alors que la différence entre les populations restera détectable (Kayser, *et al*., 2003).

Pour obtenir un catalogue le plus exhaustif possible, nous regroupons les deux jeux de signatures issues de l'analyse d_i et Si, et obtenons au total 2546 régions qui couvrent 20,61% du génome canin. En projetant ensemble les coordonnées des régions des différentes races, nous obtenons un total de 1218 régions (cf Table 11 et Table 12). Parmi ces régions 469 (39%) différencient plusieurs races du reste de la population canine. Les 749 autres régions sont spécifiques de race. Parmi les 469 régions partagées entre plusieurs races, 432 contiennent entre 1 et 72 annotations de gènes (moyenne: 7,1). Parmi ces 432 régions, 394 contiennent des gènes codant pour des protéines dont 36 contiennent un unique gène codant.

Si+ d_i	#	min(kb)	max(kb)	moyenne(kb)	écart type(kb)	couverture(%)
Communes	469	213	5373	593	427	10,99
Spécifiques	749	206	1386	325	130	9,62
Toute régions	1218	206	5373	428	312	20,61

*Table 11 : **Statistiques des régions candidates à la sélection**: effectifs et tailles en kb des régions qui différencient plusieurs races (Communes) et des régions spécifiques de races (Spécifiques).*

Si+ d_i	min	max	moyenne	écart type
#régions par race	70	105	85	9,89
moyenne de taille au sein de la race (kb)	308	450	349	33
couverture (%)	0,9	1,4	1,12	0,12

*Table 12 : **Statistiques des régions candidates à la sélection par race***

Chaque race contient entre 14 et 37 régions spécifiques. La taille moyenne des régions de différenciation spécifique de race est inférieure à la taille moyenne de l'ensemble des régions de différenciation de chaque race (p valeur du test Mann Withney = 0,046; Table 13). Parmi les 749 régions spécifiques d'une race, 635 contiennent entre 1 et 47 annotations de gènes (moyenne: 4,3). Parmi ces 635 régions, 574 contiennent des gènes codant pour des protéines dont 98 contiennent un unique gène codant.

Si+ d_i	min	max	moyenne	écart type
#régions spécifiques par race	14	37	25	6
moyenne de taille au sein de la race (kb)	265	388	327	34
couverture (%)	0,19	0,46	0,32	0,07

*Table 13 : **Statistiques des régions candidate à la sélection spécifiques par race***

Ces régions contiennent 17 des 20 locus associés à des phénotypes connus de différenciation de race et illustrent notamment les trois catégories de traits sélectionnés que sont : le locus contenant le gène HAS2 pour les traits spécifiques de races ; le locus RSPO2 pour les traits dichotomiques partagés par un petit groupe de races et le locus IGF1 pour les caractères à variation continue entre races. L'intégration des analyses de d_i et Si permet d'établir un catalogue exhaustif de régions de différenciation et de perte de diversité génétique qui constitue le catalogue de régions candidates ciblées par la sélection artificielle. L'analyse du contenu en gènes de ces locus permet la recherche des principales voies fonctionnelles à priori ciblées par la sélection artificielle.

II.7. Recherche d'enrichissements fonctionnels

II.7.1. Les ontologies de gènes

Parmi les régions candidates à la sélection artificielle identifiées, 2191 contiennent en moyenne 4,5 gènes (de 1 à 69 gènes dans une même région). Quand une région contient plusieurs gènes, il n'est pas possible d'identifier quel est le ou les gènes qui sont potentiellement la cible de la sélection artificielle. Par conséquent nous avons sélectionné, au sein de chaque race les régions Si et d_i qui contiennent un seul gène afin de limiter les risques de biais lors des analyses d'enrichissements fonctionnels. Les régions d_i ne contenant qu'un seul gène permettent de dresser une liste de 119 gènes. Les régions Si permettent de dresser une liste de 272 gènes. Nous avons recherché les orthologues humains de ces gènes dans la base de données Ensembl (Ensembl v.62) à l'aide de l'outil Biomart (Durinck, *et al*., 2005). Nous avons obtenu 72 gènes humains en orthologie 1:1 avec les gènes détectés dans les régions d_i et 176 gènes humains en orthologie 1:1 avec les gènes détectés dans les régions Si. Enfin, nous avons utilisé l'outil en ligne WebGestalt (Zhang B., *et al*., 2005a) pour retrouver les termes GO associés à ces gènes othologues et tester la significativité des enrichissements en termes GO. Cet enrichissement est testé en comparaison avec une liste de référence constituée de l'ensemble des gènes humains en relation d'orthologie 1:1 avec les gènes canins et de longueur moyenne similaire à la longueur moyenne des gènes testés (230 kb). Seules les catégories GO enrichies significativement (p valeur du test hypergeometrique $< 0,05$) sont considérées.

La liste des gènes détectés avec l'indice Si est enrichi en gènes de 40 catégories GO et la liste des gènes détectés avec l'indice d_i est enrichi en gènes de 6 catégories GO. Au delà de la différence quantitative, les deux indices permettent la détection de catégories

qualitativement distinctes. D'une part les gènes détectés avec l'indice Si appartiennent préférentiellement aux catégories de processus développementaux, de développement des organes, liées au système nerveux central et aux voies de la pigmentation. D'autre part les gènes détectés avec l'indice d_i incluent la communication cellulaire et la transduction du signal qui sont représentées par 16 et 15 gènes respectivement. Parmi ces gènes, nous constatons la présence du gène codant pour le récepteur de IGF1, le polymorphisme du gène IGF1 a déjà été montré comme associé à des différences de tailles entre races canines (Sutter, *et al*., 2007), IGF1R est donc un bon candidat pour expliquer une autre part des variations de croissance. Le gène ANGPT1 est aussi candidat à la sélection artificielle. ANGPT1 joue un rôle dans le développement vasculaire et a été démontré sous sélection positive dans les populations humaines tibétaines (Wang Binbin, *et al*., 2011).

II.7.2. Corrélation avec les termes OMIM

Les races canines sont prédisposées à de nombreuses maladies génétiques, soit par la sélection directe d'un phénotype de maladie (comme pour la chondrodysplasie) ou par effet d'auto-stop génétique. Nous avons recherché si les gènes inclus dans les régions de différenciation et de perte de diversité sont plus fréquemment impliqués dans des maladies génétiques. Pour cela, nous avons extrait l'ensemble des orthologues humains des gènes chevauchant les régions détectées par l'indice d_i ou par l'indice Si lorsque ceux-ci présentent une orthologie 1:1 entre le chien et l'humain. Nous avons extrait les statuts OMIM (Online Mendelian Inheritance in Man) (McKusick, 1998) de chacun de ces gènes. Globalement 1829 gènes présents dans les régions Si et 1126 présents dans les régions d_i sont annotés en tant que gènes modulateurs de maladies humaines. Plus précisément 384 et 255 orthologues de gènes présents dans les régions Si et d_i respectivement ont un statut décrit de "morbidité". Nous avons téléchargé les statuts OMIM de tous les gènes humains en orthologie 1:1 avec le chien et de même taille moyenne que les orthologues des gènes des régions Si et d_i (230 kb). Cette liste de gènes a été utilisé comme référence pour tester la corrélation entre l'appartenance des gènes à une région d_i ou Si et leur statuts OMIM. Le nombre de gènes ayant une annotation OMIM ou un statut "morbidité" parmi les orthologues 1:1 chien:humain des gènes des régions d_i ou Si ne diffère par significativement du nombre d'orthologue 1:1 chien:humain total ayant de telles annotations (chi-deux ; $p>0.05$).

Discussion et perspectives

I. Sélection positive naturelle chez le chien

I.1. Le contexte phylogénétique

Pour établir le catalogue complet des gènes sous sélection positive dans le génome canin, nous avons utilisé des alignements de séquence entre les gènes canins et les orthologues 1:1 (collaboration avec l'équipe DYOGEN de l'ENS Paris) de 9 espèces (cinq primates, deux rongeurs, deux laurasiatheria) et identifié 633 gènes sous sélection positive dans le génome canin. Ces analyses se basent sur deux prérequis indispensables : l'identification de gènes en relation d'orthologie 1:1 entre les 10 espèces et l'alignement de séquence fiable de ces gènes.

L'identification des orthologues vrais est indispensable pour toute analyse en génomique comparative. L'utilisation de gènes en orthologie 1:1 est à la base d'un paradoxe dans la recherche de sélection positive. Un gène en relation d'orthologie 1:1 entre espèces implique une forte conservation au cours de l'évolution sans duplication. Les analyses sont réalisées sur des gènes fortement conservés dans leur structure et à priori dans leurs fonctions au cours de l'évolution. Nous recherchons donc les modifications par le calcul du dN/dS parmi les gènes les plus fixés au cours de l'évolution des mammifères. Un changement dans l'un de ces gènes doit induire des conséquences phénotypiques qui peuvent être la cible de la sélection positive et expliquer des différences importantes entre les adaptations des différentes espèces. Le contexte phylogénétique utilisé ici comporte plus d'espèces proches de l'Homme que du chien. Ce biais est dû au nombre de génomes ayant bénéficié d'un séquençage complet de bonne qualité. En parallèle, l'ajout d'espèces permet d'affiner le contexte phylogénétique et d'améliorer la détection des gènes sous sélection positive dans la lignée menant au chien. L'apport d'un génome plus proche du génome canin tel que le génome du chat améliorerait la fiabilité des détections. En contrepartie l'apport d'un génome réduit le nombre d'orthologues 1:1 et implique d'analyser une plus faible proportion du catalogue des gènes codant pour des protéines.

I.2. Limitations de l'analyse de sélection positive par le calcul du dN/dS

L'analyse du ratio dN/dS se focalise sur la séquence codante des gènes conservés et ignore donc les séquences régulatrices de ces gènes, les gènes créés (par duplication, fusion ou néo-création), les gènes perdus, les gènes en relation d'orthologie plus complexe tels que les familles des gènes ainsi que les gènes non-codants qui peuvent être autant de substrats de

l'évolution. Un prolongement de ce travail consisterait à étudier les régions codantes de familles de gènes et de tester la sélection dans des régions non-codantes tel que les miRNA qui sont des cibles de la sélection dans l'espèce humaine (Quach, *et al*., 2009). Cette étude pourra se faire en utilisant par exemple l'outil baseML du package PAML qui permet d'analyser l'évolution des séquences nucléotidiques à partir de l'optimisation par maximum de vraisemblances de modèles de Markov de substitutions nucléotidiques. Un projet d'identification et d'annotation des gènes canins non-codants est en cours au laboratoire. Parmi les perspectives de ce projet, l'étude de la sélection positive sur cette nouvelle catégorie de séquences fonctionnelles sera envisagée.

Le test LRT utilisé dans cette analyse est une méthode de comparaison de modèles qui permet de comparer la pertinence des hypothèses d'absence et de présence de sélection positive dans un gène donné pour une branche donnée. Pour obtenir des résultats pertinents de ce test de comparaison de taux de substitution synonyme et non-synonyme, l'exactitude des séquences et la justesse des alignements sont déterminants. En effet, une erreur dans un alignement multiple de séquence peut fausser le décompte de mutations synonymes et non synonymes pour une partie de la séquence et ainsi induire la détection d'un excès artificiel de gènes en sélection positive, comme cela a été le cas chez le chimpanzé (Mallick, *et al*., 2009). Ainsi, lors d'une analyse initiale nous avons extrait les séquences codantes des gènes en relation 1:1 entre six espèces (chien, Homme, chimpanzé, rat, souris et vache), traduit ces séquences, aligné les traductions avec le programme clustalW et remplacé les acides aminés par les codons issus des séquences de départ. Nous avons obtenu alors un grand nombre de gènes détectés sous sélection positive pour certaines espèces (1330 chez le chien, 996 chez le rat après correction des p valeurs par la méthode de Bonferroni). Les alignements multiples de séquence redéfinis par nos collaborateurs D. Enard et H. Roest Crollius issus de l'intégration des alignements nucléotidiques et protéiques des exons des orthologues apportent une plus grande fiabilité à la détection de la sélection positive. La méthode de détection des sites sous sélection positive au sein des gènes implique une correction de type tests multiples. Les résultats peuvent être considérés comme robustes mais sur-corrigés.

I.3. Les gènes canins sous sélection positive

Après l'identification des gènes sous sélection positive, nous avons recherché quels gènes ont été positivement sélectionnés indépendamment dans la lignée menant au chien et dans les autres lignées. Les co-occurrences des gènes sous sélection positive dans l'espèce canine et au moins une autre espèce ont été dénombrées. De manière intéressante le chien

présente plus de gènes sous sélection positive en commun avec plusieurs autres espèces qu'à l'attendu. Le chien partage des gènes sous sélection positive avec les autres laurasiatheria représentés par le cheval et la vache et qui sont les espèces les plus proches du chien dans notre jeu de données et qui partagent avec le chien le fait d'avoir été domestiqué. Le chien partage aussi des gènes sous sélection positive avec les rongeurs ; le rat et la souris sont, comme le chien, présents dans l'environnement défini par l'Homme dans le monde entier. Par contre le chien ne présente pas plus de gènes sous sélection positive en commun avec les primates (le chimpanzé mis à part) qui sont plus éloignés phylogénétiquement que les autres Laurasiatheria. Le chimpanzé présente une co-occurrence de gènes sous sélection positive avec les Laurasiatheria et les rongeurs du jeu de données; Ceci peut révéler une évolution particulière du chimpanzé ou être un artéfact des analyses comme Mallick et collaborateurs l'ont montré pour 59 gènes qui étaient des faux positifs dans la recherche de gènes sous sélection positive chez le chimpanzé par rapport au génome humain (Mallick, *et al*., 2009). Les analyses fonctionnelles à venir permettront de mettre en évidence les voies métaboliques co-sélectionnées entre le chien et chaque autre espèce ou groupe d'espèces.

I.4. Développement d'un serveur d'analyse des contraintes sélectives des séquences codantes : OMEGA

Il existe plusieurs types d'outils en ligne pour rechercher la sélection positive à l'aide du ratio ω (dN/dS). Des bases de données des gènes positivement sélectionnés sont consultables telles que la base SELECTOME (Proux, *et al*., 2009) ou TAED (Roth, *et al*., 2005) qui contiennent des données de ratio dN/dS pré-calculées. D'autres outils permettent de calculer en ligne le ratio ω pour un gène tel que Selecton (Doron-Faigenboim, *et al*., 2005), PAL2NAL (Suyama, *et al*., 2006) ou le "Ka/Ks Calculation tool" de l'université de Bergen (Norvège). Enfin, l'outil PhyleasProg (Busset, *et al*., 2011) qui permet d'obtenir et de calculer des informations sur l'évolution de gènes mammifères présents dans les bases de données. Le serveur OMEGA développé pendant ma thèse est dédié au calcul des ratios dN/dS et au test LRT de sélection positive. Ces calculs sont effectués à partir de séquences fournies par l'utilisateur et offre la possibilité de lancer ces calculs pour plusieurs centaines de gènes simultanément. OMEGA retourne à l'utilisateur les résultats des ratios dN/dS pour les différents branches de l'arbre phylogénétique des gènes fournis et les informations de site qui accompagnent le résultat du test LRT. La procédure d'alignement étant déterminante dans le calcul du ratio dN/dS, nous avons inclus au serveur OMEGA le méta-aligneur T-coffee qui utilise la combinaison d'alignements réalisés à l'aide d'un algorithme d'alignement local et

d'un algorithme d'alignement global afin d'obtenir des alignements les plus fiables possible. OMEGA bénéficiera de deux améliorations principales : l'utilisation des capacités de calcul parallèle des machines de la plate-forme bioinformatique GenOuest, ce qui permettra d'augmenter la rapidité de calcul des tests en répartissant les différents jeux de données des utilisateurs ou les calculs des différents modèles sur différents processeurs et l'ajout d'une option permettant de choisir entre la procédure d'alignement actuelle et l'utilisation un outil d'alignement qui tient compte des décalages du cadre de lecture (qui sont alors considérés comme des erreurs sans signification biologique) tel que MACSE (Ranwez, *et al*., 2011).

II. Différenciation génétique entre races canines

II.1. La puce CanineHD

La recherche pan-génomique des régions de différenciation génétique entre 30 races canines s'est effectuée à partir de l'analyse du polymorphisme de 170.000 SNPs. Les SNPs sont espacés de 13 kb en moyenne avec 133 régions inter SNPs de plus de 100 kb dont 17 de plus de 200 kb. Une seule vaste région du génome canin n'est pas couverte par les SNPs, correspondant à un intervalle de plus de 600 kb entre les marqueurs en position 48.498.895 et 51.734.271 sur le chromosome X. La séquence de cette zone contient une région non séquencée : chrX:48668193-51668192 qui correspond au centromère du chromosome X, le seul chromosome canin non acrocentrique. Le jeu de données que nous avons utilisé couvre donc le génome canin avec la plus grande densité disponible à ce jour. L'étendue du déséquilibre de liaison dans une race de chien étant de 100 à 1000 kb, les données analysées reflètent bien le polymorphisme du génome canin. Quelques régions sont non ou très faiblement polymorphes dans l'espèce canine et ne sont donc pas informatives pour détecter la différenciation génétique des races. Cependant l'ensemble des locus des régions jamais différenciées entre les races sera analysé afin de déterminer quelles sont les régions du génome canin qui se différencient du loup pour lequel un jeu de données de génotypage est en cours de production. La perspective d'identifier des régions de différenciation entre le loup et le chien permettra d'établir un catalogue des locus candidats ayant été la cible de la sélection lors de la période de domestication du chien.

II.2. Recherche de régions de différenciation génétique : la méthode Fst-di

II.2.1. Impact de l'échantillonnage

Le jeu de données utilisé pour la recherche de régions de différenciation est constitué de génotypes de SNPs sur 456 chiens de 30 races différentes. Ces races sont représentées par 10 à 26 individus. L'effectif minimum de 10 individus par race que nous nous sommes imposés permet de disposer des données issues de 20 chromosomes au moins et constitue un seuil arbitraire. Initialement il s'agissait d'établir un équilibre entre la volonté d'établir des cohortes de plus grands effectifs possibles et la volonté de travailler à partir d'un jeu de données représentant le plus grand nombre de races possibles. Ce choix de 10 individus est appuyé par deux éléments. Premièrement, nous disposions dans le jeu de données complet de cinq races représentées par moins de 10 individus (5 cavalier king charles, 6 terriers, 7 dalmatiens, 8 boxer et 8 bull terrier anglais). Pour chacune de ces races un jeu de génotypes simulé a été calculé par Erik Axelsson qui a constaté que les déséquilibres de liaison entre les marqueurs simulés pour ces races étaient beaucoup moins proches des déséquilibres de liaison réels que pour les races de plus fort effectif. Deuxièmement, lors du stage de Master2 "Modélisation des systèmes biologiques" au laboratoire, Kevin Druet a évalué l'impact de l'échantillonnage d'un plus faible nombre de chromosomes sur la détection des régions de différenciation. La race "braque de weimar" qui est représentée par 26 individus a été ré-échantillonnée en deux populations de 13 individus (paire de ré-échantillonnage). Cette analyse a été conduite 10 fois. Pour chaque paire de ré-échantillonnage, nous avons analysé et détecté les régions de différenciation avec le pipeline de calcul de d_i. Près de 70% des fenêtres de différenciation sont détectées en commun à partir des 2 ré-échantillonnages. La même analyse effectuée sur un échantillonnage de 6 et 8 individus permet de n'identifier que 45% et 56% respectivement de fenêtres communes. Dans le cadre du consortium LUPA, près de 10.000 chiens ont été génotypés pour différents projets de recherche des bases génétiques de maladies. L'utilisation de ces données permettra de disposer de génotypes pour plus de 80 races comportant 100 individus en moyenne. Ces effectifs vont permettre de mener des analyses statistiques plus robustes et de préciser le catalogue de régions de différenciation.

II.2.2. Crible du génome par fenêtres

La méthode de recherche de régions de différenciation repose sur le fait qu'en l'absence de pression évolutive, les valeurs de Fst et de d_i de deux marqueurs consécutifs, varient de manière aléatoire. En plus de la variabilité réelle du Fst entre deux marqueurs, un SNP peut

avoir une valeur de Fst altérée par un problème de génotypage. Pour limiter ces artefacts, il est important de prendre en compte simultanément les valeurs de plusieurs marqueurs consécutifs. Nous avons choisis d'utiliser la moyenne des valeurs de d_i contenues dans des fenêtres d'environ 10 SNPs. En effet si une fenêtre n'est pas la cible d'une sélection, la moyenne des valeurs de d_i des SNPs présents dans la fenêtre va être faible, la valeur moyenne va niveler les variations individuelles. En revanche, si une forte différence moyenne de polymorphisme entre races est observée, cela indique que les valeurs de Fst et de d_i de tous les SNPs de cette fenêtre sont élevées. Le choix de cribler le génome par une méthode de fenêtres glissantes et chevauchantes permet de détecter des signaux forts et résolutifs car de petites tailles. L'utilisation de fenêtres chevauchantes permet de définir de manière plus précise les limites des régions potentiellement cibles de la sélection, de définir le centre de ces régions et de faciliter la sélection des signaux en maximisant les valeurs de d_i obtenus pour la meilleure fenêtre de chaque signal.

L'utilisation de l'indice d_i permet d'obtenir des résultats reproductibles entre jeux de données et laboratoires indépendants puisque nous retrouvons 150 des 155 régions de différenciations de races retrouvées par le groupe de J. Akey (Akey J. M., *et al*., 2010) qui dans son étude a utilisé 10 races génotypées avec 20.000 SNPs et une approche de fenêtres adjacentes de 1 Mb. Le d_i est calculé par normalisation des comparaisons par paire et l'addition des comparaisons concernant une race donnée. Une valeur élevée de d_i d'une fenêtre indique l'importance de cette fenêtre dans la définition de la race, mais ne garantit pas une valeur absolue de différenciation de races. Ainsi une région sélectionnée de la même manière dans deux races peut avoir des valeurs de d_i différentes selon la présence d'autres sélections importantes pour l'une des deux races. Une alternative est de formaliser un nouveau calcul des valeurs de Fst qui prenne en compte l'ensemble des marqueurs de la fenêtre analysée. Les valeurs du Fst calculées pour une fenêtre ne correspondent pas à la moyenne des valeurs de Fst des SNPs contenus dans cette fenêtre ou région. Une analyse préliminaire nous a permis d'observer un plus fort contraste entre les valeurs élevées et faibles de Fst que par l'utilisation de la moyenne des Fst des différents SNPs.

II.2.3. Du génotype à la séquence

À plus long terme, les nouvelles technologies de séquençage (NGS) permettront de réaliser des analyses qui s'appuieront sur la séquence pour identifier les régions de différenciation des races comme cela a déjà été initié pour d'autres espèces domestiquées

comme le poulet (Rubin, *et al*., 2010). Par exemple, le séquençage d'un individu chez un grand nombre de races pourra permettre de dégager les régions communes par groupe de races (pointers, retrievers, terriers etc.). Le séquençage de plusieurs individus de quelques races pourra permettre de rechercher finement les régions spécifiques de différenciation entre races. Enfin le séquençage des exomes de nombreux individus dans un grand nombre de races permettra de capturer la totalité de la fraction codant pour des protéines du génome, d'établir les catégories des polymorphismes qui sont sélectivement neutres, délétères ou avantageux, et d'identifier directement les gènes ciblés par la sélection. L'absence de données non-codantes limite l'application de l'approche 'exome' (Tennessen, *et al*., 2011).

II.3. Étude des régions identifiées

II.3.1. Sélection statistique

En parallèle de l'analyse de différenciation de race par approche Fst-di, nos collaborateurs de l'université d'Uppsala ont réalisé une analyse similaire basée sur la perte d'hétérozygotie. L'indice Si utilisé correspond au niveau d'hétérozygotie d'une race en relation avec le niveau d'hétérozygotie des autres races. Nous avons réalisé les mêmes analyses sur les données simulées afin d'obtenir des distributions des variables aléatoires sous l'hypothèse H_0 de non sélection. En terme de recherche de signatures génomiques, la principale variable en mesure de différencier un jeu de données comportant des régions de sélection et un jeu de données non-sélectionné est la taille des régions obtenues. Nous avons défini les p valeurs des régions de chaque race en utilisant la distribution des tailles des régions simulées comme distribution des tailles des régions sous l'hypothèse H_0 de non sélection. Nous obtenons 530 régions Fst-di de différenciation de races et 990 régions de perte d'hétérozyosité pour lesquelles la taille est significativement plus grande que dans le jeu de données simulées. Lorsque nous appliquons la correction pour test multiple Benjamini-Hochberg (BH), aucune région de différenciation Fst-di n'a une taille significativement assez grande pour passer la correction BH et 350 régions de perte d'hétérozygotie restent significativement plus grande qu'à l'attendu.

Le critère de comparaison de taille de signal est bien adapté à une statistique basée sur la perte d'hétérozyosité car cette mesure révèle principalement des signaux liés à des sélections fortes de variants rares ou récents, chacun lié à un ou peu d'haplotypes. Les régions obtenues après sélection peuvent regagner lentement en polymorphisme par mutations mais présentent trop peu de polymorphismes pour regagner rapidement en diversité par recombinaison. Ceci est accentué par le niveau des déséquilibres de liaison dans les races

canines qui favorisent la transmission de grands blocs. En revanche les régions de différenciation Fst-di révèlent d'autres types de signaux pour lesquels la sélection et les pratiques d'élevage ont changé le polymorphisme. Ces régions sont différentes car le changement de nature de fréquence allélique n'entraîne pas nécessairement une perte significative d'hétérozygotie. En effet un variant sélectionné assez ancien pour être présent dans plusieurs haplotypes va faire augmenter la fréquence de plusieurs haplotypes. Ces haplotypes continuent de présenter une diversité génétique propre à rétablir le niveau de polymorphisme perdu par la sélection. Les marqueurs dont le polymorphisme affecté est détectable en recherche de différenciation sont donc plus proche du variant d'intérêt que pour les régions de perte d'hétérozygotie. Les tailles des régions de différenciation seront donc proches des tailles des régions issues de la dérive génétique et donc des données simulées. Ceci explique que les p valeurs basées sur la taille des régions soient moins significatives pour les régions de différenciation que pour les régions de perte d'hétérozygotie et ainsi perdent leur significativité après correction pour les tests multiples. De plus, les distributions théoriques des tailles des régions Si et d_i ont été définies à partir d'une simulation unique des génotypes. Par conséquent le nombre de régions issues des données simulées est comparable au nombre de régions issues des données réelles. Les p valeurs les plus significatives sont au mieux de 1/nombre de régions, le nombre moyen de régions simulées en d_i pour une race étant de 287 et le nombre moyen de régions issues des données expérimentales en d_i étant de 213, les p valeurs les plus significatives sont de l'ordre de 1/287 = 0,00348 et la base de la correction est l'effectif de 213. La procédure de correction BH est construite de telle sorte que pour obtenir des p valeurs corrigées significatives dans ces conditions, plusieurs p valeurs doivent être égales à la plus faible p valeur, c'est-à-dire que plusieurs régions issues du jeu de données doivent êtres plus grandes que toutes les régions issues du jeu de données simulées. À partir d'un seul jeu de données simulées, les procédures de corrections des p valeurs pour tenir compte de la réalisation des tests multiples représentent une sur-correction des données que ce soit pour les régions de différenciation ou les régions de perte d'hétérozygotie.

II.3.2. Intégration des résultats di et Si

Les deux indices utilisés dans cette étude identifient des régions différentes car les signaux détectés sont différents. En effet, l'indice d_i détecte un changement de fréquence des allèles d'une région, par exemple si un phénotype variable dans une population devient la cible d'une sélection visant à l'accentuer. L'indice d_i est indiqué pour la détection des

balayages sélectifs modérés mais perdra en puissance si la sélection porte sur un nouveau variant apparu dans un haplotype déjà majoritaire. L'indice Si détecte les signaux de balayage sélectifs forts liés à une sélection rapide d'une mutation présente dans un seul haplotype qui devient fixé. Cependant les sélections les plus fortes doivent avoir laissé des traces dans le génome détectables par les deux méthodes. Nous avons considéré que les régions détectées par les deux méthodes sont les meilleures régions candidates pour être la cible de la sélection artificielle. Ces régions pourront être étudiées plus en détail pour déterminer la présence ou l'absence de sélection comme dans le cas des phénotypes et gènes étudiés en génétique humaine (Bersaglieri, *et al*., 2004; Patin, *et al*., 2006; Barreiro, *et al*., 2008). Les régions détectées par seulement un indice restent candidates pour la détection de la sélection artificielle et permettent de capturer un maximum d'événement.

II.3.3. Recherche enrichissements fonctionnels

II.3.3.1. Gene Ontology

Parmi les régions candidates à la sélection artificielle identifiées, 2191 contiennent en moyenne 4,5 gènes (1 à 69 gènes). Une région peut comporter un ou plusieurs gènes soumis à la sélection artificielle voire un groupe de gènes participant à un même phénotype. Mais ces régions contiennent aussi des gènes ayant une valeur sélective neutre et il n'existe pas de moyen trivial pour déterminer quel gène de cette région est candidat pour être impliqué dans un phénotype sélectionné. Ainsi nous avons restreint l'analyse des catégories fonctionnelles des gènes aux seules régions qui contiennent un seul gène.

Nous avons observé un enrichissement des catégories GO (gene ontology) liés au développement, fonctions qui peuvent être liées aux différences morphologiques entre races tel que les différences de taille, forme de la queue, la présence des plis (shar-pei), la courbure des oreilles, etc. L'enrichissement des voies de la pigmentation corrèle avec la sélection des chiens sur les couleurs de leur robe. L'enrichissement de catégories liées au développement du système nerveux peuvent être mis en parallèle avec les aptitudes cognitives et comportementales sélectionnées chez le chien telles que le comportement stéréotypé des races de type pointer qui se figent pour indiquer une direction ou les aptitudes d'apprentissage des chiens de travail. Ces aptitudes font du chien un bon modèle pour étudier les mécanismes d'apprentissage et les gènes présents dans les régions candidates à la sélection artificielle sont un point d'entrée pour l'étude des mécanismes liés à la cognition. En revanche nous observons l'absence d'enrichissements de régions de différenciation associées au système olfactif alors que les capacités olfactives différencient les races (Tacher, *et al*., 2005; Robin, *et al*., 2009). L'absence de ce type de catégorie fonctionnelle peut

indiquer que les différences de capacités olfactives entre races s'expliquent par d'autres facteurs que les récepteurs olfactifs tels que les neurones du bulbe olfactif et les capacités d'apprentissage propre à certaines races.

II.3.3.2. GO : sélection naturelle vs sélection artificielle

Les catégories fonctionnelles GO identifiées ne correspondent pas aux catégories retrouvées classiquement dans les recherches de sélection positive entre espèces basées sur les gènes codant pour des protéines. Ces résultats suggèrent que la sélection artificielle des polymorphismes présents dans l'espèce canine a pu se réaliser sur des gènes et des fonctions en sélection négative. Ces différences fonctionnelles entre sélection naturelle et artificielle peuvent aussi s'expliquer par les différentes échelles de temps considérées. Les sélections sur le polymorphisme SNP sont des sélections récentes alors que les sélections sur les gènes orthologues doivent agir sur une plus grande échelle de temps. Une autre explication est que nous détectons à l'aide du polymorphisme des SNP analysés sur l'ensemble du génome les régions régulatrices des gènes mais pas les régions codantes qui sont sélectionnées négativement ou pas encore affectées. Une perspective de ce travail est de mener une analyse fonctionnelle des régions qui ne contiennent pas de gènes annotés. Les variants sélectionnés peuvent correspondre à des éléments régulateurs tels que les régions promotrices, des sites de liaisons aux facteurs de transcription ou des ARN non-codant pour des protéines. Par ailleurs, les signatures de sélection dépourvues de gènes connus identifient des régions génomiques pour lesquelles un effort de réannotation du génome canin doit être considéré (Derrien, *et al*., 2011).

II.3.3.3. Corrélation avec les termes OMIM

Chaque race canine présente une ou plusieurs maladies génétiques spécifiques. L'hypothèse d'une corrélation entre la création des races et la prépondérance de maladies génétiques doit se refléter par le nombre de gènes responsables de maladies présents dans les régions de différenciations détectées. Nous avons testé pour la présence d'enrichissement en gène possédant un terme OMIM parmi les gènes des régions détectées. Le test du χ^2 n'a pas permis de rejeter l'hypothèse H_0 de proportion similaire de gènes connus pour être impliqués dans des maladies génétiques. L'absence de différence en annotation OMIM des gènes présents dans les régions Si et d_i suggère que la plupart des sélections qui ont conduit à la création des races sont des sélections associées à des critères non pathologiques. Cependant l'observation de la co-ségrégation de sélection avec des maladies ou des prédispositions aux

maladies génétiques reste vraie. Cette sélection s'observe par association entre un trait sélectionné et un trait pathologique comme le phénotype peau plissée du shar-pei qui est associée à une duplication (CNV) de 16.1 kb sur le chr13 associée à une fièvre spécifique de cette race (Olsson, *et al*., 2011). Les auteurs de cette étude suggère que la duplication contient un ou plusieurs éléments régulateurs qui altèrent la régulation du gène voisin HAS2 impliqué la synthèse de l'acide hyaluronique, un composant majeur de la peau. La sélection de certaines couleurs de robe est aussi liée à des maladies par exemple la couleur de la robe 'merle' qui est caractérisée par des taches claires sur des robes sombres est souvent associée à des troubles de la vue et de l'audition (Clark, *et al*., 2006; Hédan, *et al*., 2006). La sélection de maladies ou de prédisposition lors de la sélection des races est illustrée par l'exemple de la sélection de races de petite taille due aux membres courts mais chondrodysplasique chez le teckel, ou la sélection du phénotype "croupe basse" chez les bergers allemands qui induit une dysplasie de la hanche.

II.3.4. Utilisation de la méthode Fst-di en support des études d'associations génétiques

Les régions d_i, ont une taille d'environ 200 kb, et sont beaucoup plus précises qu'un locus identifié lors d'une d'étude d'association qui est typiquement de quelques Mb. La possibilité d'analyser les groupes de chiens cas et contrôles d'une étude d'association comme deux populations qui ne se différencieraient que sur le locus associé à la maladie est une perspective de ce travail. Lors de son stage de M2 au laboratoire, Kevin Druet a testé cette hypothèse sur des données de 48K SNP d'une population de Golden retriever pour laquelle notre équipe a identifié la mutation causale du gène responsable d'une genodermatose - l'ichtyose- chez le chien et chez l'Homme (Grall, *et al*., 2011). L'analyse d'association menée lors de cette étude a identifié un locus de 5 Mb alors que l'approche Fst-di a identifié un locus de 200 kb qui contient le gène porteur de la mutation causale identifié par le laboratoire. Nous utiliserons le pipeline Fst-di en complément des études d'associations menées dans notre laboratoire.

III. Intégration des signatures des sélections naturelle et artificielle

Nous avons recherché indépendamment les signatures que la sélection naturelle a imprimé sur le génome canin au cours des 89 millions d'années d'évolution de la branche

carnivore et les signatures liées à la sélection artificielle et aux pratiques d'élevages des races canines de ces derniers siècles. Ces deux sélections ont agit sur des échelles de temps totalement différentes. De plus, les signatures de la sélection naturelle ont été recherchées sur la séquence codante d'environ la moitié des gènes annotés (n=10.730) chez le chien, alors que les signatures de la sélection artificielle ont été définies à partir d'un jeu de données de 170.000 SNP qui interroge la quasi totalité du génome (>90%). Pour comparer de manière exhaustive les signatures des sélections naturelles et artificielles, nous devons i) obtenir des données de sélection naturelle recherchées sur l'ensemble du génome ; ii) définir une méthode pour comparer ces régions et détecter les co-occurrences ; iii) tester la significativité des co-occurrences.

Pour obtenir les données de balayages sélectifs de sélection naturelle chez le chien, nous avons poursuivi notre collaboration avec le Dr. H. Roest Crollius. En effet, en 2010 cette équipe (Enard*, et al*., 2010) a défini une méthode de détection des hot-spots de sélection naturelle partagées entre espèces à partir de la séquence génomique d'un seul individu. Le principe de cette recherche de signature de sélection naturelle est de détecter les régions de perte de niveau d'hétérozygotie par rapport au niveau de divergence avec une espèce proche. Ces zones de perte de polymorphisme peuvent-être dues à un balayage sélectif, une dérive génétique ou à une particularité de l'individu séquencé. Si une région génomique présente une perte d'hétérozygotie dans une espèce ainsi que dans les régions orthologues des autres espèces, cette région sera alors considérée comme soumise à un balayage sélectif récurrent entre différentes espèces.

Pour comparer les régions de différenciation des races canines avec les régions de perte de polymorphisme définies entre espèces, nous avons utilisé l'ensemble des fenêtres de différenciation que nous avons classées par valeur de d_i. À partir de ce classement nous avons défini 16 jeux de régions, le premier est issu de la sélection des 50 fenêtres de plus fort d_i pour chaque race. Le second jeu de régions est issu de la sélection des 100 fenêtres de plus fort d_i pour chaque race et ainsi de suite jusqu'au 16^e jeu qui correspond à la totalité des fenêtres identifiées. Chacun de ces 16 jeux est alors comparé indépendamment aux régions de sélection naturelle identifié en co-occurrence par le laboratoire DYOGEN.

Pour analyser la co-occurrence entre les régions de sélection naturelle et artificielle, le test développé par David Enard consiste à diviser le génome en 20 intervalles dont les bornes sont en dehors des régions sous sélection qui sont permutées aléatoirement. Lors d'une permutation, les régions de sélection naturelle sont redistribuées aléatoirement. La co-occurrence entre les gènes présents dans les régions de sélection artificielle et les régions de

sélection naturelle est alors due au hasard. En réalisant 100.000 permutations, il est alors possible de définir une p valeur marginale qui correspond à la proportion de co-occurrence aléatoire supérieure à la co-occurrence réelle. Des résultats préliminaires indiquent une co-occurrence entre la sélection naturelle et artificielle. Cette co-occurrence est d'autant plus significative que les jeux de données de sélection artificielle sont les plus forts et avec un effectif de régions suffisant pour conserver la puissance statistique (jeux n°3-5). La co-occurrence augmente également lorsque l'on considère uniquement les régions partagées par plusieurs races et lorsque l'on considère les gènes localisés au centre des régions qui les contiennent. Les calculs complets des tests de co-occurrence sont en cours de réalisation au laboratoire DYOGEN.

Les résultats préliminaires de co-occurrence entre la sélection naturelle et la sélection artificielle pose la question de la signification des locus impliqués dans les événements de sélection naturelle et artificielle. La question est d'importance à la fois pour l'étude du génome canin et pour l'étude de l'évolution. En effet les phénotypes canins sélectionnés peuvent être considérés comme une intensification rapide des variations présentes dans les populations naturelles (taille, pelage, tempérament, etc.), soit comme des phénotypes 'innovants' mais 'mal-adaptés' dans une population naturelle (petite taille des membres par rapport au corps, brachycéphalie, prédisposition à certaines maladies, etc.). Si un gène a été ciblé par la sélection naturelle et par la sélection artificielle de manière indépendante, il devient moins probable que ce gène contribue à un phénotype innovant et spécifique de la lignée canine. D'un point de vue évolutif, le catalogue des gènes ciblés par les deux types de sélection permet d'établir le patron des hot-spots des locus cibles de la sélection.

D'autre part, les locus canins affectés par la sélection artificielle mais non ciblés par la sélection naturelle qui a opéré pendant plus de 90 millions d'années chez 10 espèces de mammifères, ont une meilleure probabilité d'avoir contribué à un phénotype artificiel fixé dans une ou quelques races canines. Ainsi, ces locus deviennent de bons candidats à considérer pour détecter des événements de la sélection qui n'auraient pas laissé de signatures génétiques pérennes au cours de l'évolution des espèces. Nous pouvons spéculer que ces locus sont de bons candidats pour identifier les causes génétiques qui contribuent aux phénotypes présents uniquement sous sélection artificielle et qui, à priori, ne procurent pas un avantage sélectif. Cette catégorie de phénotypes est illustrée par exemple par une morphologie particulière comme la chondrodysplasie causée par la rétrotransposition du gène FGF4, les plis du Shar-Pei liés au gène HAS2 ou une aptitude particulière comme la capacité de communication Homme-chien que décrit Brian Hare et Michael Tomasello dans "Human-

like social skills in dogs" (Hare and Tomasello, 2005). Selon ces auteurs, le chien est conscient de ce que voit l'Homme. En effet, le chien comprend le geste du doigt pointé vers un objet ou la signification d'un regard ou d'un mouvement de la tête. Le chien déchiffre mieux que le chimpanzé la communication humaine, il égale dans ce domaine la capacité du jeune enfant. L'Homme, au cours du processus de domestication aurait sélectionné chez le chien une aptitude à la communication humaine qui serait fixée dans l'espèce canine. Ce type d'aptitude est inconnue dans d'autres espèces. Le catalogue des locus spécifiquement détectés sous sélection artificielle pave la voie de la recherche des causes génétiques qui expliquent ou contribuent à ces aptitudes.

Conclusion générale

Mon travail de thèse a permis l'identification des gènes codant pour des protéines sous sélection positive dans la lignée carnivore et l'établissement d'un catalogue des régions génomiques potentiellement ciblées par la sélection artificielle lors de la création des races canines. Ces deux parties de mon travail se sont déroulées dans le cadre de deux collaborations. La première collaboration au niveau national avec l'équipe du Dr. Hugues Roest Crollius (équipe DYOGEN, ENS Paris) nous a permis d'analyser les événements de sélection naturelle survenus dans la lignée canine en comparaison des événements de sélection naturelle survenus dans les lignées des autres espèces Nous avons établi le catalogue de l'ensemble des gènes sous sélection positive pour 10 espèces. L'analyse de la co-occurrence des gènes positivement sélectionnés montre que la sélection positive agit indépendamment sur des gènes communs entre le chien, les Laurasiatheria et les rongeurs analysés mais pas sur des gènes communs avec les primates. Au cours de ce travail, nous avons développé un outil en ligne, OMEGA , pour automatiser le calcul des tests de sélection positive sur plusieurs jeux de données. La seconde collaboration au niveau international avec l'équipe du Dr Matthew Webster (Université d'Uppsala en Suède) dans le cadre du consortium européen LUPA nous a permis d'établir un catalogue de régions potentiellement cibles de la sélection artificielle ayant permis la création des races et présentant un enrichissement fonctionnel différent de ce qui est attendu lorsque l'on s'intéresse à la sélection naturelle. Au cours de ce travail, nous avons développé un pipeline d'analyse de détection des régions de différenciation allèlique entre populations qui peut être utilisé pour de nombreuses études disposant de génotypages de SNP et qui peut fournir un support aux études d'associations telles que celles réalisées dans notre équipe. Le travail sur le polymorphisme des races de chiens se poursuit par le développement et l'utilisation d'un outil dérivé du premier pipeline et l'utilisation des jeux de données LUPA ouvre plusieurs perspectives. Deux collaborations sont déjà initiées, l'une avec le Dr. Jacques Nicolas de l'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes) pour rechercher une approche combinatoire exacte en définissant avec précision les SNPs qui caractérisent le mieux une race donnée. L'autre avec le Dr. Mathieu Emily (Département de statistiques, Université Rennes2), qui développe une approche statistique sur la recherche de combinaisons de régions d_i pour lesquelles des associations d'haplotypes entre régions permettent de définir les races. Dans notre équipe une étude basée sur la comparaison des régions de plus faible d_i et des polymorphismes que le loup partage avec le chien est engagée pour déterminer les régions dont le polymorphisme est commun à toute la population canine

mais se différencie du loup révélant ainsi des signatures génétiques pouvant être liées au processus de domestication.

La collaboration avec le Dr. Hugues Roest Crollius se poursuit actuellement pour établir les co-occurrences entre la sélection artificielle et la sélection naturelle afin de déterminer si il existe des régions du génome qui sont constamment affectées par la sélection.

La structure de l'espèce canine séparée en races isolées d'un point de vue reproductif, subissant des sélections aboutissant à une grande diversité de morphologie, d'aptitude et de comportement évoque la structure des mammifères, issue de 165 millions d'années d'évolution, organisée en espèces isolées d'un point de vue reproductif, subissant des sélections différentes et présentant une grande diversité de morphologie, d'aptitude et de comportement. Les caractéristiques comparées des signatures génétiques des races créées par la sélection artificielle et des signatures de la sélection naturelle nous conduira à poser la question : l'espèce canine peut-elle être considérée comme une simulation réduite mais accélérée de la radiation des mammifères ?

Bibliographie

1 Abadie, J., Hedan, B., Cadieu, E., De Brito, C., Devauchelle, P., Bourgain, C.*, et al.* (2009) Epidemiology, pathology, and genetics of histiocytic sarcoma in the Bernese mountain dog breed, *J Hered*, **100 Suppl 1**, S19-27.

2 Abitbol, M., Thibaud, J.-L., Olby, N.J., Hitte, C., Puech, J.-P., Maurer, M.*, et al.* (2010) A canine Arylsulfatase G (ARSG) mutation leading to a sulfatase deficiency is associated with neuronal ceroid lipofuscinosis. *Proc Natl Acad Sci USA*. pp. 14775-14780.

3 Akey, J.M., Zhang, G., Zhang, K., Jin, L. and Shriver, M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection, *Genome Research*, **12**, 1805-1814.

4 Akey, J.M., Ruhe, A.L., Akey, D.T., Wong, A.K., Connelly, C.F., Madeoy, J.*, et al.* (2010) Tracking footprints of artificial selection in the dog genome, *Proc Natl Acad Sci U S A*, **107**, 1160-1165.

5 Anderson, T.M., vonHoldt, B.M., Candille, S.I., Musiani, M., Greco, C., Stahler, D.R.*, et al.* (2009) Molecular and evolutionary history of melanism in North American gray wolves, *Science*, **323**, 1339-1343.

6 Auton, A. and McVean, G. (2007) Recombination rate estimation in the presence of hotspots. *Genome Research*. pp. 1219-1227.

7 Avery, O.T., Macleod, C.M. and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III, *J Exp Med*, **79**, 137-158.

8 Awano, T., Johnson, G.S., Wade, C.M., Katz, M.L., Johnson, G.C., Taylor, J.F.*, et al.* (2009) Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis, *Proc Natl Acad Sci U S A*, **106**, 2794-2799.

9 Axelsson, E., Webster, M.T., Ratnakumar, A., Consortium, L., Ponting, C.P. and Lindblad-Toh, K. (2011) Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Research*.

10 Bamshad, M. and Wooding, S.P. (2003) Signatures of natural selection in the human genome, *Nat Rev Genet*, **4**, 99-111.

11 Barreiro, L., Laval, G., Quach, H., Patin, E. and Quintana …, L. (2008) Natural selection has driven population differentiation in modern humans, *Nature genetics*.

12 Beaumont, M.A. and Balding, D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans, *Mol Ecol*, **13**, 969-980.

13 Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A practical and powerfil approach to multiple testing, *Journal of the Royal Statistical Society*, **57**, 289-300.

14 Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A.*, et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. pp. 1111-1120.

15 Berthouly, C., Leroy, G., Van, T.N., Thanh, H.H., Bed'Hom, B., Nguyen, B.T.*, et al.* (2009) Genetic analysis of local Vietnamese chickens provides evidence of gene flow from wild to domestic populations. *BMC Genet*. pp. 1.

16 Bjornerfeldt, S., Webster, M.T. and Vila, C. (2006) Relaxation of selective constraint on dog mitochondrial DNA following domestication, *Genome Res*, **16**, 990-994.

17 Bodey, A.R. and Michell, A.R. (1996) Epidemiological study of blood pressure in domestic dogs. *J Small Anim Pract*. pp. 116-125.

18 Borowsky, R. (2008) Restoring sight in blind cavefish, *Curr Biol*, **18**, R23-24.

19 Boyko, A.R., Quignon, P., Li, L., Schoenebeck, J.J., Degenhardt, J.D., Lohmueller, K.E*., et al.* (2010) A simple genetic architecture underlies morphological variation in dogs, *PLoS Biol*, **8**, e1000451.
20 Breen, M., Bullerdiek, J. and Langford, C.F. (1999) The DAPI banded karyotype of the domestic dog (Canis familiaris) generated using chromosome-specific paint probes, *Chromosome Res*, **7**, 401-406.
21 Breen, M., Jouquand, S., Renier, C., Mellersh, C.S., Hitte, C., Holmes, N.G*., et al.* (2001) Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes, *Genome Res*, **11**, 1784-1795.
22 Breen, M., Hitte, C., Lorentzen, T.D., Thomas, R., Cadieu, E., Sabacan, L*., et al.* (2004) An integrated 4249 marker FISH/RH map of the canine genome, *BMC Genomics*, **5**, 65.
23 Bright, J.M. and Dentino, M. (2002) Indirect arterial blood pressure measurement in nonsedated Irish wolfhounds: reference values for the breed. *J Am Anim Hosp Assoc*. pp. 521-526.
24 Bull, J. and Wichman, H. (2001) Applied evolution, *Annu Rev Ecol Syst* **32**, 183–217.
25 Busset, J., Cabau, C., Meslin, C. and Pascal, G. (2011) PhyleasProg: a user-oriented web server for wide evolutionary analyses. *Nucleic Acids Research*. pp. W479-485.
26 Cadieu, E., Neff, M.W., Quignon, P., Walsh, K., Chase, K., Parker, H.G*., et al.* (2009) Coat variation in the domestic dog is governed by variants in three genes, *Science*, **326**, 150-153.
27 Chakraborty, R. and Jin, L. (1993) A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In Pena, S*., et al.* (eds), *DNA fingerprinting, current state of the science.* Birkhauser, Basel, pp. 153–175.
28 Chen, G.K., Marjoram, P. and Wall, J.D. (2009) Fast and flexible simulation of DNA sequence data. *Genome Research*. pp. 136-142.
29 Clark, L.A., Wahl, J.M., Rees, C.A. and Murphy, K.E. (2006) Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog, *Proc Natl Acad Sci U S A*, **103**, 1376-1381.
30 Clutton-Brock, J. (1995) Origins of the dog: domestication and early history. In Serpell, J. (ed), *The domestic dog: its evolution, behaviour and interactions with people*. Cambridge University Press, NY.
31 Cremers, F.P., van den Hurk, J.A. and den Hollander, A.I. (2002) Molecular genetics of Leber congenital amaurosis, *Hum Mol Genet*, **11**, 1169-1176.
32 Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life*. J. Murray, London.
33 Darwin, C. (1860) *A Naturalist's Voyage Round the World; The Voyage Of The Beagle*. MURRAY, London.
34 Darwin, C. (1868) *The variations of animals and plants under domestication*. John Murray, London.
35 De Meyer, S.F., Vanhoorelbeke, K., Chuah, M.K., Pareyn, I., Gillijns, V., Hebbel, R.P*., et al.* (2006) Phenotypic correction of von Willebrand disease type 3 blood-derived endothelial cells with lentiviral vectors expressing von Willebrand factor, *Blood*, **107**, 4728-4736.
36 De Vries, H. (1902) The origin of species by mutation, *Science*, **15**, 721-729.
37 Derrien, T., Andre, C., Galibert, F. and Hitte, C. (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps, *Bioinformatics (Oxford, England)*, **23**, 498-499.

38 Derrien, T., Theze, J., Vaysse, A., Andre, C., Ostrander, E.A., Galibert, F. and Hitte, C. (2009) Revisiting the missing protein-coding gene catalog of the domestic dog, *BMC Genomics*, **10**, 62.

39 Derrien, T., Vaysse, A., André, C. and Hitte, C. (2011) Annotation of the Domestic Dog Genome Sequence: Finding the missing genes, *Mamm Genome*.

40 Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E. and Pupko, T. (2005) Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics (Oxford, England)*. pp. 2101-2103.

41 Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*. pp. 3439-3440.

42 Enard, D., Depaulis, F. and Roest Crollius, H. (2010) Human and non-human primate genomes share hotspots of positive selection, *PLoS Genet*, **6**, e1000840.

43 Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection, *Genetics*, **155**, 1405–1413.

44 Fisher, R. (1930) *The Genetical Theory of Natural Selection (1999)*. Variorum Edition. Oxford Univ. Press, Oxford.

45 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool*. pp. 99-113.

46 Fitch, W.M. (1995) Uses for evolutionary trees. *Philos Trans R Soc Lond, B, Biol Sci*. pp. 93-102.

47 Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R.*, et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science (New York, N.Y*, **269**, 496-512.

48 Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y.*, et al.* (2011) Ensembl 2011, *Nucleic Acids Research*, **39**, D800-D806.

49 Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S.*, et al.* (2009) The genome response to artificial selection: a case study in dairy cattle, *PLoS ONE*, **4**, e6595.

50 Fondon, J.W., 3rd and Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution, *Proc Natl Acad Sci U S A*, **101**, 18058-18063.

51 Franklin, R.E. and Gosling, R.G. (1953) Molecular configuration in sodium thymonucleate, *Nature*, **171**, 740-741.

52 Fu, Y.-X. and W.H., L. (1993) Statistical tests of neutrality of mutations, *Genetics*, **133**, 693-709.

53 Fu, Y.-X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection, *Genetics*, **147**, 915–925.

54 Galibert, F., André, C. and Hitte, C. (2004) [Dog as a mammalian genetic model], *Med Sci (Paris)*, **20**, 761-766.

55 Galibert, F. and André, C. (2008) The dog: A powerful model for studying genotype-phenotype relationships. *Comp Biochem Physiol Part D Genomics Proteomics*. pp. 67-77.

56 Galis, F., Van der Sluijs, I., Van Dooren, T.J.M., Metz, J.A.J. and Nussbaumer, M. (2007) Do large dogs die young? , *J Exp Zool B Mol Dev Evol*. pp. 119-126.

57 Gardner, M., Williamson, S., Casals, F., Bosch, E., Navarro, A., Calafell, F.*, et al.* (2007) Extreme individual marker F(ST)values do not imply population-specific selection in humans: the NRG1 example. *Human Genetics*. pp. 759-762.

58 Giger, U., Sargan, D.R. and McNiel, E.A. (2006) Breed-specific hereditary diseases and genetic screening. In Ostrander, E.A., Giger, U. and Lindblad-Toh, K. (eds), *The Dog ans its Genome*. Cold Spring Harbor Laboratory Press, pp. 584.

59 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular Biology and Evolution*, **11**, 725-736.
60 Gould, S.J. and Eldredge, N. (1977) Punctuated equilibria: the tempo and mode of evolution reconsidered, *Paleobiology*, **3**, 115-157.
61 Grall, A., Guaguère, E., Planchais, S., Grond, S., Bourrat, E., Hausser, I.*, et al.* (2011) PNPLA1 mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans, *Nature genetics*, **In press.**
62 Gray, M., Granka, J., Bustamante, C., Sutter, N., Boyko, A., Zhu, L.*, et al.* (2009) Linkage Disequilibrium and Demographic History of Wild and Domestic Canids, *Genetics*.
63 Gray, M.M., Sutter, N.B., Ostrander, E.A. and Wayne, R.K. (2010) The IGF1 small dog haplotype is derived from Middle Eastern grey wolves, *BMC Biol*, **8**, 16.
64 Gudbjartsson, D.F., Walters, G.B., Thorleifsson, G., Stefansson, H., Halldorsson, B.V., Zusmanovich, P.*, et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nature genetics*. pp. 609-615.
65 Guyon, R., Kirkness, E.F., Lorentzen, T.D., Hitte, C., Comstock, K.E., Quignon, P.*, et al.* (2003a) Building comparative maps using 1.5x sequence coverage: human chromosome 1p and the canine genome, *Cold Spring Harb Symp Quant Biol*, **68**, 171-177.
66 Guyon, R., Lorentzen, T.D., Hitte, C., Kim, L., Cadieu, E., Parker, H.G.*, et al.* (2003b) A 1-Mb resolution radiation hybrid map of the canine genome, *Proc Natl Acad Sci U S A*, **100**, 5296-5301.
67 Guyon, R., Pearce-Kelling, S.E., Zeiss, C.J., Acland, G.M. and Aguirre, G.D. (2007) Analysis of six candidate genes as potential modifiers of disease expression in canine XLPRA1, a model for human X-linked retinitis pigmentosa 3, *Mol Vis*, **13**, 1094-1105.
68 Haldane, J. (1927a) The comparative genetics of color in rodents and carnivora., *Biol. Rev. Camb. Philos. Soc.*, **2**, 199–212.
69 Haldane, J.B.S. (1927b) A mathematical theory of natural and artificial selection. Part V. Selection and mutation., *Proc. Cambridge Philos. Soc.*, **23**, 838-844.
70 Hardy, G.H. (1908) Mendelian proportions in a mixed population, *Science*, **28**, 49-50.
71 Hare, B. and Tomasello, M. (2005) Human-like social skills in dogs?, *Trends in cognitive sciences*, **9**, 439-444.
72 Hédan, B., Corre, S., Hitte, C., Dréano, S., Vilboux, T., Derrien, T.*, et al.* (2006) Coat colour in dogs: identification of the merle locus in the Australian shepherd breed. *BMC Vet Res*. pp. 9.
73 Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M.*, et al.* (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA*. pp. 5154-5162.
74 Herzog, R.W., Yang, E.Y., Couto, L.B., Hagstrom, J.N., Elwell, D., Fields, P.A.*, et al.* (1999) Long-term correction of canine hemophilia B by gene transfer of blood coagulation factor IX mediated by adeno-associated viral vector, *Nat Med*, **5**, 56-63.
75 Hitte, C., Madeoy, J., Kirkness, E.F., Priat, C., Lorentzen, T.D., Senger, F.*, et al.* (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping, *Nat Rev Genet*, **6**, 643-648.
76 Hitte, C., Kirkness, E.F., Ostrander, E.A. and Galibert, F. (2008) Survey sequencing and radiation hybrid mapping to construct comparative maps, *Methods Mol Biol*, **422**, 65-77.
77 Huxley, T.H. (1860) The Origin of Species, *Westmint. Rev.*, **17**, 541-570.
78 Irion, D.N., Schaffer, A.L., Famula, T.R., Eggleston, M.L., Hughes, S.S. and Pedersen, N.C. (2003) Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers, *J Hered*, **94**, 81-87.

79 Javed, A., Melé, M., Pybus, M., Zalloua, P., Haber, M., Comas, D., *et al.* (2011) Recombination networks as genetic markers in a human variation study of the Old World. *Human Genetics*.
80 Jones, P., Chase, K., Martin, A., Davern, P. and Ostrander …, E. (2008) Single-Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes, *Genetics*.
81 Jouquand, S., Priat, C., Hitte, C., Lachaume, P., Andre, C. and Galibert, F. (2000) Identification and characterization of a set of 100 tri- and dinucleotide microsatellites in the canine genome, *Anim Genet*, **31**, 266-272.
82 Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H., Zody, M.C., Anderson, N., *et al.* (2007) Efficient mapping of mendelian traits in dogs through genome-wide association, *Nature genetics*, **39**, 1321-1328.
83 Karlsson, E.K. and Lindblad-Toh, K. (2008) Leader of the pack: gene mapping in dogs and other model organisms, *Nat Rev Genet*, **9**, 713-725.
84 Kayser, M., Brauer, S. and Stoneking, M. (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*. pp. 893-900.
85 Kern, A.D. and Haussler, D. (2010) A population genetic hidden Markov model for detecting genomic regions under selection. *Molecular Biology and Evolution*. pp. 1673-1685.
86 Kimura, M. (1968) Evolutionary rate at the molecular level, *Nature*, **217**, 624-626.
87 Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., *et al.* (2003) The dog genome: survey sequencing and comparative analysis, *Science*, **301**, 1898-1903.
88 Kukekova, A.V., Acland, G.M., Oskina, I.N., Kharlamova, A.V., Trut, L.N., Chase, K., *et al.* (2006) The Genetics of Domesticated Behavior in Canids: What can Dogs and Silver Foxes tell us about each other? In Ostrander, E.A., Giger, U. and Lindblad-Toh, K. (eds), *The Dog and its Genome*. Cold Spring Harbor Laboratory Press, pp. 584.
89 Lander, E.S. and Botstein, D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children, *Science*, **236**, 1567-1570.
90 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., *et al.* (2001) Initial sequencing and analysis of the human genome, *Nature*, **409**, 860-921.
91 Langford, C.F., Fischer, P.E., Binns, M.M., Holmes, N.G. and Carter, N.P. (1996) Chromosome-specific paints from a high-resolution flow karyotype of the dog, *Chromosome Res*, **4**, 115-123.
92 Le Meur, G., Stieger, K., Smith, A.J., Weber, M., Deschamps, J.Y., Nivard, D., *et al.* (2007) Restoration of vision in RPE65-deficient Briard dogs using an AAV serotype 4 vector that specifically targets the retinal pigmented epithelium, *Gene Ther*, **14**, 292-303.
93 Leonard, J.A., Wayne, R.K., Wheeler, J., Valadez, R., Guillen, S. and Vila, C. (2002) Ancient DNA evidence for Old World origin of New World dogs, *Science (New York, N.Y*, **298**, 1613-1616.
94 Lequarré, A.-S., Andersson, L., André, C., Fredholm, M., Hitte, C., Leeb, T., *et al.* (2011) LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs, *Vet J*, **189**, 155-159.
95 Leroy, G., Verrier, E., Meriaux, J.C. and Rognon, X. (2009) Genetic diversity of dog breeds: between-breed diversity, breed assignation and conservation approaches. *Animal genetics*. pp. 333-343.
96 Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature genetics*. pp. 584-591.

97 Lewontin, R.C. and Krakauer, J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms, *Genetics*, **74**, 175-195.
98 Li, W.H., Wu, C.I. and Luo, C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*. pp. 150-174.
99 Lin, C.T., Gould, D.J., Petersen-Jonest, S.M. and Sargan, D.R. (2002) Canine inherited retinal degenerations: update on molecular genetic research and its clinical application, *J Small Anim Pract*, **43**, 426-432.
100 Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X.*, et al.* (1999) The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene, *Cell*, **98**, 365-376.
101 Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M.*, et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog, *Nature*, **438**, 803-819.
102 Lingaas, F., Sorensen, A., Juneja, R.K., Johansson, S., Fredholm, M., Wintero, A.K.*, et al.* (1997) Towards construction of a canine linkage map: establishment of 16 linkage groups, *Mamm Genome*, **8**, 218-221.
103 Lohi, H., Young, E.J., Fitzmaurice, S.N., Rusbridge, C., Chan, E.M., Vervoort, M.*, et al.* (2005) Expanded repeat in canine epilepsy, *Science*, **307**, 81.
104 Mallick, S., Gnerre, S., Muller, P. and Reich, D. (2009) The difficulty of avoiding false positives in genome scans for natural selection, *Genome Research*, **19**, 922-933.
105 Manten, A. (1963) The non-medical use of antibiotics ans the risk of causing microbial drug-resistance, *Bull World Health Organ*, **29**, 387-400.
106 Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene, *Genetical Research*, 23-35.
107 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila, *Nature*, **351**, 652-654.
108 McKusick, V.A. (1998) Online Mendelian Inheritance in Man, OMIM®. Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore.
109 Mech, L.D. (1970) *The wolf: The ecology and behavior of an endangered species.* Natural History Press, Garden City, NY. .
110 Mech, L.D. and Boitani, L. (2003) *Wolves: behavior, ecology and conservation*. Chicago.
111 Mellersh, C.S., Langston, A.A., Acland, G.M., Fleming, M.A., Ray, K., Wiegand, N.A.*, et al.* (1997) A linkage map of the canine genome, *Genomics*, **46**, 326-336.
112 Mellersh, C.S., Hitte, C., Richman, M., Vignaux, F., Priat, C., Jouquand, S.*, et al.* (2000) An integrated linkage-radiation hybrid map of the canine genome, *Mamm Genome*, **11**, 120-130.
113 Mendel, G. (1865) Versuche über Pflanzen-Hybriden., *Verh. Naturforsch. Ver. Brunn*, **4**, 3-47.
114 Merveille, A.C., Davis, E.E., Becker-Heck, A., Legendre, M., Amirav, I., Bataille, G.*, et al.* (2011) CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs, *Nature genetics*, **43**, 72-78.
115 Michell, A.R. (1999) Longevity of British breeds of dog and its relationships with sex, size, cardiovascular variables and disease. *Vet Rec*. pp. 625-629.
116 Mignot, E., Bell, R.A., Rattazzi, C., Lovett, M., Grumet, F.C. and Dement, W.C. (1994) An immunoglobulin switchlike sequence is linked with canine narcolepsy. *Sleep*. pp. S68-76.
117 Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol*, **302**, 205-217.
118 Okada, N. (1991) SINEs. *Curr Opin Genet Dev*. pp. 498-504.

119 Oleksyk, T.K., Zhao, K., De La Vega, F.M., Gilbert, D.A., O'Brien, S.J. and Smith, M.W. (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE*. pp. e1712.

120 Oleksyk, T.K., Smith, M.W. and O'Brien, S.J. (2010) Genome-wide scans for footprints of natural selection, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 185-205.

121 Olsson, M., Meadows, J.R., Truve, K., Rosengren Pielberg, G., Puppo, F., Mauceli, E.*, et al.* (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs, *PLoS Genet*, **7**, e1001332.

122 Ostrander, E.A., Sprague, G.F., Jr. and Rine, J. (1993) Identification and characterization of dinucleotide repeat (CA)n markers for genetic mapping in dog, *Genomics*, **16**, 207-213.

123 Ostrander, E.A. and Giniger, E. (1997) Semper fidelis: what man's best friend can teach us about human biology and disease, *Am J Hum Genet*, **61**, 475-480.

124 Ostrander, E.A., Galibert, F. and Patterson, D.F. (2000) Canine genetics comes of age, *Trends Genet*, **16**, 117-124.

125 Ostrander, E.A. and Kruglyak, L. (2000) Unleashing the canine genome, *Genome Res*, **10**, 1271-1274.

126 Ouyang, Z. and Liang, J. (2007) Detecting positively selected sites from amino Acid sequences: an implicit codon model. *Conf Proc IEEE Eng Med Biol Soc*. pp. 5302-5306.

127 Owen, R. (1848) *On the archetype and homologies of the vertebrate skeleton*. Murray, London.

128 Parker, H.G., Kim, L.V., Sutter, N.B., Carlson, S., Lorentzen, T.D., Malek, T.B.*, et al.* (2004) Genetic structure of the purebred domestic dog, *Science (New York, N.Y*, **304**, 1160-1164.

129 Parker, H.G. and Ostrander, E.A. (2005) Canine genomics and genetics: running with the pack, *PLoS Genet*, **1**, e58.

130 Parker, H.G., Sutter, N.B. and Ostrander, E.A. (2006) Understanding genetic relationships among purebred dogs: the PhyDo project. In Ostrander, E.A., Giger, U. and Lindblad-Toh, K. (eds), *The Dog and its Genome*. Cold Spring Harbor Laboratory Press, pp. 584.

131 Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S.*, et al.* (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs, *Science*, **325**, 995-998.

132 Patin, E., Barreiro, L.B., Sabeti, P.C., Austerlitz, F., Luca, F., Sajantila, A.*, et al.* (2006) Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. *Am J Hum Genet*. pp. 423-436.

133 Patterson, D.F. (2000) Companion animal medicine in the age of medical genetics, *J Vet Intern Med*, **14**, 1-9.

134 Patterson, E.E., Minor, K.M., Tchernatynskaia, A.V., Taylor, S.M., Shelton, G.D., Ekenstedt, K.J. and Mickelson, J.R. (2008) A canine DNM1 mutation is highly associated with the syndrome of exercise-induced collapse, *Nature genetics*, **40**, 1235-1239.

135 Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis, *PLoS Genet*, **2**, e190.

136 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*. pp. 2444-2448.

137 Pele, M., Tiret, L., Kessler, J.L., Blot, S. and Panthier, J.J. (2005) SINE exonic insertion in the PTPLA gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs, *Hum Mol Genet*, **14**, 1417-1427.

138 Peyron, C., Faraco, J., Rogers, W., Ripley, B., Overeem, S., Charnay, Y.*, et al.* (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains, *Nat Med*, **6**, 991-997.
139 Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M.J.*, et al.* (2004) The Ensembl analysis pipeline. *Genome Research*. pp. 934-941.
140 Priat, C., Hitte, C., Vignaux, F., Renier, C., Jiang, Z., Jouquand, S.*, et al.* (1998) A whole-genome radiation hybrid map of the dog genome, *Genomics*, **54**, 361-378.
141 Proux, E., Studer, R.A., Moretti, S. and Robinson-Rechavi, M. (2009) Selectome: a database of positive selection, *Nucleic Acids Research*, **37**, D404-407.
142 Przeworski, M., Coop, G. and Wall, J.D. (2005) The signature of positive selection on standing genetic variation, *Evolution*, **59**, 2312-2323.
143 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D.*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, **81**, 559-575.
144 Quach, H., Barreiro, L.B., Laval, G., Zidane, N., Patin, E., Kidd, K.K.*, et al.* (2009) Signatures of purifying and local positive selection in human miRNAs, *Am J Hum Genet*, **84**, 316-327.
145 Quignon, P., Herbin, L., Cadieu, E., Kirkness, E.F., Hedan, B., Mosher, D.S.*, et al.* (2007) Canine Population Structure: Assessment and Impact of Intra-Breed Stratification on SNP-Based Association Studies, *PLoS ONE*, **2**, e1324.
146 R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*.
147 Ranwez, V., Harispe, S., Delsuc, F. and Douzery, E.J.P. (2011) MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS One*. pp. e22594.
148 Rincon, G., Tengvall, K., Belanger, J.M., Lagoutte, L., Medrano, J.F., André, C.*, et al.* (2011) Comparison of buccal and blood-derived canine DNA, either native or whole genome amplified, for array-based genome-wide association studies. *BMC Res Notes*. pp. 226.
149 Robin, S., Tacher, S., Rimbault, M., Vaysse, A., Dréano, S., André, C.*, et al.* (2009) Genetic diversity of canine olfactory receptors. *BMC genomics*. pp. 21.
150 Robinson, D.M. (2003) Protein Evolution with Dependence Among Codons Due to Tertiary Structure. *Molecular Biology and Evolution*. pp. 1692-1704.
151 Roth, C., Betts, M.J., Steffansson, P., Saelensminde, G. and Liberles, D.A. (2005) The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics, *Nucleic Acids Research*, **33**, D495-497.
152 Rubin, C.-J., Zody, M.C., Eriksson, J., Meadows, J.R.S., Sherwood, E., Webster, M.T.*, et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication, *Nature*, **464**, 587-591.
153 Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F.*, et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**, 832-837.
154 Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C.*, et al.* (2007) Genome-wide detection and characterization of positive selection in human populations, *Nature*, **449**, 913-918.
155 Sampaolesi, M., Blot, S., D'Antona, G., Granger, N., Tonlorenzi, R., Innocenzi, A.*, et al.* (2006) Mesoangioblast stem cells ameliorate muscle function in dystrophic dogs, *Nature*, **444**, 574-579.

156 Sargan, D.R., Aguirre-Hernandez, J., Galibert, F. and Ostrander, E.A. (2007) An extended microsatellite set for linkage mapping in the domestic dog, *J Hered*, **98**, 221-231.
157 Savolainen, P., Zhang, Y.-p., Luo, J., Lundeberg, J. and Leitner, T. (2002) Genetic Evidence for an East Asian Origin of Domestic Dogs, *Science (New York, N.Y*, **298**, 1610-1613.
158 Schaffner, S. and Sabeti, P. (2008) Evolutionary adaptation in the human lineage, *Nature Education*, **1**.
159 Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*. pp. 413-429.
160 Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *Am J Hum Genet*, **52**, 506-516.
161 Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C*., et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*. pp. 1611-1618.
162 Stajich, J.E. and Hahn, M.W. (2005) Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*. pp. 63-73.
163 Starkey, M.P., Scase, T.J., Mellersh, C.S. and Murphy, S. (2005) Dogs really are man's best friend--canine genomics has applications in veterinary and human medicine!, *Brief Funct Genomic Proteomic*, **4**, 112-128.
164 Storz, J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence, *Mol Ecol*, **14**, 671-688.
165 Sutter, N.B., Eberle, M.A., Parker, H.G., Pullar, B.J., Kirkness, E.F., Kruglyak, L. and Ostrander, E.A. (2004) Extensive and breed-specific linkage disequilibrium in Canis familiaris, *Genome Res*, **14**, 2388-2396.
166 Sutter, N.B. and Ostrander, E.A. (2004) Dog star rising: the canine genetic system, *Nat Rev Genet*, **5**, 900-910.
167 Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L*., et al.* (2007) A single IGF1 allele is a major determinant of small size in dogs, *Science*, **316**, 112-115.
168 Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*. pp. W609-612.
169 Switonski, M., Reimann, N., Bosma, A.A., Long, S., Bartnitzke, S., Pienkowska, A*., et al.* (1996) Report on the progress of standardization of the G-banded canine (Canis familiaris) karyotype. Committee for the Standardized Karyotype of the Dog (Canis familiaris), *Chromosome Res*, **4**, 306-309.
170 Switonski, M., Szczerbal, I. and Nowacka, J. (2004) The dog genome map and its use in mammalian comparative genomics, *J Appl Genet*, **45**, 195-214.
171 Tacher, S., Quignon, P., Rimbault, M., Dreano, S., Andre, C. and Galibert, F. (2005) Olfactory receptor sequence polymorphism within and between breeds of dogs. *The Journal of heredity*. pp. 812-816.
172 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics*, **123**, 585–595.
173 Tennessen, J.A., O'Connor, T.D., Bamshad, M.J. and Akey, J.M. (2011) The promise and limitations of population exomics for human evolution studies. *Genome Biol*. pp. 127.
174 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. pp. 4673-4680.

175 Tsai, K.L., Clark, L.A. and Murphy, K.E. (2007) Understanding hereditary diseases using the dog and human as companion model systems, *Mamm Genome*, **18**, 444-451.
176 Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S.*, et al.* (2011) Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping, *PLoS Genet*, **7**, e1002316.
177 Venditti, C., Meade, A. and Pagel, M. (2011) Multiple routes to mammalian diversity. *Nature*.
178 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G.*, et al.* (2001) The sequence of the human genome. *Science*. pp. 1304-1351.
179 Vignaux, F., Hitte, C., Priat, C., Chuat, J.C., Andre, C. and Galibert, F. (1999) Construction and optimization of a dog whole-genome radiation hybrid panel, *Mamm Genome*, **10**, 888-894.
180 Vila, C., Amorim, I.R., Leonard, J.A., Posada, D., Castroviejo, J., Petrucci-Fonseca, F.*, et al.* (1999a) Mitochondrial DNA phylogeography and population history of the grey wolf canis lupus, *Mol Ecol*, **8**, 2089-2103.
181 Vila, C., Maldonado, J.E. and Wayne, R.K. (1999b) Phylogenetic relationships, evolution, and genetic diversity of the domestic dog, *J Hered*, **90**, 71-77.
182 Vilà, C., Seddon, J. and Ellegren, H. (2005) Genes of domestic mammals augmented by backcrossing with wild ancestors, *Trends Genet*, **21**, 214-218.
183 Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome, *PLoS Biol*, **4**, e72.
184 Vonholdt, B.M., Pollinger, J.P., Lohmueller, K.E., Han, E., Parker, H.G., Quignon, P.*, et al.* (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication, *Nature*, **464**, 898-902.
185 Wade, C.M., Karlsson, E.K., Mikkelsen, T.S., Zody, M.C. and Lindblad-Toh, K. (2006) The Dog Genome: Sequence, Evolution, and Haplotype structure. In Ostrander, E.A., Giger, U. and Lindblad-Toh, K. (eds), *The Dog and its Genome*. Cold Spring Harbor Laboratory Press, pp. 584.
186 Wang, B., Zhang, Y.-B., Zhang, F., Lin, H., Wang, X., Wan, N.*, et al.* (2011) On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS ONE*. pp. e17002.
187 Wang, W. and Kirkness, E.F. (2005) Short interspersed elements (SINEs) are a major source of canine genomic diversity, *Genome Res*, **15**, 1798-1808.
188 Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature*, **171**, 737-738.
189 Wayne, R.K. and Vilà, C. (2001) Phylogeny and Origin of the Domestic Dog. In Ruvinsky, A. and Sampson, J. (eds), *The Genetics of the Dog*. CABI Publishing, pp. 564.
190 Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M.*, et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*. pp. 575-583.
191 Weinberg, W. (1908) Uber den Nachweis der Vererbung beim Menschen., *Jahresh. Ver. Vaterl. Naturkd. Wuerttemb.*, **64**, 368-382.
192 Weir, B.S. (1996) *Genetics Data Analysis II*. Sinauer Associates Inc.
193 Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. and Hill, W.G. (2005) Measures of human population structure show heterogeneity among genomic regions, *Genome Research*, **15**, 1468-1476.

194 Werner, P., Mellersh, C.S., Raducha, M.G., DeRose, S., Acland, G.M., Prociuk, U.*, et al.* (1999) Anchoring of canine linkage groups with chromosome-specific markers, *Mamm Genome*, **10**, 814-823.
195 Wiehe, T. (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theoretical population biology*. pp. 272-283.
196 Wilkins, M.H.F., Stokes, A.R. and Wilson, H.R. (1953) Molecular structure of deoxypentose nucleic acids, *Nature*, **171**, 738-740.
197 Wong, A.K., Ruhe, A.L., Dumont, B.L., Robertson, K.R., Guerrero, G., Shull, S.M.*, et al.* (2010) A comprehensive linkage map of the dog genome. *Genetics*. pp. 595-605.
198 Wright, S. (1921) Systems of Mating. V. General Considerations, *Genetics*, **6**, 167-178.
199 Wright, S. (1931) Evolution in Mendelian Populations, *Genetics*, **16**, 97-159.
200 Wright, S. (1943) Isolation by Distance, *Genetics*, **28**, 114-138.
201 Wright, S. (1951) The genetical structure of populations, *Annals of Eugenics*, **15**, 323-354.
202 Xiong, Y. and Eickbush, T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J*. pp. 3353-3362.
203 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput Appl Biosci*, **13**, 555-556.
204 Yang, Z. and Nielsen, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals, *J Mol Evol*, **46**, 409-418.
205 Yang, Z. and Dos Reis, M. (2011) Statistical properties of the branch-site test of positive selection, *Molecular Biology and Evolution*, **28**, 1217.
206 Zeng, K., Fu, Y.X., Shi, S. and Wu, C.I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants, *Genetics*, **174**, 1431–1439.
207 Zhang, B., Kirov, S. and Snoddy, J. (2005a) WebGestalt: an integrated system for exploring gene sets in various biological contexts, *Nucleic Acids Res*, **33**, W741-748.
208 Zhang, J., Nielsen, R. and Yang, Z. (2005b) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level, *Molecular Biology and Evolution*, **22**, 2472-2479.

Annexe

I. Publications liées à la thèse

Le travail que j'ai effectué au cours de ma thèse a fait l'objet de plusieurs publications :

1) Dans le premier article, j'ai utilisé le ratio dN/dS pour évaluer le niveau de contraintes sélectives qui agissent sur les gènes prédits comme pseudogènes dans le génome du chien mais fonctionnels chez les autres espèces mammifères. Cet article est le point de départ de mon travail de thèse. Ce travail est presenté

Revisiting the missing protein-coding gene catalog of the domestic dog

Derrien, T., Théze, J., Vaysse, A., André, C., Ostrander, E.A., Galibert, F. and Hitte, C. (2009), *BMC Genomics*, 10, 62.

Among mammals for which there is a high sequence coverage, the whole genome assembly of the dog is unique in that it predicts a low number of protein-coding genes, ~19,000, compared to the over 20,000 reported for other mammalian species. Of particular interest are the more than 400 of genes annotated in primates and rodent genomes, but missing in dog.

Using over 14,000 orthologous genes between human, chimpanzee, mouse rat and dog, we built multiple pairwise synteny maps to infer short orthologous intervals that were targeted for characterizing the canine missing genes. Based on gene prediction and a functionality test using the ratio of replacement to silent nucleotide substitution rates (d(N)/d(S)), we provide compelling structural and functional evidence for the identification of 232 new protein-coding genes in the canine genome and 69 gene losses, characterized as undetected gene or pseudogenes. Gene loss phyletic pattern analysis using ten species from chicken to human allowed us to characterize 28 canine-specific gene losses that have functional orthologs continuously from chicken or marsupials through human, and 10 genes that arose specifically in the evolutionary lineage leading to rodent and primates.

This study demonstrates the central role of comparative genomics for refining gene catalogs and exploring the evolutionary history of gene repertoires, particularly as applied for the characterization of species-specific gene gains and losses.

2) Le travail de recherche des signatures de différenciation génétique entre races canines est publié dans l'article Vaysse et al. 2011 Plos Genet. Ce travail est présenté dans la partie II des résultats du manuscrit de thèse.

Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MS, Lawley CT, Karlsson EK; The LUPA Consortium, Bannasch D, Vilà C, Lohi H, Galibert F, Fredholm M, Häggström J, Hedhammar A, André C, Lindblad-Toh K, Hitte C, Webster MT. (2011), PLoS Genet, 7 Oct;7(10):e1002316.

The extraordinary phenotypic diversity of dog breeds has been sculpted by a unique population history accompanied by selection for novel and desirable traits. Here we perform a comprehensive analysis using multiple test statistics to identify regions under selection in 509 dogs from 46 diverse breeds using a newly developed high-density genotyping array consisting of >170,000 evenly spaced SNPs. We first identify 44 genomic regions exhibiting extreme differentiation across multiple breeds. Genetic variation in these regions correlates with variation in several phenotypic traits that vary between breeds, and we identify novel associations with both morphological and behavioral traits. We next scan the genome for signatures of selective sweeps in single breeds, characterized by long regions of reduced heterozygosity and fixation of extended haplotypes. These scans identify hundreds of regions, including 22 blocks of homozygosity longer than one megabase in certain breeds. Candidate selection loci are strongly enriched for developmental genes. We chose one highly differentiated region, associated with body size and ear morphology, and characterized it using high-throughput sequencing to provide a list of variants that may directly affect these traits. This study provides a catalogue of genomic regions showing extreme reduction in genetic variation or population differentiation in dogs, including many linked to phenotypic variation. The many blocks of reduced haplotype diversity observed across the genome in dog breeds are the result of both selection and genetic drift, but extended blocks of homozygosity on a megabase scale appear to be best explained by selection. Further elucidation of the variants under selection will help to uncover the genetic basis of complex traits and disease.

3) J'ai participé à l'élaboration d'une revue sur la structure du génome canin, la nécessité de ré-annotation du génome canin et les outils pour y parvenir. Dans cet article, je décris l'importance de la sélection artificielle dans l'évolution du chien. Ces concepts sont présentés dans l'introduction et la discussion du manuscrit de thèse.

Annotation of the Domestic Dog Genome Sequence : Finding the missing genes

Derrien, T., Vaysse, A., André, C. and Hitte, C. (2011), Mammalian Genome Nov 11

The domestic dog is composed of over 350 genetically distinct breeds that present considerable variations in morphology, physiology, and diseases susceptibility. Its genome sequence was assembled and released in 2005 providing an estimate of about 20,000 protein-coding genes that are a great asset for the scientific community that uses the dog system as a genetic biomedical model and for comparative and evolutionary studies. Although the canine gene set has been predicted using a combination of ab initio methods, homology studies, motif analysis and similarity-based programs, it still requires a deep annotation of noncoding genes, alternative splicing, pseudogenes, regulatory regions and gain and loss events. Such analyses could benefit from new sequencing technologies (RNA-Seq) to better exploit the advantages of the canine genetic system in tracking disease genes. Here, we review the catalog of canine protein-coding genesand the search for missing genes, and we propose rationales for an accurate identification of noncoding genes though next-generation sequencing.

II. Publication en préparation

Le serveur OMEGA fera l'objet d'une publication de type "application note".
OMEGABASE: a web server for analyzing selective constraints acting on mammalian genes.
Vaysse, A., André, C and Hitte, C. en préparation

III. Publications non liées à la thèse

Au cours de ma thèse, j'ai eu plusieurs opportunités de réaliser des analyses bioinformatiques et statistiques faisant parties des projets portés par d'autres membres de l'équipe génétique du chien et leurs collaborateurs. Ces travaux collaboratifs ne sont pas décrits dans le manuscrit de thèse :

1- Abadie J, Hédan B, Cadieu E, De Brito C, Devauchelle P, Bourgain C, Parker HG, Vaysse A, Margaritte-Jeannin P, Galibert F, Ostrander EA, André C. (2009) **Epidemiology, pathology, and genetics of histiocytic sarcoma in the Bernese mountain dog breed**. *J Hered 100 Suppl 1:S19-27*.

2- Rimbault, M., Robin, S., Vaysse, A. and Galibert, F. (2009) **RNA profiles of rat olfactory epithelia: individual and age related variations.** *BMC Genomics* Dec 2;10:572.

3- Robin, S., Tacher, S., Rimbault, M., Vaysse, A., Dréano, S., André, C., *Hitte C., Galibert F.* (2009) **Genetic diversity of canine olfactory receptors.** *BMC GenomicsJan 14;10:21*..

BMC Genomics

BioMed Central

Research article

Open Access

Revisiting the missing protein-coding gene catalog of the domestic dog

Thomas Derrien[1,3], Julien Thézé[1], Amaury Vaysse[1], Catherine André[1], Elaine A Ostrander[2], Francis Galibert[1] and Christophe Hitte*[1]

Address: [1]Institut de Génétique et Développement, CNRS UMR6061, Université de Rennes1, 2 Av du Pr. Léon Bernard, 35043 Rennes, France, [2]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Bethesda MD 20892, USA and [3]Centre for Genomic Regulation (CRG), Bioinformatics Program C/Dr. Aiguader, 88 08003 Barcelona, Spain

Email: Thomas Derrien - thomas.derrien@crg.es; Julien Thézé - theze.julien@gmail.com; Amaury Vaysse - amaury.vaysse@univ-rennes1.fr; Catherine André - catherine.andre@univ-rennes1.fr; Elaine A Ostrander - eostrand@mail.nih.gov; Francis Galibert - francis.galibert@univ-rennes1.fr; Christophe Hitte* - hitte@univ-rennes1.fr

* Corresponding author

Published: 4 February 2009

BMC Genomics 2009, **10**:62 doi:10.1186/1471-2164-10-62

Received: 28 August 2008
Accepted: 4 February 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/62

Abstract

Background: Among mammals for which there is a high sequence coverage, the whole genome assembly of the dog is unique in that it predicts a low number of protein-coding genes, ~19,000, compared to the over 20,000 reported for other mammalian species. Of particular interest are the more than 400 of genes annotated in primates and rodent genomes, but missing in dog.

Results: Using over 14,000 orthologous genes between human, chimpanzee, mouse rat and dog, we built multiple pairwise synteny maps to infer short orthologous intervals that were targeted for characterizing the canine missing genes. Based on gene prediction and a functionality test using the ratio of replacement to silent nucleotide substitution rates (d_N/d_S), we provide compelling structural and functional evidence for the identification of 232 new protein-coding genes in the canine genome and 69 gene losses, characterized as undetected gene or pseudogenes. Gene loss phyletic pattern analysis using ten species from chicken to human allowed us to characterize 28 canine-specific gene losses that have functional orthologs continuously from chicken or marsupials through human, and 10 genes that arose specifically in the evolutionary lineage leading to rodent and primates.

Conclusion: This study demonstrates the central role of comparative genomics for refining gene catalogs and exploring the evolutionary history of gene repertoires, particularly as applied for the characterization of species-specific gene gains and losses.

Background

Comparative genomics plays a key role in understanding organism evolution, refining functional annotation and identifying orthology relationships. By taking advantage of whole-genome sequence assemblies with a high level of coverage [1-4], one can seek to provide exhaustive and genome-scale level predictions regarding functional sequence [5]. The general approach relies on the exploitation of sequence similarities [6-8] phylogenetic data [9,10], evolutionary models [11,12] and evidence regarding conservation of gene order [13-15]. These often complementary comparative approaches have been developed

to estimate and improve the identification of functional sequences for both newly sequenced species as well as reference species, such as human and mouse [16-18]. Moreover, multispecies genome scale comparisons allow to refine protein-coding genes annotation [19-21] as well as better understanding of the timing and the frequency of duplication events for lineage-specific genes called in-paralogs [22,23].

Fine-scale comparative maps constructed using robust orthologous sequences are key for allowing identification, visualization and characterization of conserved segments as well as collinearity of gene order between the species [24,25]. Gene order between species is not random and this has been shown to correlate with, for example, co-expressed and co-regulated genes suggesting a functional significance [26]. Otherwise, gene order conservation between species could also be exploited to identify relocated protein-coding genes in non-syntenic chromosomal regions [27], as well as potentially retrotransposed genes given that the latter correspond mostly to pseudogenes inserted in non-syntenic regions [10]. Consequently, as part of the characterization of architecture of a genome, analysis of gene order conservation between species can be a strong indicator for both gene prediction [28] and identification of gene loss [29].

In this study, we have analyzed the sequence assembly of the domestic dog for which the annotation process identified less protein-coding genes than expected compared to predictions from the primates and rodent genomes. We focused on a set of 412 genes that are all annotated in four closely related mammals; human, chimpanzee, mouse and rat, but absent in the dog genome in the most recent assembly of the dog (CanFam 2.0). We exploited the property of gene adjacency conservation between related species to target in-depth sequence alignments over a short genomic interval. In addition, our approach includes a functionality test that investigates the ratio of amino acid replacement (nonsynonymous, d_N) to silent (synonymous, d_S) substitution rates, which indicates selective constraints acting on a given genomic regions [10]. As mutations in genes causing amino acid replacements with functional consequences are selected against in contrast to mutations occurring in pseudogenes, we took advantage of the distinctive patterns of d_N/d_S ratios to refine the identification of new gene predictions and gene losses occurring in dog.

Using the above strategies we identified 232 canine genes for which synteny conservation, cross-species sequence analysis and the neutral rate of evolution based on d_N/d_S results converged strongly to support their existence. In addition, we identified 69 gene-loss candidates of which predictions for which accumulating ORF-disrupting mutations, and significant d_N/d_S ratios support scenarios of 21 genes lost as pseudogenes in the canine species. To further characterize gene losses, we inferred their phyletic pattern in ten species from chicken to human over a period of 310 million years. Therefore, we were able to differentiate canine-specific losses from gene losses that have occurred in others lineage or genes formed after the evolutionary branchpoint leading to dog.

Results

Using all annotated genes from human, chimpanzee, mouse, rat and dog (Ensembl v42) [30], we extracted 412 genes annotated as protein-coding in all species but dog. These genes exhibit a '1:1:1:1:0' phyletic pattern, that is indicative of the presence/absence of genes with a one-to-one orthologous relationship among the five species. We refer to these as 'missing genes' for purposes of this study. We examined the structural features of the 412 missing genes in the four mammalian reference sequences and compared them to an independent and randomly selected set of 400 genes. The mean length of the protein products of the missing genes set was 722 amino acids (AA), which is significantly smaller than the random set at 905 AA (*t* test; $P = 6.8e - 11$). Similarly, the mean transcript size was ~50% smaller than observed in a random set (*t* test; $P = 2.6e - 9$). The mean number of exons in missing genes was also smaller (5.8 vs 9.8; *t* test;$P = 3.7e - 13$) than the random set and particularly single-exon genes were found to be over represented by 15%. To ensure that single-exon missing genes were functional and not processed pseudogenes, we analyzed each, using the human dataset, for accumulated degenerative mutations (frameshifts and premature stop codons) in their coding sequence and found none. In addition, we identified sequence alignment between single-exon genes and ESTs (sequence similarity > 96% for at least 150 bp) for 95% of them.

To test the underlying assumption that missing genes may be implicated in particular biological pathways, we examined their functional annotation in the context of Gene Ontology (GO) using the program GO Tree Machine [31]. Using the human sequence as a reference, the results demonstrate that the missing gene set is enriched for genes implicated in physiological pathways of immunity and organism responses to pathogens (12 genes), olfaction (16) and regulation of transcription (63). This classification comprises functional pathways that play an important role in the adaptation of organisms to their environment. Interestingly, these biological functions are often linked to large proteins families that are attractive targets for lineage-specific functions and lineage-specific loss and gain of genes [32].

Constructing synteny maps with 1:1 orthologs

We extracted pairwise sets of 14,997; 14,798; 14,667 and 14,065 one-to-one (1:1) orthologous protein-coding genes (Ensembl v42) between human and dog (H-D),

chimpanzee-dog (C-D), mouse-dog (M-D) and rat-dog (R-D), respectively. Using those 1:1 orthologs as comparative anchors, we built four fine-scale whole-genome pairwise synteny maps (Additional data file 1) with the program AutoGRAPH, which we recently developed [13]. We identified 218, 229, 326 and 325 CSOs, i.e. chromosomal segments for which markers are in the same linear order on the chromosome as noted across species [25], between H-D, C-D, M-D and R-D respectively. The mean distance between two consecutive genes was ~180 kb. In all synteny maps, CSOs cover almost the entire genome while breakpoint regions, areas delimitating CSOs, cover only ~5% of a genome and may contain single-gene segment or very short synteny blocks [33] (Additional data file 2).

In each pairwise synteny map, we localized the missing gene orthologs on the reference sequence (Figure 1). Of the 412 missing genes, the vast majority (mean of 92.3%; range 92 to 94%) mapped within CSOs with only 7.7% mapping within breakpoints. In all reference species the missing genes spanned all chromosomes, although their distribution varied greatly, i.e. one to 44 per human (HSA) chromosome in the case of the human-dog synteny map.

Targeting genomic intervals

We used multiple pairwise synteny maps described above to identify short, targeted, orthologous genomic intervals. On each reference genome, these intervals are delimited by the closest flanking 1:1 orthologs on either side of each

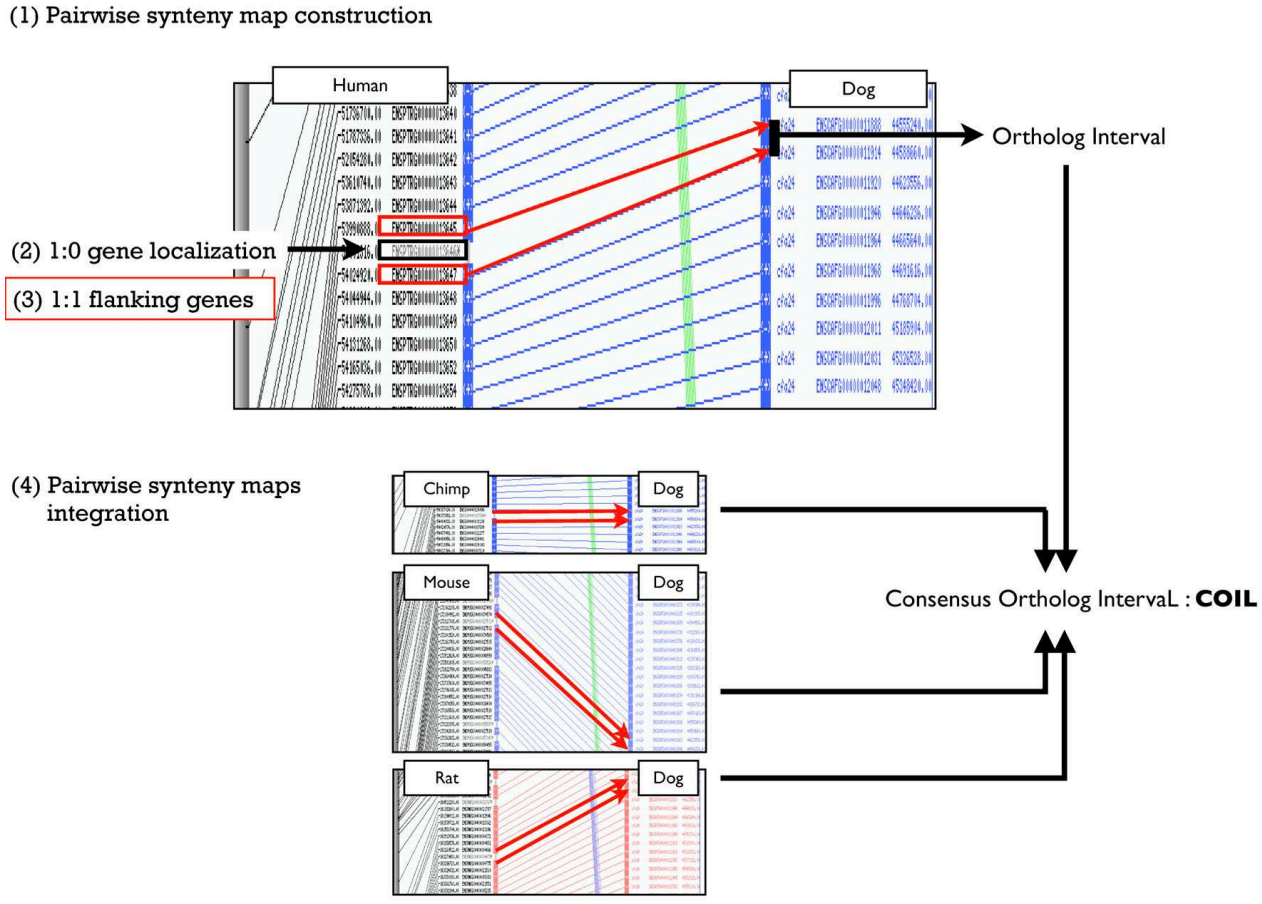


Figure 1

Consensus Ortholog IntervaL identification. The figure illustrates the 4-step method to infer targeted interval for gene prediction. (1) is the first step that build the pairwise synteny map (here a schematic Human-dog syntenic map) using 1:1 orthologs that are connected through colored lines. (2) 1:0 gene ('missing gene' in the dog) is positioned on the reference species of the synteny map. (3) indicates the identification of flanking 1:1 orthologs used to define an orthologous interval on the canine chromosome as indicated by red arrows. (4) is the last step that integrates the four orthologous intervals using all pairwise synteny maps (Chimpanzee-dog; Mouse-dog and Rat-dog) to define a Consensus Ortholog IntervaL (COIL) as shown on the right of the figure.

missing gene that in turn define orthologous intervals on the canine genome as shown in Figure 1. The use of multiple pairwise maps enabled us to identify the shortest consensus interval on the canine genome to search for genes, that we refer to as Consensus Ortholog IntervaLs (COILs) (Figure 1). From the 412 missing genes, we delimited 383 COILs (92.9%) having a mean size of 347 kb (Additional data file 3). For a set of 17 COILs (4.1%) localized in common breakpoint regions (i.e. overlapping between at least two species) [24,34] and for 12 missing genes, no COIL could be determined because of the absence of a consensus interval.

Targeted gene prediction

Within each canine COIL, we used the GeneWise program [6] to splice and align the protein sequence of each reference species in order to most accurately predict the structure of the dog gene. We retained gene predictions produced by at least two reference species protein templates. This produced 231 gene structure predictions with amino acid identity > 40% (Figure 2). Fifty-three genes were predicted using only rodent protein sequence as templates, thus illustrating the complementary contribution of multispecies analysis. We post-processed GeneWise results to detect potential gene features and found the presence of a coding start site for 53.1% of the gene predictions. In addition, amongst the 231 predicted genes, 75% of the predictions with multi-exonic structure exhibit at least a canonical splice site (GT/AG).

To address the question whether COIL delimitation is too restrictive for gene prediction, we aligned the human transcript sequences corresponding to the 383 missing genes for which we defined a COIL, against the assembly of the canine genome sequence (CanFam 2.0) with the Exonerate program [35]. We repeated the analysis with chimpan-

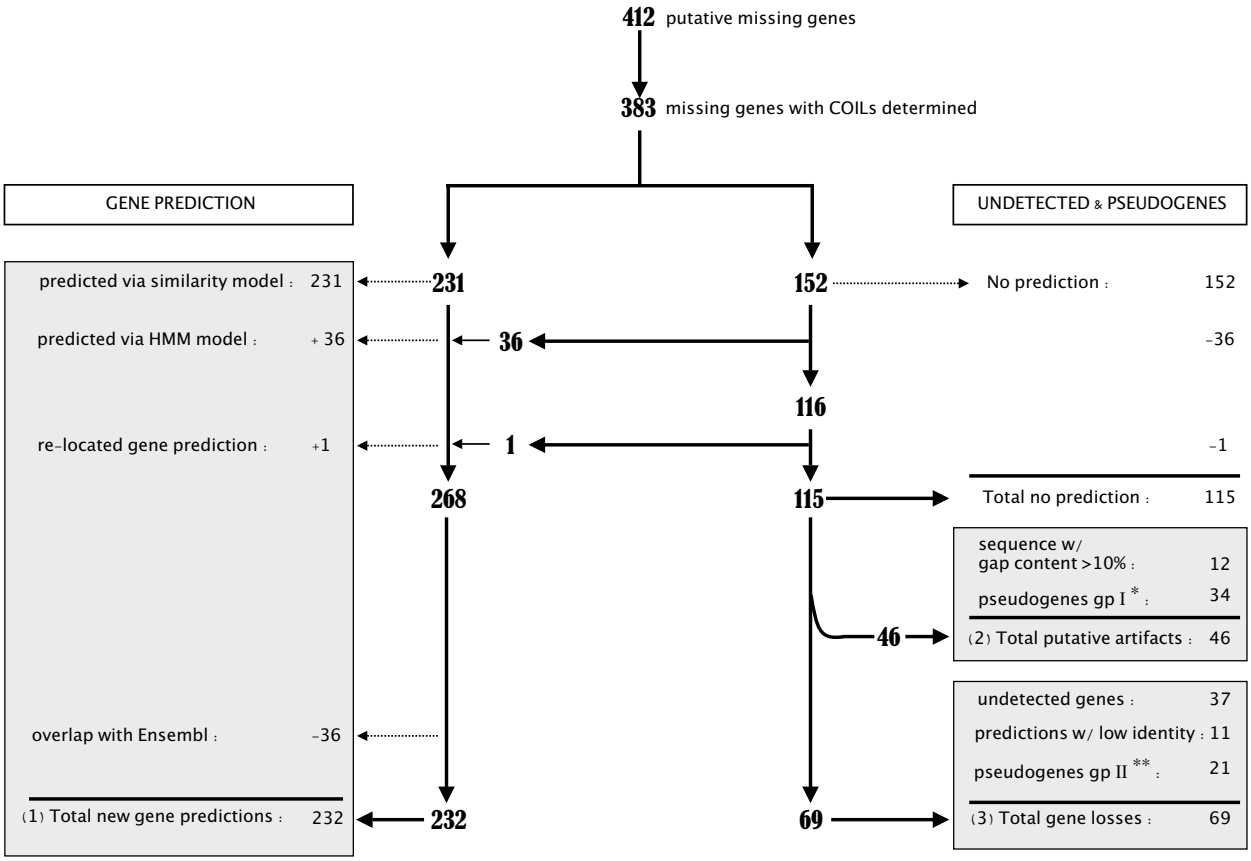


Figure 2

Flowchart of the computational analysis. The left pipeline indicates all steps in the computational analysis of gene predictions and the right pipeline shows a detailed account of the process of undetected genes and pseudogenes. Gray boxes summarize the three main categories (1) new gene predictions, (2) putative artifacts, * indicates pseudogenes identified with low confidence (group I), and (3) gene losses, (**) indicates pseudogenes identified with accumulated mutations (group II) and higher *dN*/*dS* support. See text for details.

zee, mouse and rat transcript sequences. We considered the best five matching sequences to relax the limitations of conventional best-match methods [29]. Then, we defined a concordance between the COIL approach and the whole-genome sequence analysis, when matching sequences from the Exonerate-based analysis for at least two species were totally embedded in COILs. Based on this criterion, concordance was obtained for 342 (89.2%) genes. Of the 41 instances with no agreement between the expected syntenic location and the whole-genome sequence analysis, 36 showed weak match (identity < 20%) within the canine genome assembly suggesting unspecific alignment while five showed a significant match, from at least two species suggesting that these genes may have acquired a new location in the dog. Of the latter five instances, we identified only one gene prediction (*PLA2G4C*) with conservative criteria indicating a relocated gene in a non-syntenic genomic area.

In this study, we applied Genewise program with a sequence similarity-based method that explicitly models the conservation of gene structure and a high degree of conservation. As such model is known to show a marked decrease in performance for less similar genes [36], we further investigate the undetected subset of genes using a probabilistic pair hidden Markov model (HMM) that show a weaker dependence on percent identity and performs better to pick out distant homologs. The Genewise HMM based analysis allowed to predict 36 additional genes (Figure 2). Both prediction sets were merged into a single set (n = 268) for further analysis.

Sequence alignments were next generated between gene predictions and canine transcript sequences (Unigene april 08 [37]). We identified significant alignment (sequence similarity > 96% for at least 150 bp) in 53% of cases with an average of 7.5 ESTs/mRNA per gene prediction (range 1–99). Using Interproscan, [38] protein motifs were found from InterPro database for 80.5% of the gene predictions, providing additional evidence for dog gene identification.

As a further validation step, the construction of canine predicted protein three-dimensional models was investigated based on the homologous structure of the human ortholog or paralog (>40% identity), which was used as a template. For the subset of genes for which the 3D structure is solved (n = 21), canine-human comparative modelling was determined using the SWISS-MODEL server [39]. In 16 instances of canine-human comparative modelling, the mean identity obtained between sequences was 70%. Homology-based 3D model for each canine prediction was validated using the Verify 3D graphs [40] (data not shown) that distinguish between homology models of higher and lower accuracy.

To test for possible overlap between gene predictions obtained in this study and all canine genes annotated in Ensembl (v42), we performed sequence alignment between these two sets of predictions. A total of 232 (88%) predicted genes did not overlap any Ensembl annotated protein-coding genes. Therefore, these were classified as "definite" gene identifications together with the delineation of new orthologous relationships with the four reference species (Additional data file 4). The remaining 36 gene predictions overlapped an annotated gene (protein identity > 80%) indicating that these gene predictions correspond to sequences already defined as genes, but with undetected or spurious orthologous relationships (Figure 2). [41].

Gene prediction assessment from dN/dS *analysis*

To assess the validity of gene predictions through the strength and direction of selective constraints, we used a functionality test that uses the ratio of replacement to silent nucleotide substitution rates (d_N/d_S). The ratio d_N/d_S, where d_N is the number of non-synonymous nucleotide substitution per non-synonymous site and d_S the number of synonymous nucleotide substitution per synonymous site, is used as a proxy for the evolutionary constraints that occur on nucleotide substitution [42]. The calculation of the d_N/d_S ratio requires the comparison to a homologous reference sequence. First, we constructed a benchmark set of true orthologous genes using all 1:1 orthologous genes between human and dog (n = 14,994) to obtain a representative d_N/d_S value. From this benchmark set, we calculated the median d_N/d_S ratio of 0.15 using all d_N/d_S values extracted from the pairwise alignments of transcripts (Figure 3). To assess the 232 gene predictions identified in this study with the functionality test, we determined d_N/d_S ratio for each of the gene predictions in comparison to their human functional orthologous gene from pairwise transcripts alignments. We calculated a median d_N/d_S of 0.19, a value highly similar to the benchmark set (0.15). To further assess the d_N/d_Scomparison, d_N/d_S values were analyzed through their distributions (as log d_N/d_S) between benchmark and predicted genes sets (Figure 4) and we did not detect statistically significant differences (Mann-Whitney test; $P = 0.16$). Therefore d_N/d_S similar distributions are indicative of similar high selective constraints and little or no positive selection on both benchmark and predicted genes sets, suggesting the functional properties of the canine gene predictions products involved are conserved.

To analyze the evolutionary rate of the new canine predicted gene sequences in a phylogenetic context we used the 232 mouse genes in addition to human genes and dog predicted genes to assess the levels of selective constraint of each lineage in comparison to the rest of the tree. In this way, differences or similarity in selective constraints can

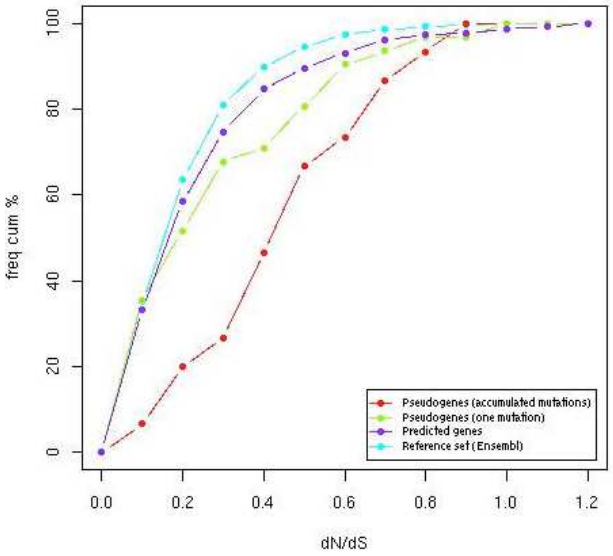


Figure 3
***D_N/d_S* cumulative frequency distribution of references, gene predictions and pseudogene predictions sets**. Benchmark, predicted genes, pseudogenes (with one mutation) and pseudogenes (with accumulated mutations) sets exhibit a median d_N/d_S of 0.15, 0.18, 0.22, 0.47, respectively, compared to their human functional orthologues. While the d_N/d_S distribution of pseudogenes with accumulated mutations sets is clearly shifted upwards to the theoretical value of 0.57 (average between 1.0 for no selection and 0.15 for selection from the benchmark set), the pseudogene set with one mutation is not significantly shifted suggesting this set may contains spurious pseudogene prediction. Predicted and benchmark gene sets have a similar d_N/d_S cumulative frequency distribution indicating comparable selective constraints level.

be predicted on all lineages within the phylogeny. For each of the 232 genes, we inferred the d_N and d_S values and calculated the d_N/d_S ratio. The median d_N/d_S for the dog lineage was found between human and mouse (Table 1), a result in agreement to these determined for 13,816 human, mouse and dog genes with 1:1:1 orthologs [2] with similar differences found across the three lineages.

Pseudogene predictions

Off the 412 missing genes, a subset of 55 predictions containing ORF-disrupting mutations lead to pseudogene identification. Among pseudogenes, we determined if protein sequences have different numbers of in-frame stop codons and/or frameshift disruptions. Using such quantitative measures, two mutation levels were apparent. A set of inactivated genes (n = 21) was predicted with accumulated mutations (mean = 4.2; range 2–11) and a second set (n = 34) was predicted with one mutation (Figure 3). To normalize the mutation rate by taking into account the coding sequence length, we expect proteins of similar lengths to now have similar numbers of stop-codons or a frameshift. We therefore examined the ratio of accumulation of ORF-disrupting mutations per 100 AA in both groups of pseudogenes. A mutation rate of 0.28 was determined for the group of pseudogenes with one mutation and a significant higher rate of 1.21 (Mann-Whitney test; P = 8.052e - 7) was found for the set of pseudogenes with accumulated mutations.

Although transcribed pseudogenes have been experimentally identified [43], a significant part of pseudogenes are thought to be transcriptionally silent in comparison to protein-coding genes. We thus searched for sequence alignment with canine transcript sequences (Unigene april 08 [37]) to assess the transcription activity of the pseudogene predictions with two and more mutations. We obtained alignment for 14%, a result in agreement with a recent report [44] showing that 19% of pseudogenes are the sources of novel RNA transcripts. These data indicate that the predicted pseudogenes are mostly undetected as expressed sequences in comparison to gene predictions with intact ORF (53%) and therefore significantly correspond to untranscribed pseudogenes [44].

Detecting nonfunctionality from dN/dS *analysis*

To assess independently of the presence of stop codons or frame-shifts, the validity of pseudogene predictions, we used the functionality test that uses the d_N/d_S ratio. Assuming a constant mutation rate, the d_N/d_S ratio between dog pseudogenes, for which a loss of function occurred, and their human functional orthologs should theoretically relax towards 0.57 (as the average of 1.0 in the absence of selection and 0.15 for negative selection as we calculated from the benchmark set) [10]. Thus, we calculated d_N/d_S ratio for each of the candidate pseudogene predictions in comparison to their human functional orthologous gene from pairwise transcripts pair alignments. For the pseudogene set with accumulated mutations, we calculated a median d_N/d_S of 0.50 indicating a considerable relaxation of selective constraints of the canine pseudogenes in comparison to their human functional orthologous (Figure 3 and Table 1). Furthermore, the d_N/d_S distributions obtained were shifted upwards in comparison to the benchmark set (Figure 4), which is significant to a Mann-Whitney test (P = 5.17e - 6), indicating relaxation of evolutionary constraints on the predicted pseudogenes. For the pseudogene set with one mutation, the median d_N/d_S of 0.18 was observed, suggesting no detectable differences in selective constraints between predicted pseudogenes from the canine sequence and their human functional counterparts. In addition, we analyzed whether the d_N/d_S ratio has an independent value before and after the stop codon among the predicted pseudogenes. In 26/28 instances, no significant differences were detected when

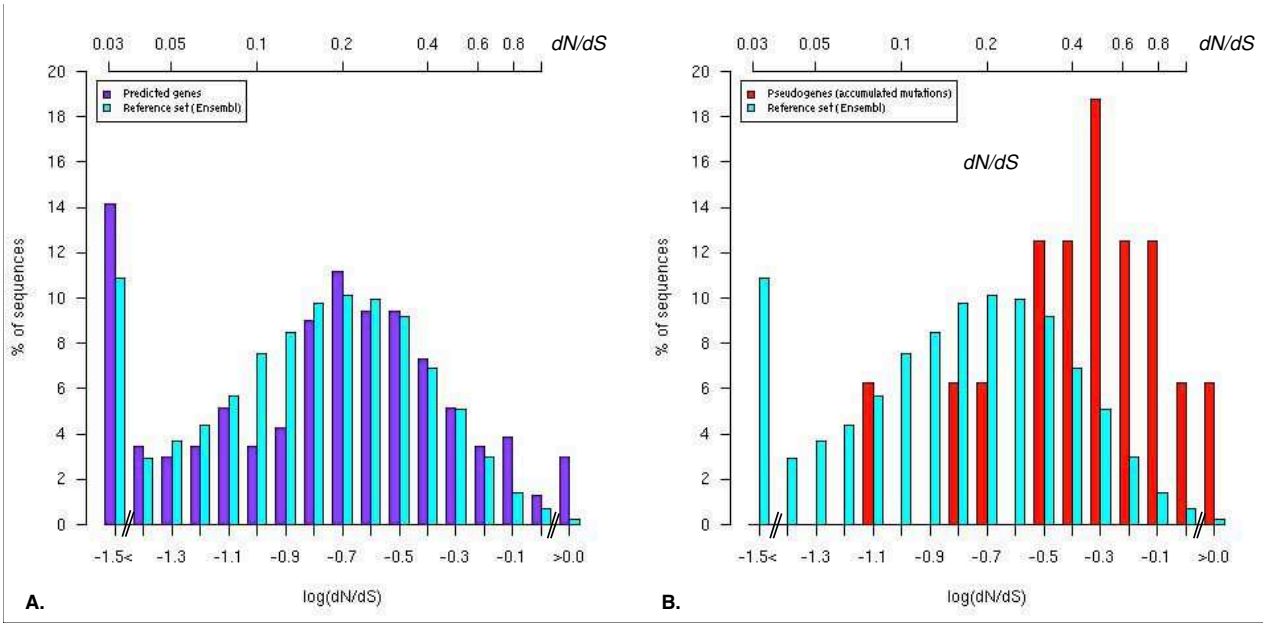


Figure 4

***D*$_N$/*d*$_S$ distributions of benchmark and test sets**. A. The d_N/d_S distribution (as log d_N/d_S) of the test set (new predicted genes) is represented in purple and benchmark set (human-dog 1:1 orthologous) is represented in blue. Test set exhibits a d_N/d_S distribution similar to the benchmark set (Mann-Whitney; $P = 0.16$) suggesting comparable selective constraints for both sets. B. In contrast the d_N/d_S distribution of the pseudogene (with accumulated mutations) set (red) is significantly shifted upwards (Mann-Whitney; $P = 5.17e - 6$) in comparison to the benchmark set, indicating relaxation of selective constraints on the predicted pseudogenes.

comparing d_N/d_S ratio for the two parts of each gene. In two cases, the d_N/d_S value before the stop was indicative of strong selective constraints (<0.1), in comparison to the value detected after the stop (>0.9), which suggest that the biological function may have been preserved.

We next searched to determine if the canine predicted pseudogenes showed any deviations from the expected rate of evolution using a phylogenetic context that includes human and mouse gene sequences. Such variation in rate may reflect relaxation of constraints in the dog lineage. The deviation between dog predicted pseudogenes with multiple mutations and the human and mouse lineages differs clearly (d_N/d_S = 0.41 for dog, 0.19 for mouse and 0.26 for human; Kruskal-Wallis test: $P = 1.04e - 2$) while no significant deviation ($P = 0.36$) was observed for the set of pseudogenes with one mutation (Table 2). We therefore retained the 21 pseudogene predictions with both the higher d_N/d_S value as characterized by pairwise and phylogenetic approaches and high mutation rate as gene loss candidates.

Gene loss identification

In addition to pseudogene identification, 11 gene predictions could not be detected with sufficient protein identity (average = 21.7%), both in the targeted genomic region

Table 1: Median and mean *dS* and *dN*/*dS* values of pseudogenes, predicted genes and reference set of human-canine orthologues

value	Pseudogenes with one mutation	Pseudogenes with several mutations	Predicted genes	Benchmark set 1:1 dog-human orthologs
dS median	0.45	0.44	0.39	0.39
dS mean	0.48	0.46	0.40	0.38
dN/*dS* median	0.18	0.50	0.19	0.15
dN/*dS* mean	0.28	0.50	0.26	0.20

Table 2: Evolutionary constraints (*dS* and *dN*/*dS*) for 1:1:1 orthologs among human, mouse and dog

dN/*dS* median	Predicted genes	Pseudogenes with several mutations	Pseudogenes with one mutation
Human	0.21	0.26	0.19
Dog	0.17	0.41	0.16
Mouse	0.15	0.19	0.13

(COIL) and in the whole canine sequence. For these predictions with no readily identifiable counterparts in dog, we searched for sequence alignment with canine expressed sequences (Unigene april 08) to address the underlying assumption that genes are not transcribed when placed in the context of highly degraded sequence. We identified sequence alignment in only three cases. These results showed that the gene predictions with poor sequence similarity were largely undetected as expressed sequences in comparison to gene predictions with intact ORF.

For the last subset of 49 canine genes that remained undetected in this study, we address the possibility that gene predictions could have been prevented because of a gap in the canine sequence assembly. We searched for gap content in the COILs that lack canine orthologous genes. For 12 COILs, the gap content was found to account for >10% of the total size of the COIL, seven-fold more than a random expectation set (n = 1000, gap = 1.32%) and manual inspection of sequence content resulted in identifying multiple sequence gaps. The 12 missing genes in those short targeted regions were therefore not retained in further analysis. Based on these results, a total of 37 undetected genes was considered and merged with the 11 gene predictions that could not be detected with sufficient protein identity and the 21 pseudogenes into a single set (n = 69) of gene loss candidates for further analyses (Figure 2 and Additional data file 5).

Evolutionary scenarios of the canine gene losses

Do we detect losses of genes that occur specifically in the dog or do such losses occur in other mammalian lineages as well? If so, do such losses correspond to the time the dog branch diverged from the Euarchontoglires (rodent/primate) lineage? One way to analyze these possibilities is to determine their phyletic pattern using ten species from chicken to human and to define the amount of time between gene origin and present. The timing of genes origin was defined by searching for 1:1 orthologs between human and nine species. In addition to human, chimp, mouse and rat genome sequence assemblies, we used scaffold assemblies of elephant, tenrec and armadillo from the Afrotheria and Xenarthra superorder and two non-placental genome assemblies of opossum and platypus. We also included the chicken sequence to infer gene origins that occurred as long as 310 million years ago (MYA) (Figure 5).

Orthologous genes were detected between human and all species (except dog) for 11 genes. Therefore, they have an origin that occurred before the separation of the mammals and birds lineages and have been functional for 310 million years (My). In addition, 17 genes were identified in all species of the opossum/platypus, elephant-tenrec-armadillo and Euarchontoglires branches, a period of 170 My, 17 in all species of the elephant-tenrec-armadillo and Euarchontoglires branches (100 My), and 10 in Euarchontoglires only (87 My) (Figure 5) [45].

Overall, 28 canine gene losses could be characterized as being functional in other species for more than 170 My

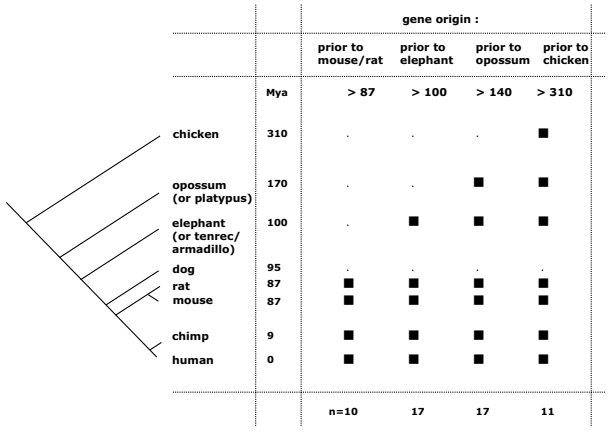


Figure 5
Gene origin timing. Timing of gene origin is assessed by determining the one-to-one orthologs between human and nine species listed on the left side of the figure. The species belong to Euarchontoglire (Primates and rodents), Xenarthra (Armadillo), Afrotheria (elephant and tenrec), Marsupial and Monotreme (opossum and platypus). Time of species divergence from the lineage leading to human is shown in MYA (million years ago). Filled squares represent the presence of the ortholog in the species. Numbers at the bottom of the figure denote the number of genes that display the presence/absence pattern across species.

and 10 genes were not detected before 87 My and therefore specifically arose in rodent and primate lineages. For these genes, postulating that they arose through duplication events of a parental gene, we searched for paralogs among all human genes. For seven genes (*ZNF426*, *WFDC12*, *ZIK1*, *HLA-SX-alpha*, *PNMA5*, *PNMA3*, *ZNF251*) we identified at least one paralog (sequence identity >30%) in the close vicinity of the parental gene (mean of 71 kb; range: 16–128 kb).

We further used the Ensembl reconciliation tree method [46] for checking possible duplication events specific of the primates and rodents lineages. Indeed, assuming that all homologous genes are known, the reconciliation of the gene tree with the species tree allows to distinguish duplication from speciation events and therefore orthologous from paralogous genes. Five genes (*ZNF426*, *ZIK1*, *HLA-SX-alpha*, *PNMA5*, *PNMA3*) have in-paralogs in the reference species suggesting a pattern of duplication event (Additional data file 6).

These results suggest that tandem duplication events have occurred and lead to specific in-paralogs in the branch leading to human species. Another contribution of this analysis is that it permits identification of 10 losses that occur in several lineages indicating multiple and independent gene loss events [47].

Functional characteristics of gene losses

For the 28 canine-specific gene losses that have been functional for more than 170 My, we determined the functional annotation of the human genes using WebGestalt, a Web-based gene set analysis toolkit [48]. The classification using the GOTree sub-module includes seven genes that belong to the biological process of response to stimulus with *PROZ*, a vitamin K-dependent protein Z precursor involved in blood coagulation pathway and *SERPINA10* a protein Z-dependant protease inhibitor that regulates factor Xa involved in blood coagulation. Moreover, it includes five genes involved in response to stimulus pathways that play a role in sensory function such as *UGT2A* which encodes an enzyme with transferase activity that may catalyze inactivation and facilitate elimination of odorants, *OR1Q1*, *OR1B1*, *ORN1* which arethree olfactory receptors, and Noggin, a secreted polypeptide encoded by the *NOG* gene that appears to have pleiotropic effect, both early in development as well as in sensory perception of sound. Other genes of interest belong to families with at least six members such as *TBX22* a transcription factor involved in the regulation of various aspects of embryonic development, in particular cell type specification and regulation of morphogenetic movements [49], and *MS4A3* which is a subset of the superfamily of tetraspan transmembrane protein encoding genes. Several genes were classified with function involved in DNA repair, apoptosis and tumor formation such as *BOK* which encodes a Bcl-2 related protein and *PDE1B* which may play a role in apoptosis. To address the question of which tissue might be significantly affected by gene loss, we determined a gene-expression profile characterization per tissue based on the occurrence frequency of the ESTs profiles of human genes corresponding to the gene lost set using the tissue expression profile sub-module of WebGestalt. Testis-expressed gene expression profiles showed a significant over or under representation and, to a lesser extent, expression profiles related to placenta and kidney tissues did as well (Additional data file 7).

Discussion

This study describes a multispecies comparative genomics approach that provides a methodology for improving genes prediction and detecting putative gene losses. When coupled to a strategy of phyletic pattern analysis, the approach allows differentiation of species-specific gene loss from multiple independent gene loss. Here, focusing on genes that were not detected in the whole-genome assembly of the dog but annotated in four rodents and primates species, we identified 232 new gene and we predicted 69 canine gene loss candidates of which 21 are identified as pseudogenes,

Targeted gene prediction: strengths and limitations

A basic application of gene order-based approaches is the capacity to detect short conserved genomic context based on robust orthologous gene pair annotation. Therefore, results are limited by the source of gene annotation. In this study, we used the Ensembl annotation because of its good gene prediction coverage of the four species used as reference genomes. Since annotation of mammalian genome is a continuous process, our gene order-based approach may be improved over the course of time.

The use of short orthologous genomic intervals filtering has been well documented [28]. First, it reduces the cost of detecting false-positives as it filters out paralogs, with the exception of those caused by tandem gene duplication, and alignments to processed pseudogenes. Second, it allows a balance between sequence alignment sensitivity versus accuracy [50]. Alternatively, for more divergent sequences, alignment criteria may be relaxed in short predefined space where the background noise is significantly reduced compared to a genome scale search.

In our analysis, predictions may not provide an exhaustive list of gene predictions as inaccuracies may be generated by sequence artifacts that typically exist in draft sequence assemblies. Another issue related to prediction accuracy is the unexpected and unknown level of highly divergence at the nucleotide level. While scenarios of functional

sequences with different evolutionary rate in different species exist [51], we postulated that using protein coding genes with a comparable evolutionary rate amongst four reference species reduces the possibility that a gene evolves independently in the dog species.

Computational prediction of gene loss

A corollary to targeted gene prediction is that the absence of prediction strongly predicts gene relocation to a different region or chromosome or a gene loss event. Gene losses arise through retrotransposition or segmental or tandem duplications events followed by inactivation of one copy, or by degenerative mutations. We used a computational analysis to identify genes lost as pseudogenes based on various detrimental sequence mutations such as in-frame stop codons and frameshifts causing or resulting from loss of function. In this study, pseudogenes were separated in two groups, with the group of pseudogenes with one mutation (showing a low mutation rate) and the second group with an elevated mutation rate (>4 mutations, on average). Pseudogene predictions with one mutation could be overstated due to sequence artifacts that exist in the assembly. Indeed, stop codons and frameshifts are accommodated by algorithm like GeneWise. Other programs specifically designed for aligning pseudogenes such as GeneMapper [52] may be useful for addressing this problem. Another hypothesis is that pseudogene predictions have existed as pseudogenes (i.e. inactivated) for different amounts of time in the carnivore lineage. The formation of pseudogenes present in the canine genome could have been initiated by different or multiple events rather than have resulted from a continuous process over the course of time. Pseudogene characterization through the ratio of silent to replacement nucleotide substitution rates (*dN*/*dS*) may be a good indicator of changes in selective constraint that tend to be recent [53]. It is clear from our analysis that the *dN*/*dS* approach is useful to assess the evolutionary constraints that occur on nucleotide substitution. However, inferences of selection need to be treated with extreme caution.

Functional impact of gene loss

We identified 28 gene losses that have been functional for more than 170 million years, a time period that extends from platypus to human (Figure 5). Losses of gene in a given species can be considered an adaptive event that may confer selective advantages to an organism [54]. Similarly to neutral losses, adaptive losses occurring ~95 MYA (for lineage leading to canid) are expected to leave genomic signatures with ORF-disrupting sequence mutations accumulation due to sequence degeneration. Here, the losses identified are based on ORF-disrupting sequence mutations, absence of EST validation and absence of significant similarity at the protein level. Although highly speculative, one hypothesis is that species-specific gene loss may confer a selective advantage in dog. Among the gene losses we identified were *PROZ*, a vitamin K-dependent protein Z precursor gene involved in response to stimulus that plays a role in blood coagulation. Mammalian blood coagulation is initiated and regulated by a complex network of interactions involved in normal hemostasis. Interestingly, Lindberg *et al*. describes a decrease of the expression of heme and globin related genes that correlate with the tameness trait in silver foxes suggesting that differences in behavior have a genetic basis [55]. A second hypothesis, is that gene loss may be a direct reflection of the loss of redundancy, where functionally overlapping genes cover for the loss of function as for genes involved in sensory functions [56,57].

Conclusion

Among mammals, one-to-one orthologous correspondence can be defined for a large part of gene repertoires. Complex homologous relationships such as one-to-zero and many-to-many ones remain to be deciphered within gene families, for genes with divergent sequence as well as for species-specific genes that have emerged or have been lost through evolution. The combination of multispecies comparative genomics with in-depth gene prediction, accurate consideration of phylogenetic relationship, and timing of gene origin events can predict both gene structure and gene losses in newly sequenced genomes. This, in turn, enhances the integrity of reference genomes. The end result is a higher quality product for all sequenced genomes, regardless of the depth of sequence. We aim to see this approach applied to many other model organisms, thus enhancing the utility of the new sequencing resources throughout the comparative genomics community.

Methods

Gene datasets

Biomart [58] version 0.5 (Ensembl v.42) was used to collect orthologous protein-coding genes from the five genomes of interest: human (NCBI 36), chimp (Chimp 2.1), mouse (NCBI m36), rat (RGSC 3.4) and dog (CanFam 2.0). Ensembl Gene Id, orthologous relationships, locations in base pair for each species were downloaded and deposited into a MySQL database (v.4.1.12). The set of 412 protein-coding genes not annotated on the dog genome assembly with a 1:1:1:1:0 Human:Chimp:Mouse:Rat:Dog match was then extracted from the MySQL database.

Synteny maps

We used the program AutoGRAPH [13] to construct pairwise synteny maps between reference genomes and tested genome. AutoGRAPH has been designed to construct syn-

teny maps using genomic coordinates of ortholog pairs. The program transposes genomic coordinates into sequence of ordinal numbers and positions genes on an ordinal scale in relation to others on their respective chromosomes. Conserved segments ordered (CSO) can then be identified with respect to the ranking order. We only considered CSO containing a minimum of three genes. AutoGRAPH inferred the collinearity rate within CSO corresponding to the longest increasing gene order sequence between the two species divided by the total number of orthologs. We discarded CSO that had a collinearity rate less than 0.5. All synteny maps (n = 88) built in this work are presented in Additional data file 1 and can be downloaded.

Gene structure prediction

The GeneWise program [6] (wise2-2-0) was used with default parameters to align each reference protein on the dog COIL forward and reverse strands (option -both) sequence. Predictions were post processed to pick up the highest genewise prediction, to compute sequence identity/similarity against reference proteins and to analyze splice sites conservation. Only predictions exhibiting at least 40% identity with reference proteins were retained. GeneWise was also used with the Hidden Markov model that uses HMM profiles generated with the HMMER package [59]. HMM-based prediction considers exons, introns and UTR regions as different states of gene structure that occupy subsequences of a sequence. A gene structure can be considered as an ordered set of state/sub-sequence pairs. A HMM-based prediction is considered as a predicted gene structure if probability of generating a gene structure is maximal over all possible states. Dynamic programming method for finding an optimal parse, or the best sequence of states has [10] been computed with the HMMER package.

Homology searches

Reference transcript sequences were collated from Ensembl (v.42) and aligned against the canine sequence assembly (CanFam2) with the program Exonerate v1.2 [35]. Exonerate includes various models for aligning splice sites, combining speed and accuracy. We used the est2genome model, with a minimum perfect match of 18 bases to trigger alignments (dnawordlen 18). For each reference transcript, we retained the best five matching sequences.

Canine proteins inferred from the gene predictions were aligned against all canine transcripts with Exonerate using the coding2coding model. Canine predicted proteins were aligned on canine dbEST (est.fa 05/19/07 from UCSC) and UNIGENE (April 2008) using Exonerate with the protein2genome model.

The protein three-dimensional structure was available for 21 human genes. The sequences were retrieved via the Protein Data Bank. The amino sequences for the corresponding canine predictions were obtained from the genewise program prediction. Canine-human comparative modelling was determined using the SWISS-MODEL server [39]. Amino acid sequences are aligned between the primary structure of the human and the canine sequence. The three-dimensional model is constructed through the process implemented in the SWISS-MODEL server.

DN/dS *analysis*

DN/*dS* analyses were conducted using the maximum-likelihood-based CODEML program (model = 0; PAML package) [60]. Sequence alignments of the whole coding region of the human orthologous sequence with canine prediction were realized with clustalW program. *Ds* values were calculated from pairwise alignments using all transcripts. To filter for possible inconsistencies among orthologous trancripts, we selected the transcript with the smallest phylogenetic distance using the smallest *dS*. For each dataset, we calculated a threshold on *dS* which two fold the median *dS*; all *dS* larger than this threshold were not used for the *dN*/*dS* calculation. *DN*/*dS* values of the benchmark set were extracted from Ensembl. *DN*/*dS* ratio in the phylogenetics context were calculated using CODEML program using the branch model set as model = 1 and run mode = 0. Sequence alignments of the whole coding region of the human, mouse and canine prediction orthologous sequence were realized with clustalW program

Gene Ontology annotation

The Gene Ontology Tree Machine (GOTM) and WebGestalt programs [31,48] were used to retrieve GO term associated with ensembl gene ID. A hypergeometric test computes the statistical significance of overrepresentations of GO term compared to a reference complete list of genes. Only GO terms that were significantly over-represented ($P < 1.0e - 3$) were considered.

Determining gene origin

For each of the 69 candidate gene losses, one-to-one orthologous gene was searched between human and nine species using the complete collection of orthologous protein-coding genes (Ensembl). Genome sequence assemblies were used for human, chimp, mouse, rat, monodelphis, platypus and chicken and scaffold assemblies for elephant, tenrec and armadillo. Timing of gene origin was inferred by determining the longest serie of one-to-one orthologs between the human and each of the nine species.

P value calculation

We used the R package (R Development Core Team 2006. R: A language and environment for statistical computing. http://www.R-project.org) to test the statistical significance in comparing distinct distributions at each step of the method (Mann-Whitney, Kruskal-Wallis and Student's test).

Abbreviations

ESTs: Expressed Sequence Tag; dbEST: database of EST; ORF: Open Reading Frame.

Authors' contributions

TD prepared the data, carried out the comparative data analysis and contributed to the writing of the manuscript, JT worked on gene prediction analysis, AV carried out dN/dS analysis, CA participated in study design, EAO provided feedback throughput, suggested various analysis and worked on all drafts of the paper, FG participated in the data interpretation, and contributed to the writing of the manuscript, CH conceived of the study, participated in the data analysis and interpretation, and contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Human-dog synteny map: Example of human chromosome 5. *An example of the synteny map built between human chromosome 5 and the dog genome.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S1.pdf]

Additional file 2

Synteny maps characteristics. *The data indicates the main characteristics of the synteny maps.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S2.pdf]

Additional file 3

Characterization of Consensus Orthologous IntervaLs (COILs) containing missing genes. *These data file lists the characteristics of the Consensus Orthologous Intervals.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S3.pdf]

Additional file 4

List of the 232 new predicted canine genes. *This table lists the 232 new gene predictions using the human gene identifiers from Ensembl.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S4.pdf]

Additional file 5

List of the 69 candidate gene losses. *This table lists the gene losses using the human gene identifiers from Ensembl.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S5.pdf]

Additional file 6

Gene/species tree reconcilation. *These data provide the gene/species tree reconcilation that show the possible duplication events specific of the primates and rodents lineages.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S6.pdf]

Additional file 7

Gene-expression profile characterization per tissue with significant over and under representation. *The data provided show gene-expression profile characterization per tissue.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-62-S7.pdf]

Acknowledgements

We are grateful to Roderic Guigo and to the reviewers for providing useful suggestions and helpful comments. We thank the OUEST-genopole bioinformatics plate-form for technical help and assistance. We acknowledge for support the Centre National de la Recherche Scientifique (JT, AV, CA, FG and CH) and the Conseil Régional de Bretagne for supporting TD with a fellowship and the Intramural Program of the National Institutes of Health (EAO).

References

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915):**520-562.
2. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, *et al.*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438(7069):**803-819.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822):**860-921.
4. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, *et al.*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982):**493-521.
5. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nat Rev Genet* 2008, **9(1):**62-73.
6. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14(5):**988-995.
7. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17(Suppl 1):**S140-148.
8. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R: **Comparative gene prediction in human and mouse.** *Genome Res* 2003, **13(1):**108-117.
9. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G: **Tree pattern matching in phylogenetic trees: automatic**

search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 2005, **21(11):**2596-2603.
10. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLoS Comput Biol* 2006, **2(9):**e133.
11. Lunter G, Ponting CP, Hein J: **Genome-wide identification of human functional DNA using a neutral indel model.** *PLoS Comput Biol* 2006, **2(1):**e5.
12. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, *et al.*: **An RNA gene expressed during cortical development evolved rapidly in humans.** *Nature* 2006, **443(7108):**167-172.
13. Derrien T, Andre C, Galibert F, Hitte C: **AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps.** *Bioinformatics* 2007, **23(4):**498-499.
14. Peng Q, Pevzner PA, Tesler G: **The fragile breakage versus random breakage models of chromosome evolution.** *PLoS Comput Biol* 2006, **2(2):**e14.
15. Tesler G: **GRIMM: genome rearrangements web server.** *Bioinformatics* 2002, **18(3):**492-493.
16. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci USA* 2007, **104(49):**19428-19433.
17. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, *et al.*: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.** *Proc Natl Acad Sci USA* 2003, **100(3):**1140-1145.
18. Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, Comstock CL, Davis C, Ewing B, Oommen S, Lau C, *et al.*: **Targeted discovery of novel human exons by comparative genomics.** *Genome Res* 2007, **17(12):**1763-1773.
19. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, *et al.*: **Evolution of genes and genomes on the Drosophila phylogeny.** *Nature* 2007, **450(7167):**203-218.
20. Heger A, Ponting CP: **Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes.** *Genome Res* 2007, **17(12):**1837-1849.
21. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, *et al.*: **Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes.** *Genome Res* 2007, **17(12):**1823-1836.
22. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL: **InParanoid 6: eukaryotic ortholog clusters with inparalogs.** *Nucleic Acids Res* 2008:D263-266.
23. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18(12):**619-620.
24. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, *et al.*: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps.** *Science* 2005, **309(5734):**613-617.
25. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA: **The promise of comparative genomics in mammals.** *Science* 1999, **286(5439):**458-462.
26. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5(4):**299-310.
27. Bhutkar A, Russo SM, Smith TF, Gelbart WM: **Genome-scale analysis of positionally relocated genes.** *Genome Res* 2007, **17(12):**1880-1887.
28. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24(5):**637-644.
29. van Baren MJ, Brent MR: **Iterative gene prediction and pseudogene removal improves genome annotation.** *Genome Res* 2006, **16(5):**678-685.
30. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2008:D707-714.
31. Zhang B, Schmoyer D, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.** *BMC Bioinformatics* 2004, **5:**16.
32. Goodstadt L, Heger A, Webber C, Ponting CP: **An analysis of the gene complement of a marsupial, Monodelphis domestica: evolution of lineage-specific genes and giant chromosomes.** *Genome Res* 2007, **17(7):**969-981.
33. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Res* 2003, **13(1):**37-45.
34. Alekseyev MA, Pevzner PA: **Are there rearrangement hotspots in the human genome?** *PLoS Comput Biol* 2007, **3(11):**e209.
35. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.
36. Meyer IM, Durbin R: **Gene structure conservation aids similarity based gene prediction.** *Nucleic Acids Res* 2004, **32(2):**776-783.
37. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, *et al.*: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31(1):**28-33.
38. Mulder N, Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison.** *Methods Mol Biol* 2007, **396:**59-70.
39. Arnold K, Bordoli L, Kopp J, Schwede T: **The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.** *Bioinformatics* 2006, **22(2):**195-201.
40. Luthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356(6364):**83-85.
41. Yu P, Ma D, Xu M: **Nested genes in the human genome.** *Genomics* 2005, **86(4):**414-422.
42. Torrents D, Suyama M, Zdobnov E, Bork P: **A genome-wide survey of human pseudogenes.** *Genome Res* 2003, **13(12):**2559-2567.
43. Yao A, Charlab R, Li P: **Systematic identification of pseudogenes through whole genome expression evidence profiling.** *Nucleic Acids Res* 2006, **34(16):**4477-4485.
44. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, *et al.*: **Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.** *Genome Res* 2007, **17(6):**839-851.
45. Hallstrom BM, Janke A: **Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations.** *BMC Evol Biol* 2008, **8:**162.
46. Page RD, Charleston MA: **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.** *Mol Phylogenet Evol* 1997, **7(2):**231-240.
47. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.** *Genome Res* 2003, **13(10):**2229-2235.
48. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005:W741-748.
49. Tada M, Smith JC: **T-targets: clues to understanding the functions of T-box proteins.** *Dev Growth Differ* 2001, **43(1):**1-11.
50. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14(5):**942-950.
51. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET: **Fast-evolving noncoding sequences in the human genome.** *Genome Biol* 2007, **8(6):**R118.
52. Chatterji S, Pachter L: **Reference based annotation with GeneMapper.** *Genome Biol* 2006, **7(4):**R29.
53. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ: **Comparisons of dN/dS are time dependent for closely related bacterial genomes.** *J Theor Biol* 2006, **239(2):**226-235.
54. Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D: **Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage.** *PLoS Comput Biol* 2007, **3(12):**e247.
55. Lindberg J, Bjornerfeldt S, Bakken M, Vila C, Jazin E, Saetre P: **Selection for tameness modulates the expression of heme related genes in silver foxes.** *Behav Brain Funct* 2007, **3:**18.
56. Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS: **Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss.** *Mol Syst Biol* 2007, **3:**86.

57. Hughes T, Liberles DA: **The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.** *J Mol Evol* 2007, **65(5):**574-588.
58. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14(1):**160-169.
59. Eddy SR, Mitchison G, Durbin R: **Maximum discrimination hidden Markov models of sequence consensus.** *J Comput Biol* 1995, **2(1):**9-23.
60. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.

OPEN ACCESS Freely available online PLoS GENETICS

Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping

Amaury Vaysse[1☯], Abhirami Ratnakumar[2☯], Thomas Derrien[1], Erik Axelsson[2], Gerli Rosengren Pielberg[2], Snaevar Sigurdsson[3], Tove Fall[4], Eija H. Seppälä[5], Mark S. T. Hansen[6], Cindy T. Lawley[6], Elinor K. Karlsson[3,7], The LUPA Consortium, Danika Bannasch[8], Carles Vilà[9], Hannes Lohi[5], Francis Galibert[1], Merete Fredholm[10], Jens Häggström[11], Åke Hedhammar[11], Catherine André[1], Kerstin Lindblad-Toh[2,3], Christophe Hitte[1], Matthew T. Webster[2]*

1 Institut de Génétique et Développement de Rennes, CNRS-UMR6061, Université de Rennes 1, Rennes, France, **2** Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, **3** Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **4** Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden, **5** Department of Veterinary Biosciences, Research Programs Unit, Molecular Medicine, University of Helsinki and Folkhälsan Research Center, Helsinki, Finland, **6** Illumina, San Diego, California, United States of America, **7** FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, **8** Department of Population Health and Reproduction, School of Veterinary Medicine, University of California Davis, Davis, California, United States of America, **9** Department of Integrative Ecology, Doñana Biological Station (CSIC), Seville, Spain, **10** Faculty of Life Sciences, Division of Genetics and Bioinformatics, Department of Basic Animal and Veterinary Sciences, University of Copenhagen, Frederiksberg, Denmark, **11** Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

Abstract

The extraordinary phenotypic diversity of dog breeds has been sculpted by a unique population history accompanied by selection for novel and desirable traits. Here we perform a comprehensive analysis using multiple test statistics to identify regions under selection in 509 dogs from 46 diverse breeds using a newly developed high-density genotyping array consisting of >170,000 evenly spaced SNPs. We first identify 44 genomic regions exhibiting extreme differentiation across multiple breeds. Genetic variation in these regions correlates with variation in several phenotypic traits that vary between breeds, and we identify novel associations with both morphological and behavioral traits. We next scan the genome for signatures of selective sweeps in single breeds, characterized by long regions of reduced heterozygosity and fixation of extended haplotypes. These scans identify hundreds of regions, including 22 blocks of homozygosity longer than one megabase in certain breeds. Candidate selection loci are strongly enriched for developmental genes. We chose one highly differentiated region, associated with body size and ear morphology, and characterized it using high-throughput sequencing to provide a list of variants that may directly affect these traits. This study provides a catalogue of genomic regions showing extreme reduction in genetic variation or population differentiation in dogs, including many linked to phenotypic variation. The many blocks of reduced haplotype diversity observed across the genome in dog breeds are the result of both selection and genetic drift, but extended blocks of homozygosity on a megabase scale appear to be best explained by selection. Further elucidation of the variants under selection will help to uncover the genetic basis of complex traits and disease.

Citation: Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, et al. (2011) Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLoS Genet 7(10): e1002316. doi:10.1371/journal.pgen.1002316

Editor: Joshua M. Akey, University of Washington, United States of America

Received February 1, 2011; **Accepted** July 30, 2011; **Published** October 13, 2011

Funding: This research was mainly supported by The LUPA Consortium, which is a Collaborative Research Project funded by the European Commission under the 7th Research Framework Programme (www.eurolupa.org). The consortium consists of research groups in more than 20 European institutes, with the goal of unraveling the genetic basis of inherited disease in dogs with relevance to human health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: matthew.webster@imbim.uu.se

☯ These authors contributed equally to this work.

Introduction

There are more than 400 breeds of domestic dog, which exhibit characteristic variation in morphology, physiology and behavior. This astonishing phenotypic diversity has been molded by two main phases of evolution: 1) the initial domestication from wolves more than 15,000 years ago, where dogs became adapted to life in closer proximity to humans and 2) the formation of distinct breeds in the last few hundred years, where humans chose small groups of dogs from the gene pool and strongly selected for novel and desirable traits [1,2]. A by-product of these processes has been that many dog breeds suffer from a high incidence of inherited disorders [3,4].

Its unique population history makes the dog an ideal model organism for mapping the genetic basis of phenotypic traits due to extensive linkage disequilibrium (LD) and a reduction in

Author Summary

There are hundreds of dog breeds that exhibit massive differences in appearance and behavior sculpted by tightly controlled selective breeding. This large-scale natural experiment has provided an ideal resource that geneticists can use to search for genetic variants that control these differences. With this goal, we developed a high-density array that surveys variable sites at more than 170,000 positions in the dog genome and used it to analyze genetic variation in 46 breeds. We identify 44 chromosomal regions that are extremely variable between breeds and are likely to control many of the traits that vary between them, including curly tails and sociality. Many other regions also bear the signature of strong artificial selection. We characterize one such region, known to associate with body size and ear type, in detail using "next-generation" sequencing technology to identify candidate mutations that may control these traits. Our results suggest that artificial selection has targeted genes involved in development and metabolism and that it may have increased the incidence of disease in dog breeds. Knowledge of these regions will be of great importance for uncovering the genetic basis of variation between dog breeds and for finding mutations that cause disease.

haplotype diversity due to genetic drift in isolated populations [3-5]. Another major advantage of the canine model is that much of the variation in morphological characteristics in dogs appears to be governed by a relatively small number of genetic variants with large effect [6]. This is likely because novel variants with large effects are preserved by artificial selection. This is in strong contrast to humans where morphological variation in traits such as height appears to be controlled by hundreds of loci with small effects, which have proven extremely difficult to catalogue [7]. Identifying the targets of artificial selection in dog breeds is therefore an extremely promising approach for identifying genetic variants involved in phenotypic variation, which could greatly facilitate the identification of similar variants and novel molecular pathways in humans.

Several loci have now been identified that control variation in morphological traits between dog breeds. In some cases, variation in a trait occurs within a breed, and long blocks of LD can be used to identify the locus responsible using genome wide association studies (GWAS). Using this approach loci involved in traits including size (IGF1) [8], coat type (RSPO2, FGF5, KRT71) [9] and coat color (MITF, CBD103) [10,11] were identified in single breeds, and it was shown that variation in these loci is also correlated with phenotypic variation between breeds. An alternative approach, when a particular trait is shared by several breeds, is to perform across-breed GWAS. In general, levels of LD decay much faster between breeds, and this reduces the power to detect association [11]. However, selection acts to fix long haplotypes bearing the causative variant, thus increasing levels of LD between breeds in regions under selection. Jones *et al.* [12] used a sparse marker set and across-breed GWAS to identify correlations with a number of morphological traits, such as size, height, and shape of ears, snout and limbs, which was further refined by Boyko *et al.* [6] using 80 dog breeds and ~61,000 SNPs. Across-breed GWAS have also been used to identify an FGF4 retrogene associated with chondrodysplasic breeds [13] and the THBS2 locus associated with brachycephalic breeds [14].

Genomic regions with a high degree of genetic differentiation between breeds are also indicative of selection. A large proportion of SNPs with high F_{ST} between dog breeds are found in loci associated with phenotypic traits such as size, ear morphology and coat color [6]. Akey *et al.* [15] scanned patterns of variation in 10 dog breeds and ~21,000 SNPs using a 1 Mb sliding windows to identify larger regions with elevated F_{ST} in particular breeds. This scan identified many regions likely to be under selection in one or more of the breeds in their dataset. Notably, a highly differentiated interval in Shar-Pei on chromosome 13 contains the HAS2 gene and is likely associated with the wrinkled skin phenotype of this breed [15,16].

Although a large number of loci under selection have now been identified, the genetic basis of much of the phenotypic variation in dog breeds and particularly behavioral traits remains unexplained. One drawback of previous studies is the use of SNP arrays with relatively low coverage of the genome. With the development of a new high-density array it is now possible to examine the dog genome at much higher resolution, allowing a comprehensive characterization of regions under selection. Genetic variants under selection in dogs can be loosely divided into two categories: 1) those that control variation in common traits such as size and ear carriage, which segregate across many breeds [6,8] and 2) those that encode rare traits that present in one or a few breeds, such as brachycephaly, chondrodysplasia and skin wrinkling [13,14,16].

Here we implement a variety of approaches to identify both these types of loci. In cases where a common trait has been identified, it is possible to search for genotype-phenotype correlations. We attempt to identify both behavioral and morphological traits that vary between breeds using across-breed GWAS. We also use F_{ST} statistics to identify additional SNPs that have high variability in frequency between breeds. These methods identify known loci and indicate new regions that may be involved in common trait variation.

The action of selection can potentially be identified by examining patterns of variation in individual breeds in order to detect the characteristic signature of selective sweeps. This signature is characterized by the presence of long haplotypes, a skew in allele frequency, reduced heterozygosity, and elevated population differentiation. A large number of statistical methods have been developed to detect sweeps based on these different patterns [17-22]. The formation of dog breeds occurred during an extremely brief evolutionary time, and likely involved rapid fixation of haplotypes under strong artificial selection. Under this scenario, simulations suggest that statistics based on F_{ST} and differences in heterozygosity are likely to be most powerful. [23]. Furthermore, dog breeds are known to be characterized by extensive LD and limited haplotype diversity, including long blocks of homozygosity, which reflect the action of population bottlenecks and selective breeding. This suggests that tests based on allele frequency spectrum and haplotype length will be of limited applicability, as many genomic regions are essentially devoid of genetic variation. We therefore base our approach to identify selective sweeps on pairwise comparisons of both F_{ST} and heterozygosity between breeds.

The presence of long blocks of homozygosity in the dog genome [1,11] is likely to reflect the action of both selection and genetic drift. We therefore conduct extensive coalescent simulations in order to distinguish between these processes. These simulations incorporate a realistic model of dog population history under neutrality to provide null distributions to compare with the real data. We also conduct a comprehensive characterization of SNP variation in a 3 Mb region encompassing several loci with extreme population differentiation that are associated with at least two morphological traits.

Results

High-density canine array design and evaluation

Our first goal was to develop a high-density, high-accuracy mapping array with uniform SNP coverage across the whole genome. Since the SNP map from the canine genome project, although containing >2.8 million SNPs at fairly even coverage, still contained gaps, we first performed targeted resequencing within 1,555 regions that lie within intervals >40 kb containing no known SNPs in unique sequence. We performed Roche NimbleGen array capture to enrich these regions followed by sequencing using the Illumina Genome Analyzer on 4 pools containing multiple samples of a single dog breed (Irish Wolfhounds, West Highland White Terrier, Belgian Shepherds and Shar-Pei) and one pool of wolf samples. In total, we discovered 4,353 additional high-quality SNPs using this method. We selected SNPs from this improved map to form the "CanineHD" array panel. We generated an initial panel of 174,943 SNPs that were included on the array of which 173,622 (99.2%) give reliable data. These loci are distributed with a mean spacing of 13 kb and only 21 gaps larger than 200 kb. Loci with unreliable SNP calls, potentially due to copy number polymorphism, were not included in the analysis. In total, 172,115 are validated for SNP genotyping and 1,547 are used only for probe intensity analyses. This is a significant improvement compared with the largest previously existing array, which has 49,663 well performing SNPs, with a mean spacing of 47 kb and 1,688 gaps larger than 200 kb. Figure S1 shows the distribution of SNPs in 100 kb windows across the genome. The improvement in coverage is particularly striking on the X chromosome, where >75% of 100 kb windows contain no SNPs on the previous array, but <5% of windows do not contain SNPs on the CanineHD array.

Of all the SNPs on the array, 0.9% are novel SNPs discovered by the targeted resequencing experiment. The remaining SNPs have been previously described: 65.1% of them were present in a comparison of the boxer reference genome with a previously sequenced poodle, 21.7% were present in alignments of low coverage sequencing reads from various dog breeds to the boxer reference genome, 25.4% were present within the boxer reference and 1.2% were present in alignments of wolf and/or coyote sequencings with the reference boxer genome. There is therefore a bias in the way that SNPs were ascertained: all of them were identified in a comparison involving the boxer reference assembly. However this has not had a great impact on the number of SNPs polymorphic in different breeds (see below). The array was initially evaluated using 450 samples from 26 breeds termed the "Gentrain" dataset. Within this dataset, average call rates were 99.8% and reproducibility and Mendelian consistency were both >99.9%. A subset of 24 samples generated by whole genome amplification (WGA) of 12 blood and 12 cheek swab samples produced slightly lower call rates (blood-WGA 99.3%; buccal-WGA 98.9%). Probe intensities from the array can also be used to analyze copy number polymorphisms, although this is not evaluated here.

Dataset construction

To perform a broader analysis of canine breed relationships and selective sweeps, we constructed a larger dataset consisting of unrelated samples from the Gentrain dataset, and unrelated control dogs genotyped for disease gene mapping studies from multiple breeds as part of the LUPA consortium. This dataset, which we refer to here as the "full LUPA genotype dataset" consists of 509 dogs from 46 diverse breeds and 15 wolves, genotyped on the CanineHD array. These include 156 dogs from 13 breeds derived from LUPA control dogs and 353 dogs from 33 breeds from the Gentrain dataset (See Table S1 for full details). A subset of this dataset, referred to here as the "reduced LUPA genotype dataset" is made up of all the samples in the 30 breeds (plus wolf) with more than 10 samples in the full dataset (471 samples in total).

Table 1 shows patterns of polymorphism in the reduced LUPA genotype dataset. In total, 157,393 SNPs on the array were polymorphic (90% of SNPs on the array). A mean of 119,615 SNPs (69%) were polymorphic within a single dog breed. Hence although there is a bias in the way that SNPs were ascertained, there is a substantial amount of variation within all breeds surveyed. On average 39 SNPs were polymorphic only in one breed, although this figure shows large variation between breeds. A subset of 1,471 SNPs were variable in wolves but not within any dog breed. However, most of these SNPs were originally discovered by comparisons of sequences from different dog breeds, which suggests that they are also variable between (but not within) dog breeds.

Evolutionary relationships between dog breeds

We used the CanineHD array to investigate breed relationships by constructing a neighbor-joining tree [24] of raw genetic distances in the full LUPA genotype dataset (Figure 1). Three main features are obvious: 1) Dogs from the same breed almost invariably cluster together. This reflects the notion that modern breeds are essentially closed gene pools that originated via population bottlenecks. 2) Little structure is obvious in the internal branches that distinguish breeds. This is consistent with the suggestion that all modern dog breeds arose from a common population within a short period of time and that only a very small proportion of genetic variation divides dog breeds into subgroups. 3) The internal branches leading to boxer and wolf are longer than those leading to other breeds. The long boxer branch can be explained by the fact that a large proportion of the SNPs were assayed by comparing boxer with other breeds, which implies that the dataset is enriched for SNPs that differ between boxer and other breeds. The longer wolf branch probably reflects more distant relatedness.

Some breeds show a tendency to group together in the tree, such as breeds of retrievers, spaniels, setters, and terriers. However, the length of the internal branches leading to these clusters is only a small fraction of the average total length of branches in these clusters, which indicates that genetic variation in dogs is much more severely affected by breed creating bottlenecks than it is by historical origins of various breeds, although detailed analysis of these data has power to reveal their historical origins [25]. The most obvious clustering of breeds is exhibited by two wolf hybrids: Sarloos and Czechoslovakian wolf dog, which exhibit a closer relationship to the wolf than other breeds as predicted by their known origin [26]. The German shepherd also clusters with this group, although this is likely to be a result of its close relationship with the Czechoslovakian wolf dog, rather than with wolf. The tree is consistent with previous studies and supports the accuracy and reliability of the array. Although the long boxer branch likely reflects SNP ascertainment bias on the array, the tree reflects extensive polymorphism both within and between breeds. This suggests the SNP ascertainment scheme is not problematic and that the array is well suited for both within and across breed gene mapping.

We performed coalescent simulations modeling the ascertainment bias, sample size, and inferred recombination rate in the true dataset (see Materials and Methods) in order to predict the expected patterns of genetic diversity that we expect to observe

Table 1. Levels of genetic variation in breeds with 10 or more samples.

Breed	Abbreviation	No. Samples	Seg. sites	Private seg. sites
Belgian Tervuren	BeT	12	115,154	0
Beagle	Bgl	10	115,254	16
Bernese Mountain Dog	BMD	12	106,152	15
Border Collie	BoC	16	127,491	7
Border Terrier	BoT	25	108,344	15
Brittany Spaniel	BrS	12	130,115	11
Cocker Spaniel	CoS	14	126,118	19
Dachshund	Dac	12	131,372	5
Doberman Pinscher	Dob	25	112,627	19
English Bulldog	EBD	13	111,720	19
Elkhound	Elk	12	127,066	82
English Setter	ESt	12	121,196	24
Eurasian	Eur	12	120,360	6
Finnish Spitz	FSp	12	109,510	20
Gordon Setter	GoS	25	134,615	12
Golden Retriever	Gry	11	112,144	10
Greyhound	GRe	14	128,907	45
German Shepherd	GSh	12	108,614	11
Greenland Sledge Dog	GSl	12	102,899	19
Irish Wolfhound	IrW	11	92,718	61
Jack Russell Terrier	JRT	12	137,837	12
Labrador Retriever	LRe	14	129,951	23
Newfoundland	NFd	25	127,503	13
Nova Scotia Duck Tolling Retriever	NSD	23	118,387	36
Rottweiler	Rtw	12	107,022	15
Schipperke	Sci	25	126,530	21
Shar-Pei	ShP	11	124,828	93
Standard Poodle	StP	12	132,289	123
Yorkshire Terrier	TYo	12	129,768	388
Weimaraner	Wei	26	111,958	21
Wolf	Wlf	15	118,256	1,471
Total	-	471	157,393	-

doi:10.1371/journal.pgen.1002316.t001

within and between breeds in the absence of selection. The bottleneck population sizes at breed creation used in the simulations are presented in Table S2. The decay of LD in the simulated data closely matches the real decay in LD (Figure S2).

Across-breed GWAS: morphological traits

To identify genetic variation associated with common traits that vary among breeds, we performed across-breed GWAS using the full LUPA genotype dataset. A list of traits and their variation between breeds is in Table S3. Each sample was given a value corresponding to the standardized breed phenotype for the trait under study. We performed quantitative association studies for size and personality traits whereas other traits were binary coded. For each GWAS, we assayed genome-wide significance by permuting the phenotype of each breed, assigning each dog of the same breed with identical phenotype values. The true significance of genotype-phenotype correlation at each SNP was compared with the maximum permuted value of all SNPs across the array in order to estimate genome-wide significance (see Materials and Methods). This permutation procedure corrects for the extreme population substructure present in dog breeds.

Using this method we were able to replicate several known associations. We first performed a GWAS comparing 4 breeds with furnishings (a coat type with moustache and eyebrows [9]) compared to 42 without them. Genome-wide significant associations were observed at 3 SNPs distributed located between 10.42 - 11.68 Mb on chromosome 13. The most strongly associated SNP is at 11,678,731 ($P_{genome}<0.001$), 44 kb from the causative SNP previously identified in RSPO2 [9]. We next scanned the genome for associations with size, using weight in kilograms as a proxy (data taken from [8]; see Table S3). The most strongly associated SNP was located on chromosome 15 at 44,242,609 ($P_{genome}=0.004$), which is within the IGF1 gene, previously implicated in size variation [8]. Genome-wide significant associations ($P_{genome}<0.05$) were observed at 7 SNPs within an interval between 44.23 - 44.44 Mb. In addition, we observed an association within a previously

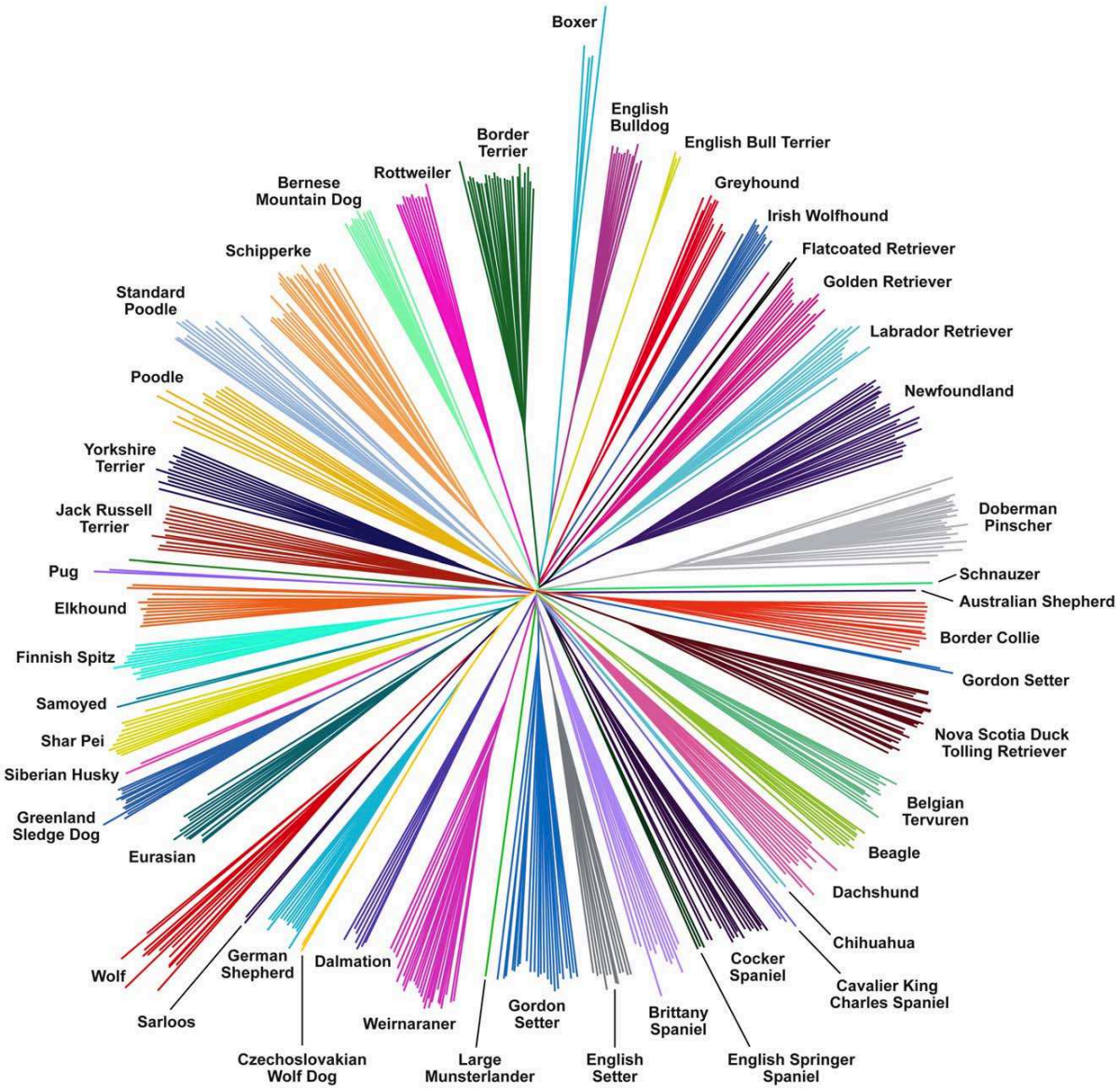


Figure 1. Neighbor-joining tree constructed from raw genetic distances representing relationships between samples. More than 170,000 SNPs were genotyped in 46 diverse dog breeds plus wolves using the CanineHD array. The boxer branches are longer, which likely represents the influence of ascertainment bias, as the SNPs were discovered from sequence alignments involving the boxer reference sequence.
doi:10.1371/journal.pgen.1002316.g001

defined region on chr10 (11,169,956 bp; $P_{genome} = 0.036$). The SNP at chr10:11,169,956 is about 500kb upstream of HMGA2, which has been established to be associated with body size variation in other species [27–29].

The frequency of the SNP (chr15:44,242,609) most strongly associated to size shows a steady decline according to the size of the breed. However, the differences in allele frequency at the SNP chr10:11,169,956 are more striking, as one allele appears at very low frequencies in all breeds apart from a number of very small breeds (Yorkshire Terrier, Border Terrier, Jack Russell Terrier, Schipperke), where it is at or close to fixation (Figure S3). Hence, there appears to be relatively continuous variation in frequency in a variant affecting IGF1 between breeds, whereas a variant upstream of HMGA2 appears to have been fixed in a subset of small breeds but shows little variation in allele frequencies in other breeds.

Dog breeds show extreme variation in ear morphology ranging from pricked ears to low hanging dropped ears. We performed a GWAS using 12 breeds with pricked ears and 15 breeds with dropped ears. Within an interval between 10.27 - 11.79 Mb, 23 SNPs had genome-wide significant associations ($P_{genome}<0.05$; Figure 2). The most strongly associated SNP was chr10: 11,072,007 ($P_{genome}<0.001$), which lies between the HGMA2 and MSRB3 genes. This region has been associated with ear type and body size in previous studies [6,12]. Using the CanineHD array, we are able to type SNPs at a much higher density in the associated region. There is also large variation between dog breeds in degree of tail curl. We classified breeds in our dataset into 11

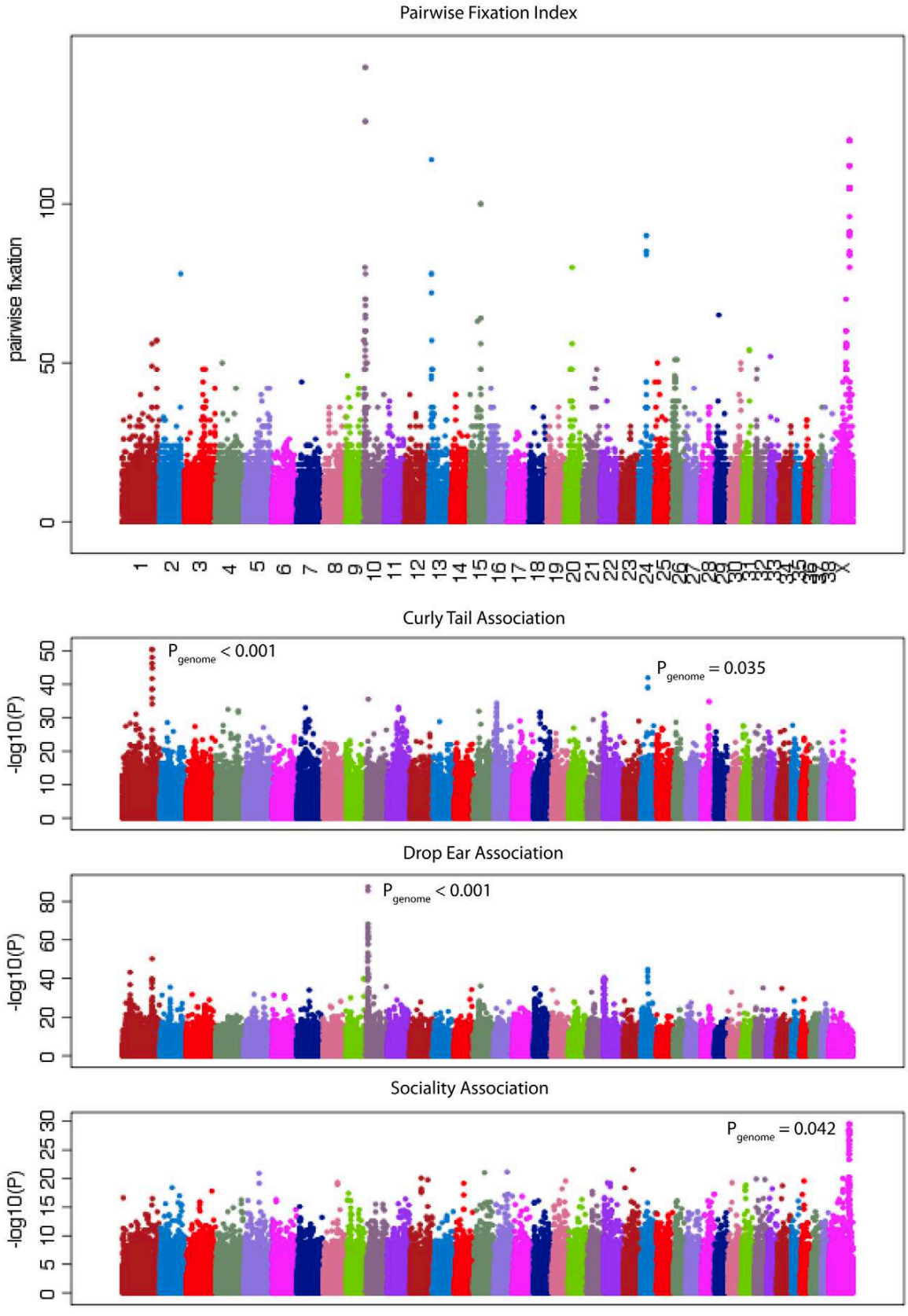

Pairwise Fixation Index
pairwise fixation
0
50
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 X
Curly Tail Association
-log10(P)
0 10 20 30 40 50
P genome < 0.001
P genome = 0.035
Drop Ear Association
-log10(P)
0 20 40 60 80
P genome < 0.001
Sociality Association
-log10(P)
0 5 10 15 20 25 30
P genome = 0.042

Figure 2. Identification of variants with large differences in allele frequencies between breeds that are associated with phenotypic variation. The top panel shows the variation in pairwise fixation index (see text for definition) at SNPs across the genome on the CanineHD array. The bottom panel shows GWAS for three traits (curly tail, drop ear, and sociality) with signals that correspond to SNPs with high population differentiation. P-values from breed permutations are also shown.
doi:10.1371/journal.pgen.1002316.g002

with curly tails and 7 with straight tails and performed a GWAS. Six SNPs on chromosome 1 were most significantly associated within an interval 96.26 - 96.96 Mb ($P_{genome}<0.05$; Figure 2), which are downstream of RCL1 and upstream of JAK2 (Figure S4). This region has not been previously associated with tail curl.

Across-breed GWAS: behavioral traits

We performed GWAS to search for variants that affect breed differences in behavior. We first performed a GWAS by comparing 18 bold and 19 non-bold breeds using phenotypic definitions from ref. [12]. Highly significant associations were found at two SNPs on chromosome 10, 11,440,860 ($P_{genome}< 0.001$) and 10,804,969 ($P_{genome} = 0.006$), in the same region associated with both drop ear and size. Variation within this region is therefore associated with at least two morphological and one behavioral trait, which may be correlated. The region contains several genes including WIF1, HMGA2, GNS, and MSRB3 (see Figure S5). However, the most significant associations for each trait appear to occur in different places. The SNPs most associated with drop ear and size occur 98 kb apart between the MSRB3 and HMGA2 genes, with the drop ear association closer to MSRB3, whereas the top boldness association occurs within an intron of HMGA2, 271 kb 3′ of the size association. There is however a strong correlation between the bold and non-bold breed classifications and the drop ear and size classifications. All prick eared and small dogs were classified as bold in the dataset, whereas all drop eared dogs were classified as non-bold, with the exception of Bernese Mountain Dog (see Table S3).

Breed averages for five personality traits measured objectively under controlled conditions were obtained from the Swedish Kennel Club. The traits are defined as sociability, curiosity, playfulness, chase-proneness and aggressiveness [30] and have been shown to be consistent among multiple tests of the same dog [31]. We performed quantitative GWAS using the breed-average trait values presented in Table S3. We observed significant associations at a number of SNPs for the trait sociability, which measures a dog's attitude toward unknown people (Figure 2). No SNPs reached genome-wide significance, but a large number of SNPs on the X chromosome also showed strong association. In order to accurately measure genome-wide significance in the sex chromosomes compared to autosomes we removed male dogs from the analysis. This analysis identified 10 SNPs with genome-wide significant associations ($P_{genome}<0.05$) in the interval 106.03–106.61 Mb on the X chromosome (see Figure S6). This region was also identified in a previous study [6] to be highly differentiated between breeds and correlated with body size and skull shape.

Single-SNP F_{ST} statistics identify SNPs under selection in multiple breeds

Across breed GWAS is a powerful approach for identifying genotype-phenotype relationship for traits shared among breeds. The variants identified by this approach, by definition, have large variation in allele frequencies between breeds. However, there may be many more such SNPs that have been subjected to similar selective pressures for common traits between breeds where the trait is not identified. In order to find such loci, we identified SNPs that exhibit high levels of differentiation between dog breeds using the F_{ST} statistics calculated for the >173,000 SNPs in the reduced LUPA genotype dataset.

A total of 240 SNPs have a $F_{ST}>0.55$ and overall minor allele frequency >0.15 in the reduced dataset containing breeds with at least 10 samples. These cut offs are identical to those used by Boyko et al. [6] and are chosen for comparison. In the simulated data, no SNPs pass this cut off ($p<0.0001$; χ^2 test). We then generated a list of highly differentiated regions, by merging all SNPs in this list within 500kb of each other into single regions, resulting in 44 regions containing between 1 and 94 SNPs with elevated F_{ST}. Regions with two or more SNPs are presented in Table 2 and the complete list is presented in Table S4. Figure 2 presents a value for each SNP (used for illustration purposes only) that we term "pairwise fixation index" to highlight differences in allele frequencies between breeds. This is defined as pq, where p is the number of breeds where allele A is fixed or close to fixation (frequency >0.95) and q is the number of breeds where allele B is fixed or close to fixation (frequency >0.95). In total 53,944 out of 154,034 variable SNPs have a pq value > 0, indicating that they are fixed for different alleles in at least 2 breeds. The regions of high F_{ST} correspond strongly to loci where trait associations have been reported. In particular, 8 of the 9 regions comprised of more than 3 high-F_{ST} SNPs overlap known trait-associated regions, and it is likely that most or all of the remaining regions with high F_{ST} show a correlation with an as yet undefined trait. Three of these regions were not previously reported in a study based on a less dense array [6] including a region on chromosome 7 (27.99 - 28.15 Mb) containing five highly differentiated SNPs that encompasses the DMD gene. The locations of all regions are marked in Figure 3, which presents a comprehensive map of regions that are likely to contain major loci influencing phenotypic variation between dog breeds.

Three regions longer than 1Mb are identified by this measure, likely signifying regions under strong selection in many breeds. These consist of a 2.6 Mb region on chromosome X that associates with body size, skull shape and sociability, a 2.0 Mb region on chromosome 10 that associates with drop ear, size and boldness and a 2.1 Mb region on chromosome X associated with limb and tail length (see also [6]). Other loci identified include three loci involved in coat type (RSPO2, FGF5, KRT71) [9]. In particular the RSPO2 gene associated with furnishings is found within an extended 0.6 Mb region. The MITF and ASIP (Agouti) genes known to be involved in coat color in dogs [11] are also identified. The region on chromosome 1 identified here as associated with curly tail and previously associated with snout ratio [6] is associated with 4 SNPs with high F_{ST} across 50 kb. Other genes of note identified are LCORL, known to associate with human height [27,29], KITLG, associated with coat color in other species [32] and several genes with key developmental roles, such as sonic hedgehog (SHH) involved in patterning in the early embryo, msh homeobox 1 (MSX1), involved in embyrogenesis and bone morphogenic protein 1 (BMP1) involved in bone development.

Genome-wide scans for signatures of selective sweeps in single breeds

Rare selective sweeps corresponding to regions of the genome under selection in only one or a small number of breeds in our

Table 2. Description of regions with at least two nearby SNPs with high F_{ST} (>0.55) and high minor allele frequency (>15%).

no.	chr	start (bp)	end (bp)	no. SNPs	length (kb)	max F_{ST}	association	candidate genes
1	X	104,640,567	107,235,825	96	2,595	0.75	sociality*, size, skull shape	
2	10	9,836,009	11,792,711	33	1,957	0.81	drop ear, size, boldness*	WIF1, HMGA2, MSRB3
3	X	85,365,233	87,444,776	29	2,080	0.58	limb/tail length	
4	13	11,095,120	11,678,731	10	584	0.73	furnishings	RSPO2
5	15	44,216,576	44,267,011	6	50	0.68	size	IGF1
6	24	26,270,399	26,370,499	5	100	0.70	coat color	ASIP
7	X	27,990,332	28,152,042	5	162	0.63	*	DMD
8	20	24,841,077	24,889,547	4	48	0.63	coat color	MITF
9	1	96,286,007	96,335,577	4	50	0.58	snout ratio, curly tail*	RCL1
10	25	3,603,872	4,065,978	3	462	0.63	*	FOXO1, BRD2
11	20	20,449,477	20,539,359	3	90	0.62	*	KLF15, ZXDC, UROC1, TXNRD3
12	13	10,210,459	10,225,305	3	15	0.59	*	OXR1
13	X	120,769,286	121,212,627	2	443	0.70	*	MAGEA, THEM185A
14	31	14,888,449	14,944,938	2	56	0.61	*	NRIP1
15	3	68,103,223	68,260,652	2	157	0.60	*	CPEB2
16	10	5,221,427	5,440,236	2	219	0.60	size*	CDK4
17	3	93,933,450	93,944,095	2	11	0.60	size	LCORL
18	15	32,638,117	32,853,840	2	216	0.57		KITLG
19	16	3,198,732	3,212,612	2	14	0.56	*	PKD1L1

Associations (or regions with no suggested association) marked with an asterisk are novel to this study. Others are summarized in [6].
doi:10.1371/journal.pgen.1002316.t002

dataset cannot be detected by across-breed GWAS due to lack of power. They also have a weak effect on F_{ST} values at single SNPs across all breeds compared to regions under selection in many breeds. In order to identify such rare sweeps, we scanned patterns of variation in the reduced LUPA genotype dataset to identify extended regions where haplotypes had become fixed in one or more breeds, leading to a local reduction in genetic variation and increase in population differentiation. We analyzed 150 kb sliding windows, overlapping by 25 kb in each breed compared with other breeds using two statistics. The first statistic, S_i, is calculated by summing regional deviations in levels of relative heterozygosity across the genome between two breeds compared to the genomic average and summing across all pairwise comparisons. Relative heterozygosity is defined as the number of SNPs segregating in a genomic window in one breed divided by the number of SNPs segregating in that window in two breeds under comparison. Hence, regions with low S_i in a breed contain few segregating SNPs compared to other breeds. The second statistic, d_i, was implemented by Akey et al. [15], and is based on pairwise F_{ST} values normalized for a given breed relative to the genome-wide average, summed across all pairwise combinations involving the given breed. Regions of high d_i in a particular breed exhibit a large difference in allele frequencies compared with other breeds.

We first identified windows with S_i or d_i values in an extreme 1% tail of their respective distributions (the bottom 1% for S_i and top 1% for d_i). Overlapping windows were then collapsed into larger regions (see Materials and Methods). These regions represent a map of blocks of reduced heterozygosity or elevated population differentiation in each breed. We repeated our analysis of d_i and S_i on the simulated data (see above). For both statistics, the average length of regions identified was similar in real versus simulated datasets. However, there was a strong excess of regions >250 kb in the real compared with simulated datasets, which likely reflects regions influenced by selection. In order to distinguish regions generated by genetic drift compared with those generated by selective sweeps we first estimated a marginal p-value for each block, equal to the proportion of simulated blocks with longer lengths in the same breed. We then adjusted these p-values using a 5% False Discovery Rate (FDR; see Materials and Methods and ref. [33]). In total 524 high confidence putative sweeps (an average of 17 per breed) were identified using the S_i statistic, with a mean size of 475 kb. However, none of the regions identified by the d_i statistic remained significant after FDR correction. Figure S7 shows the distribution of significant S_i regions in the dog genome.

Full lists of regions identified by the S_i and d_i analyses including the marginal and FDR corrected p-values are presented in Table S5 and summary statistics of these regions are presented in Table S6. These regions are also available as a UCSC annotation dataset (see Materials and Methods for URLs). The UCSC browser offers a graphical display of S_i and d_i regions as well as d_i values for all SNPs analyzed [34]. Table S7 shows the overlap between these regions and those identified in previous studies (refs. [6] and [15]).

Long regions of reduced heterozygosity are identified by the S_i statistic

The S_i test identifies blocks of the genome where one breed has little or no variation consistent with fixation of a long haplotype by a selective sweep. On average, only 19.9% of SNPs have segregating variants in these regions in the breed where they are identified compared with the genome average of 74.5%. Among the 524 putative sweeps are several loci already implicated in breed-defining characters. Notably, a 590 kb region of low S_i overlapping the FGF4 retrogene on chromosome 18 associated with chondrodysplasia in Dachshunds. A 1.4 Mb region of low S_i overlapping the HAS2 gene implicated in skin wrinkling [16] is observed in Shar-Pei. Regions in the vicinity of the RSPO2 locus

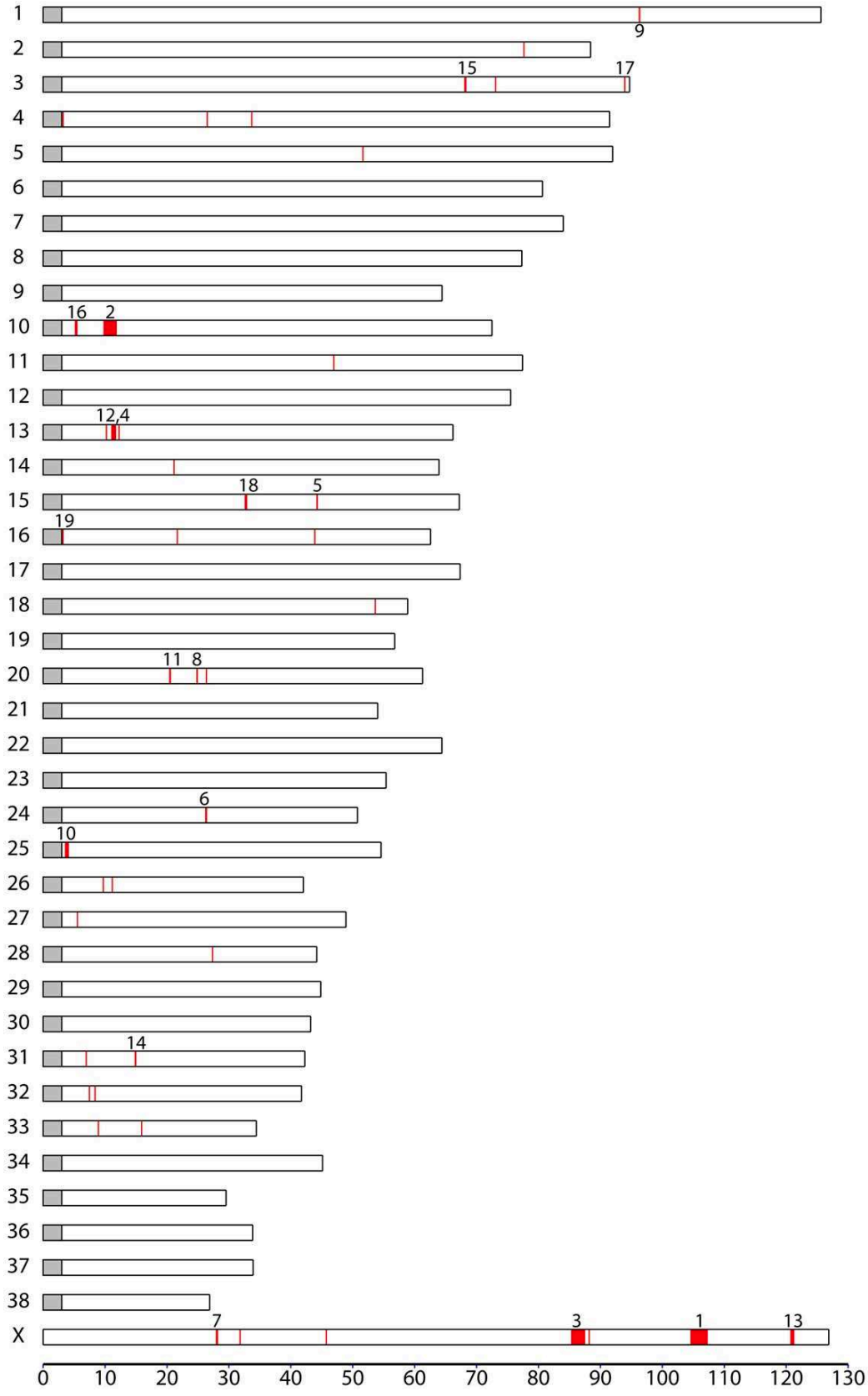

1
9
2
15
17
3
4
5
6
7
8
9
16
2
10
11
12
12,4
13
14
18
5
15
19
16
17
18
19
11
8
20
21
22
23
6
24
10
25
26
27
28
29
30
14
31
32
33
34
35
36
37
38
7
3
1
13
X
0
10
20
30
40
50
60
70
80
90
100
110
120
130

Figure 3. Map of regions with extreme differentiation between dog breeds as identified by single-SNP F_{ST}. All regions with at least one SNP with F_{ST} >0.55 and minor allele frequency >15% are shown. Numbers correspond to the regions in Table 2, which contain at least two nearby SNPs that pass these thresholds.
doi:10.1371/journal.pgen.1002316.g003

implicated with furnishings are observed in 2 breeds, which both have furnishings (Yorkshire Terrier and Standard Poodle). However, many variants implicated in phenotypic variation between breeds are not strongly associated with regions of reduced S_i. No putative sweeps overlapping the IGF1 locus are identified in small breeds using this statistic. This is likely to be because there appears to be continuous variation in allele frequency at this locus between breeds rather than complete fixation of certain haplotypes in several breeds (see Figure S3).

Table 3 shows the top 20 longest regions of significantly reduced S_i. It should be noted that two pairs of putative sweep regions occur at contiguous locations in the same breed (no. 2 and 12 in Beagle and no. 5 and 6 in Irish Wolfhound), which could potentially represent single selective sweeps. The longest region we identified is 3.1 Mb long (chr22: 5.3–8.4 Mb) in beagles. This region overlaps 3 other putative sweeps within the top 20 in other breeds (Gordon Setter, Rottweiler, and Newfoundland) whereas no other regions in the top 20 are overlapping. As this and other regions with strongest evidence for sweeps are long and contain many genes, it is not possible to identify the locus under selection in a single sweep. However, it is interesting to note that they contain genes associated with disease in humans and dogs including epilepsy (KCNQ5), cancer (NPM1, FGR), and autoimmune disease (IL6). A long sweep on chromosome 30 in Golden retrievers spans the RYR1 gene, involved in the skeletal muscle calcium release channel and implicated in canine malignant hyperthermia by linkage analysis [35]. We also identified a number of genes involved in spermatogenesis and fertilization (SPAG1, FNDC3A, CLGN) which is a category often enriched in genes under positive selection in other species [36].

In cases where multiple breeds are affected by selection acting on the same variant, it may be possible to narrow an interval containing the causative mutation by identifying a core region of identity by state (IBS) between all breeds where haplotypes are shared, most likely reflecting common ancestry. We searched our dataset for regions with significant drops in S_i that overlapped between different breeds. We then identified the maximal region where the same haplotype was fixed in all breeds identified. For many significant long regions we were able to identify shorter regions of IBS. The regions shared by 3 or more breeds are shown in Table 4 and a full list is in Table S5. As a validation of this method, we identified a 187 kb region where an identical haplotype is fixed among the 3 breeds with furnishings where we identified a sweep spanning the previously defined causative indel (region 14). Hence this method is able to identify interval containing the causative mutation in shared region of identity by descent.

The inferred selective sweep shared by the most breeds in this analysis was a 485 kb haplotype on chromosome 22 (5.4–5.9 Mb) shared by 8 breeds (Beagle, Border Terrier, English Bulldog, Gordon Setter, Irish Wolfhound, Newfoundland, Rottweiler, Weimaraner). This region contains 2 genes: FNDC3A, fibronectin type III domain containing 3A [37], which is involved in spermatogenesis and also expressed in odontoblasts indicating a role in odontogenesis, and CYSLTR2 cysteinyl leukotriene receptor 2, a member of the superfamily of G protein-coupled

Table 3. Description of longest regions with significant drop in S_i.

no.	breed	chr	start	end	length	no. genes	overlaps
1	Bgl	22	5,286,218	8,423,791	3,137,574	27	S_i (1,11,13,17) d_i (1,8,9)
2	EBD	26	8,813,638	11,587,844	2,774,207	70	d_i (7,14)
3	ESt	25	27,590,228	30,101,489	2,511,262	29	
4	ShP	13	3,036,181	5,264,386	2,228,206	24	d_i (3,12)
5	IrW	2	21,323,001	23,188,692	1,865,692	18	
6	IrW	2	19,483,009	21,068,544	1,585,536	20	
7	Bgl	14	39,147,833	40,601,429	1,453,597	22	
8	GRe	30	3,901,787	5,329,451	1,427,665	12	
9	ShP	13	23,047,599	24,433,840	1,386,242	6	
10	Gry	25	7,271,327	8,636,861	1,365,535	10	
11	GoS	22	4,904,331	6,215,828	1,311,498	14	S_i (1,11,13,17) d_i (1,8,9)
12	EBD	26	12,121,983	13,432,228	1,310,246	24	
13	Rtw	22	4,654,448	5,937,680	1,283,233	18	S_i (1,11,13,17) d_i (1,8,9)
14	Gry	19	4,473,961	5,756,046	1,282,086	10	
15	BMD	2	75,308,805	76,571,807	1,263,003	35	
16	Bgl	12	37,413,405	38,655,096	1,241,692	13	
17	NFd	22	4,752,012	5,980,115	1,228,104	17	S_i (1,11,13,17) d_i (1,8,9)
18	Rtw	13	50,613,526	51,755,732	1,142,207	16	
19	Dob	20	39,397,583	40,517,365	1,119,783	34	
20	GRe	8	5,436,616	6,533,588	1,096,973	42	

doi:10.1371/journal.pgen.1002316.t003

Table 4. Description of regions with identical fixed haplotypes across multiple breeds.

no.	chr.	start	end	length	no. breeds	no. genes	candidate genes
1	22	5,466,185	5,950,731	484,547	8	2	FNDC3A,CYSLTR2
2	37	3,453,815	3,830,321	376,507	7	6	MSTN
3	X	101,638,881	101,992,724	353,844	5	1	UBE2I
4	6	26,881,144	27,285,067	403,924	4	7	CRYM,ZP2
5	1	7,427,961	7,811,190	383,230	4	2	ZNF407
6	23	5,694,045	5,971,463	277,419	4	4	HSP90AA1
7	21	4,828,296	5,034,632	206,337	4	1	CNTN5
8	17	3,190,961	3,672,468	481,508	3	4	TMEM18
9	19	5,046,934	5,463,342	416,409	3	7	UCP1
10	2	80,709,265	81,050,769	341,505	3	2	EIF4G3
11	37	14,809,394	15,087,734	278,341	3	6	NBEAL1
12	X	112,830,694	113,040,282	209,589	3	3	ATP11C
13	11	49,319,964	49,523,469	203,506	3	1	LINGO2
14	13	11,509,194	11,695,899	186,706	3	1	RSPO2
15	12	36,463,722	36,557,249	93,528	3	0	B3GAT2
16	8	24,377,123	24,415,610	38,488	3	0	CCT6P1

doi:10.1371/journal.pgen.1002316.t004

receptors. A 402 kb haplotype on chromosome 37 (3.5–3.8 Mb) is shared among 7 breeds (Bernese Mountain Dog, Beagle, Border Terrier, Doberman, Elkhound, Finnish Spitz, Golden Retriever). This haplotype contains 7 genes including the MSTN (myostatin) gene. This gene is associated with double muscling in cattle [38] and in a similar phenotype observed in whippets [39]. It is therefore plausible that this region has been a target of selection in multiple dog breeds in order to modify muscle mass. A 354 kb haplotype on chromosome X is fixed in 5 breeds (101.6–102.0 Mb) and contains only one gene: UBE2I, an ubiquitin-conjugating enzyme. This enzyme has been shown to interact with MITF, involved in coat color, and is suggested to be a key

Table 5. Description of the longest regions with elevated d_i.

no.	breed	chr	start	end	length	no. of genes	overlaps
1	ShP	22	4,828,721	6,233,045	1,404,325	15	d_i (1,8,9) S_i (1,11,13,17)
2	JRT	10	10,265,925	11,644,756	1,378,832	10	d_i (2,4,10,18)
3	Elk	13	3,778,724	5,152,585	1,373,862	12	d_i (3,12) S_i (4)
4	TYo	10	10,265,925	11,559,700	1,293,776	8	d_i (2,4,10,18)
5	EBD	9	4,125,282	5,418,477	1,293,196	22	
6	BoC	22	15,982,263	17,193,960	1,211,698	2	
7	EBD	26	10,491,787	11,629,631	1,137,845	29	d_i (7,14) S_i (2)
8	FSp	22	4,300,860	5,397,353	1,096,494	19	d_i (1,8,9) S_i (1,11,13,17)
9	NFd	22	4,300,860	5,361,506	1,060,647	18	d_i (1,8,9) S_i (1,11,13,17)
10	Elk	10	10,707,193	11,644,756	937,564	7	d_i (2,4,10,18)
11	LRe	13	40,738,270	41,665,437	927,168	36	
12	Bgl	13	3,266,021	4,176,521	910,501	13	d_i (3,12) S_i (4)
13	TYo	3	40,471,308	41,367,554	896,247	11	
14	TYo	26	10,744,458	11,629,631	885,174	23	d_i (7,14) S_i (2)
15	ESt	19	46,765,322	47,637,012	871,691	5	
16	ShP	3	3,069,417	3,922,440	853,024	4	
17	TYo	24	25,532,482	26,370,499	838,018	20	
18	BrS	10	10,832,919	11,644,756	811,838	5	d_i (2,4,10,18)
19	BoT	19	10,239,039	11,031,247	792,209	3	
20	IrW	37	3,170,534	3,960,864	790,331	12	

doi:10.1371/journal.pgen.1002316.t005

regulator of melanocyte differentiation [40] although it also has a number of other features.

Regions of elevated population differentiation are identified by the d_i statistic

There are many extremely differentiated regions although none of them passed the 5% FDR correction for length (see Table S5 for full list). Variation in S_i and d_i statistics in the 10 longest regions identified by the S_i test is presented in Figure S8. This comparison of the d_i and S_i tests reveals that the increases in d_i often occur within a more restricted region of a large block of fixed haplotype from the S_i tests, indicating that they represent regions where an otherwise rare ancestral sub-haplotype has been fixed in a certain breed. It therefore appears that many regions detected by d_i and S_i tests are complementary. Among the top 20 longest putative sweeps identified by the d_i statistic (Table 5) are 3 overlapping sweeps that also overlap the common sweep containing FNDC3A and CYSLTR2 identified by the S_i test. We also identify putative sweeps in 4 breeds overlapping the region associated with drop ear, size and boldness among the top 20 d_i sweeps. Two putative sweeps in this list overlap a region on chromosome 13 (3.3–5.2 Mb), which is also identified by the S_i test. One gene of note in this region is VPS13B, which may have an important role in development and is associated with Cohen syndrome, which has an effect on development of many parts of the body [41]. The second longest putative sweep identified by S_i on chromosome 26 is also identified in two of the top 20 longest d_i regions.

XP-EHH

We performed an additional validation of our results using a third statistic, XP-EHH, which identifies regions where a long haplotype has reached fixation, or is close to fixation in one breed compared with other breeds [18]. We calculated the mean XP-EHH for all of the regions identified by the S_i and d_i tests. For the regions constructed from the top 1% of d_i (6404 regions) and S_i (7618 regions) statistics, mean XP-EHH was -0.94 and -1.13 respectively across all breeds compared with a genome average of zero. This difference is consistent across all 30 breeds and is highly significant (binomial test: $P<10^{-9}$). This confirms that regions identified by the S_i and d_i tests are associated with unusually long haplotypes at or near fixation in the breeds under selection compared with other breeds.

Functional categories

We analyzed genes closest to all singleton SNPs with high F_{ST} for enrichment in gene ontology (GO) categories. The six most significantly overrepresented GO categories were all involved in development. 11 of the 22 genes were found in the "developmental processes category" ($P=0.00036$) and tissue, system, organ, anatomical structure and multicellular organismal development were all significantly overrepresented ($P<0.0007$). These highly differentiated SNPs therefore highlight a number of regions involved in development that are likely to have been modified by artificial selection and contribute to the high diversity of dog breeds.

We next analyzed gene content of all of the regions constructed from the top 1% of d_i and S_i distributions that pass the marginal p-value <0.05 for each breed. We only considered regions containing a single gene, in order to enrich the analysis for true targets of selection, although this list is still expected to contain false positives. There were 119 d_i regions and 272 S_i regions containing one gene only (29 genes shared). We performed GO analysis using human genes with 1:1 human-dog orthologous relationship. As longer genes are over-represented within long genomic segments containing only one gene, we compared these candidate selection genes to a background dataset with similar length (see Materials and Methods). A total of 40 GO categories were significantly enriched in the S_i analysis and 6 in the d_i analysis (Table 6). Developmental processes, central nervous system, organ development and pigmentation pathways are significantly enriched in S_i regions whereas cell communication and signal transduction are the most represented in d_i regions. These differences in enriched GO categories could potentially reflect differences in the form of selection detected by the two statistics.

A large number of genes detected by the S_i analysis are significantly over-represented in several GO categories, which may reflect pleiotropic effects. A total of 23 of the genes belong to at

Table 6. Enriched GO categories with 5 or more genes in S_i and d_i candidate selection regions.

GO ID	GO category	no. genes	enrichment	adjusted p-value
S_i regions				
GO:0032501	multicellular organismal process	40	1.3	0.034
GO:0048869	cellular developmental process	22	1.5	0.032
GO:0030154	cell differentiation	21	1.6	0.021
GO:0048468	cell development	14	1.7	0.029
GO:0051239	regulation of multicellular organismal process	12	1.9	0.021
GO:0007186	G-protein coupled receptor protein signaling pathway	10	2.1	0.018
GO:0007507	heart development	6	2.9	0.016
GO:0030097	hemopoiesis	5	2.9	0.025
GO:0048534	hemopoietic or lymphoid organ development	5	2.7	0.033
GO:0048514	blood vessel morphogenesis	5	2.7	0.033
d_i regions				
GO:0007154	cell communication	16	1.5	0.028
GO:0007165	signal transduction	15	1.6	0.027
GO:0007166	cell surface receptor linked signal transduction	8	2	0.04

doi:10.1371/journal.pgen.1002316.t006

least 10 enriched GO categories. As an example, one candidate selection gene the thyroid stimulating hormone receptor (TSHR) is involved in 25 enriched GO categories, including central nervous system and regulation of nucleotide biosynthetic process. This gene is suggested to have an essential role in photoperiod control of reproduction in vertebrates, in organ development and in metabolic regulation and has been recently been implicated as an important domestication gene in chicken [42].

The two larger biological processes over-represented by d_i regions are 'cell communication' and 'signal transduction', which are represented by 16 and 15 genes, respectively. A region on chromosome 3 with strong statistical support contains the gene for insulin-like growth factor receptor1 (IGF1R), also detected by S_i statistics. This is a strong candidate gene in relation to selection for growth, a phenotype that has been strongly selected in dog. Another example is ANGPT1, which plays roles in vascular development and angiogenesis and contributes to blood vessel maturation and stability. This gene has been identified in a set of positively-selected genes in human Tibetan populations for which selection may have occurred to allow for more efficient oxygen utilization [43]. The presence of TSHR and ANGPT1 in enriched GO categories may suggest that these pathways are commonly involved in recent adaptation.

Fine-scale analysis of a region associated with multiple traits

The region containing the most highly differentiated SNPs identified by the single-SNP F_{ST} analysis is 9.8 – 11.8 Mb on chromosome 10. Variation in this region was also found to correlate with multiple traits: drop ear, size and boldness. As boldness shows a strong correlation with the other traits, we focused on analyzing the contribution of variants in this region to drop ear and size. We first analyzed the variation in allele frequencies of the SNPs most associated with size and drop ear across breeds scored for these traits. Size was measured as the breed average in kg, and drop ear was scored on a scale of 1-5 (Figure 4A). The SNP most associated with ear type (chr10: 11,072,007) showed correlation with this trait, but little association with size. Allele frequencies display continuous variation between breeds. In contrast, the minor allele at the SNP most associated with size (chr10: 11,169,956) was not present in most breeds, but close to fixation in a subset of small breeds (Chihuahua, Yorkshire Terrier, Border Terrier, Schipperke, Jack Russell Terrier). All of these breeds were also fixed for the prick ear allele at the ear type SNP. Based on this analysis, we hypothesize that combinations of two alleles at these two SNP loci result in three main haplotypes affecting ear type and body size segregate among dogs (Figure 4B). The small size-pricked ear combination is present in the small (non-chondrodysplasic) breeds mentioned. All other breeds genotyped possess the large-prick or large size-drop ear haplotype, and the small size-drop ear combination is not observed in our dataset.

In order to identify variants potentially responsible for these traits, we comprehensively characterized variation in a genomic segment encompassing this region (chr10: 9.5 Mb – 12.5 Mb) using Roche NimbleGen hybrid capture and sequencing using an Illumina Genome Analyzer. We choose 3 breeds with the dropped ear phenotype (Lagotto Romagnolo, Leonberger, and Bernese Mountain Dog) and 3 with the pricked ear phenotype (Chinese Crested, Schipperke, and Finnish Spitz). Two of the pricked ear breeds are small, with breed average <6 kg: Chinese Crested and Schipperke. We sequenced each breed independently, using a pool of 5 dogs from each breed. On average 8 million reads per pool were produced, of which 48% mapped to the 3 Mb region on chr10. In total, 61% of this region was mapped by at least one read. In the 68% of the region defined as non-repetitive, reads mapped to 98% of bases, at an average coverage depth of 114x. By comparison with the reference sequence, we identified fixed differences and polymorphic sites within each breed. Differences in the pattern of polymorphism between dropped and pricked eared dogs are clearly apparent, and drop eared breeds exhibit a lower level of variation on average compared with prick eared dogs, which is mainly restricted to a ~2 Mb region between 9.5 Mb and 11.5 Mb (Figure 5).

We next identified SNPs in this region that were completely fixed for different alleles in dropped and pricked eared breeds. These SNPs are distributed unevenly across the region, and a peak in the number of such fixed SNPs occurs around 11.3-11.5 Mb. In total 287 SNPs or small indels were completely fixed for different alleles in the drop ear compared with pricked ear breeds. Twenty-five of these SNPs reside in regions that show evidence for sequence conservation and are therefore candidates for being the causative mutation (Table S8). Of the 6 breeds, only Chinese Crested was completely fixed for the small size allele at chr10: 11,169,956 in our dataset. We therefore identified SNPs fixed for different alleles in Chinese Crested compared with all other breeds except Schipperke (this breed was excluded because it is small but was not completely fixed for the size-associated SNP from the GWAS). In total 297 SNPs or small indels were completely fixed for different alleles in these two groups. Of these, 17 were in conserved regions and are therefore candidates for affecting size (Table S9).

Discussion

Here we present a comprehensive catalogue of genomic regions that are candidates for being affected by artificial selection in dogs using the densest panel of SNPs to date. We focus on two main types of variant: 1) common variants that affect variation in a trait in many breeds and 2) rare variants that have undergone selective sweeps in one or a few breeds. For the first category, we identify loci where variation correlates with morphological traits such as body size and tail curl, and behavioral traits such as sociability and boldness. We also identify several loci with evidence for a high degree of population differentiation between breeds, for which the connection with phenotypic traits in dogs is not known, but that are known to associate with traits such as pigmentation and body size. To identify loci in the second category, we searched for regions with reduced heterozygosity and high population differentiation, characteristic of selective sweeps. This analysis identified loci known to be associated with breed-defining characteristics such as chondrodysplasia, skin wrinkling, and furnishings. In addition, we identify several extended regions with reduced heterozygosity $>$ 1 Mb consistent with recent selective sweeps in one or more breeds, including striking examples such as a region containing the FNDC3A and CYSLTR2 genes, and a region containing the MSTN (myostatin) gene that both bear the signal of selection in multiple breeds.

The candidate selection loci we identified are strongly enriched for genes involved in developmental and metabolic processes. In general, the GO terms we find to be significantly enriched are different from analyses of selection in natural populations, in which genes commonly targeted by positive selection include those involved in immunity and defense, olfaction and responses to external stimuli [36]. These results are consistent with the idea that artificial selection in domestic animals target different functional categories than natural selection. This result contrasts with that of Akey et al. [15] who found genes involved in immunity and

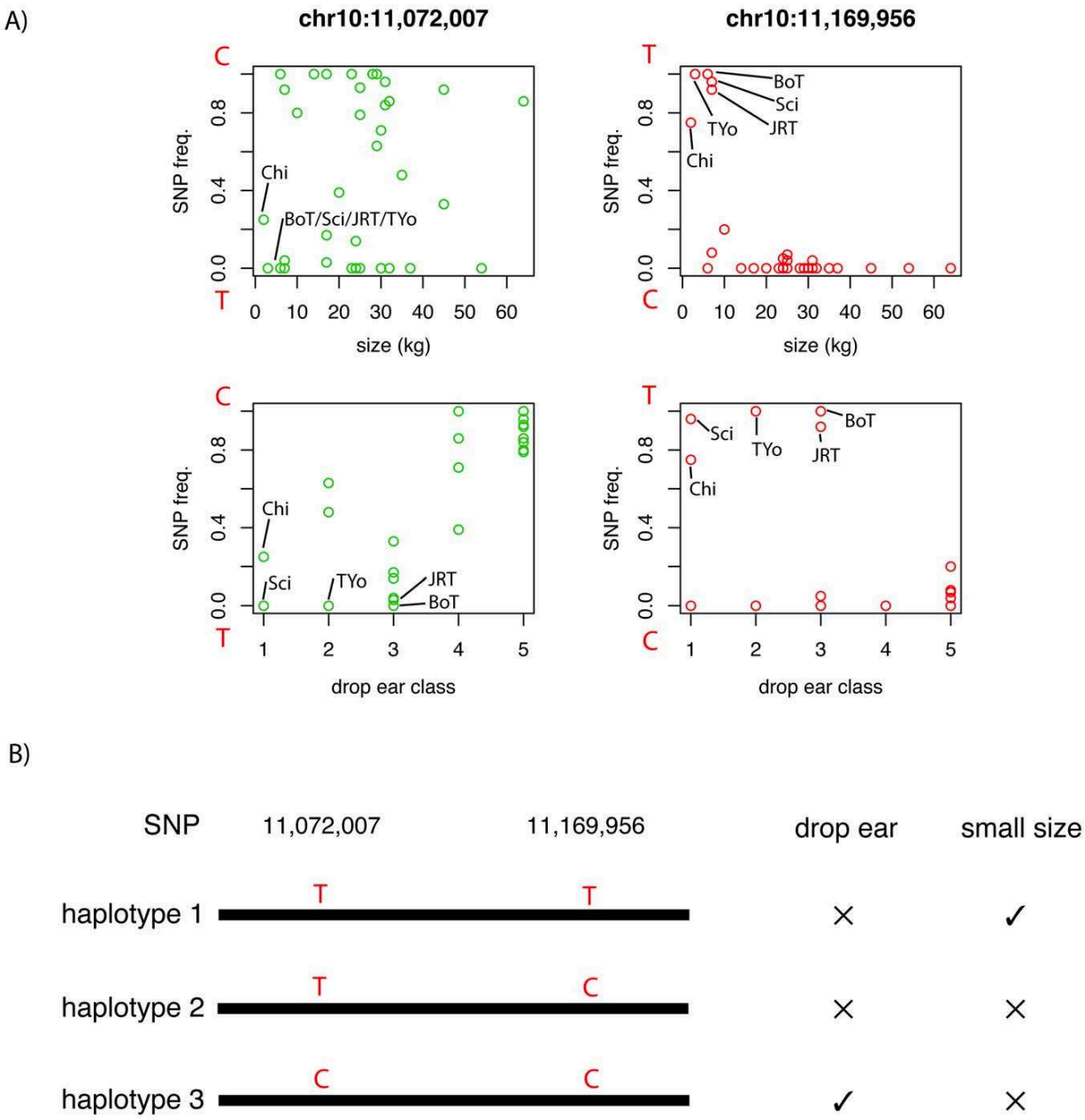


Figure 4. Variation in allele frequencies of the SNPs with the strongest association to drop ear (chr10:11,072,007) and body size (chr10:11,169,956). A) The frequency of these two SNPs in each breed is plotted against the classification of each breed according to body size and drop ear phenotype. The first SNP shows continuous variation in frequency between breeds, and correlates with drop ear class (1 = pricked ear, 5 = dropped ear). At the second SNP, one allele has very high frequency in some small breeds, but very low frequency in all other breeds. A set of small breeds with high minor allele frequency at this SNP are marked. B) The allele frequencies at these SNPs are consistent with the presence of three haplotypes, associated with different combinations of these traits.
doi:10.1371/journal.pgen.1002316.g004

defense to be overrepresented among their candidate selection regions.

Artificial selection on dog breeds coincided with breed creation bottlenecks leading to genetically distinct breeds fixed for novel traits [1,3,4]. Hence a large proportion of phenotypic and genetic variation is apportioned between but not within breeds. It is notable that 35% of polymorphic SNPs we analyzed are fixed or almost fixed for alternative alleles in two or more breeds. This is in sharp contrast to the differences between human populations, where only 78 near-fixed differences, that are all strong candidates for being under selection, were observed between four populations among 15 million SNPs identified using whole-genome resequencing [44]. The strong influence of genetic drift on genetic variation in dog breeds has also led to random fixation of long haplotypes and it is estimated that on average ~25% of the genome lies within a homozygous block >100kb in an average breed. This

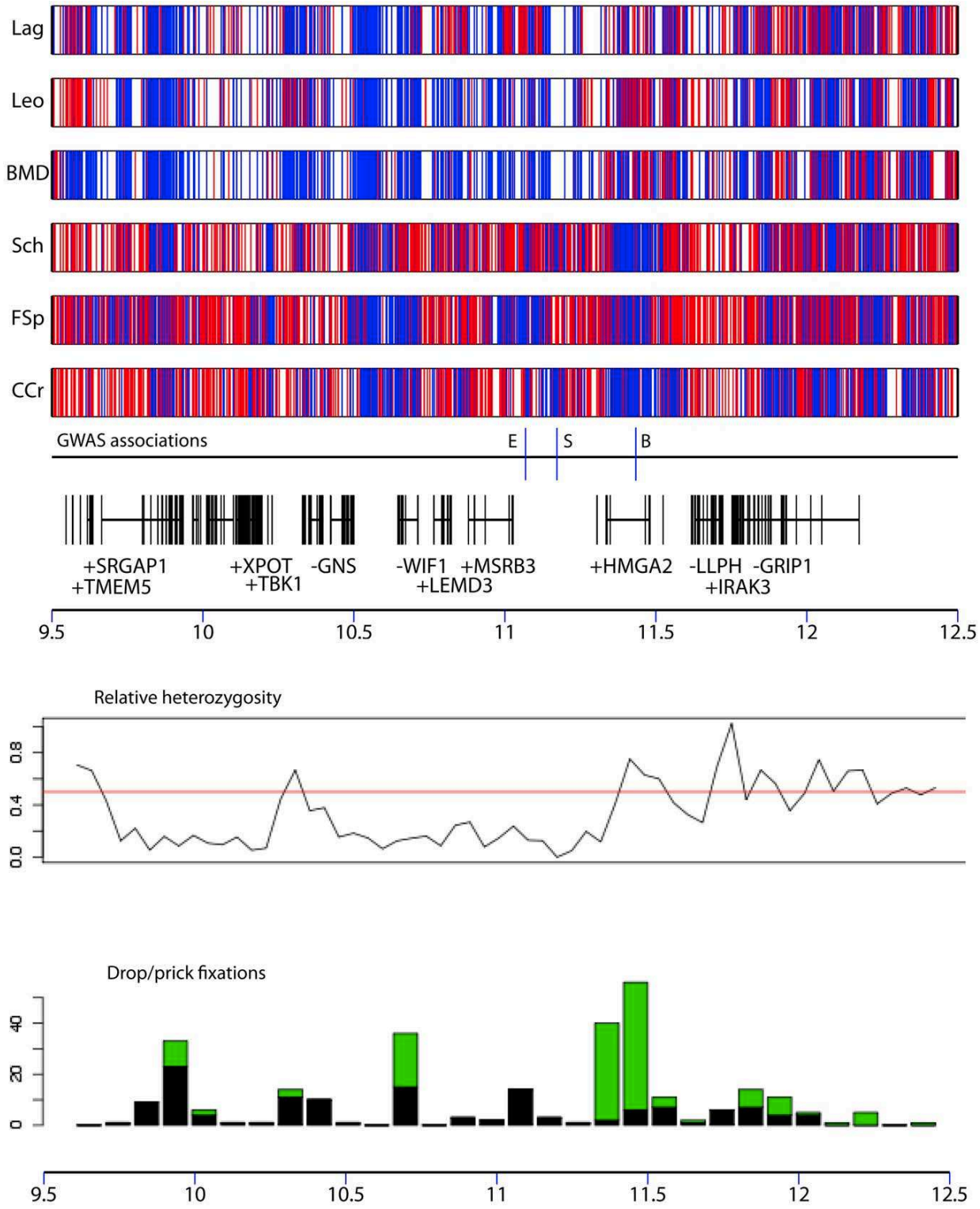


Figure 5. Patterns of polymorphism in a 3 Mb region where variation is associated with drop ear, body size, and boldness phenotypes. The top panel describes variation compared to the reference sequence in pools of 3 drop ear (Lagotto, Lag; Leonberger, Leo; Bernese Mountain Dog; BMD) and 3 pricked ear breeds (Schipperke, Sch; Finnish Spitz, FSp; Chinese Crested, CCr). Blue lines represent homozygous (fixed) differences from the reference sequence and red lines represent SNPs that are polymorphic in the breed pool. The positions of GWAS associations for drop ear (E) body size (S) and boldness (B) are shown. The positions of genes are also displayed (vertical bars correspond to exons). The second panel displays the levels of relative heterozygosity in all drop ear breeds compared with all prick ear breeds in 100 kb windows. The third panel shows the

number of SNPs that are fixed for different alleles in drop and prick ear breeds. Green segments represent SNPs fixed for the reference allele in drop ear breeds, and black segments represent SNPs fixed for the reference allele in pricked ear breeds.
doi:10.1371/journal.pgen.1002316.g005

suggests that functional genetic variation has also been affected by genetic drift. This background of fixation of haplotypes by drift makes it extremely difficult to distinguish the signal of a selective sweep from background variation, and they may often be indistinguishable.

We performed coalescent modeling using realistic estimates of recombination and demographic parameters in order to compare the length distributions of genomic segments identified by our analyses with those expected under neutrality. These simulations are by necessity an approximation of the actual evolutionary and demographic forces that shaped patterns of genetic variation in dog breeds. In particular, we do not model selection, which may reduce effective population size. Secondly, we assume a simplified demographic model, involving a single domestication bottleneck, and simultaneous breed creation. The true history of dog evolution is likely to be more complex than this, with some breeds showing closer relatedness than others. Nevertheless, long segments identified by the S_i and d_i that pass the 5% FDR cut off are strong candidates for selective sweeps, and contain a number of regions already associated with phenotypic traits.

Simulations indicate that large segments of reduced heterozygosity and elevated F_{ST} are expected under neutrality but longer segments of reduced heterozygosity, particularly those longer than 1 Mb, are not expected to occur due to drift alone and hence are more likely to reflect selection. In general we expect segments of reduced heterozygosity to contain causative variants under selection, however, in some cases we observe large blocks of reduced heterozygosity that appear to be broken up into adjacent regions separated by more variable regions. This pattern may reflect heterogeneity in ancestral haplotypes, which makes it difficult to pinpoint the focus of selection. Smaller blocks of elevated d_i often occur within extended regions of reduced heterozygosity. These probably reflect the fixation of variants that are otherwise rare in the dog population due to hitchhiking on the selected haplotype. However, most variants that are fixed by hitchhiking during a selective sweep are likely to be already common in the population, and therefore will not have a big effect on the d_i value of a region. This leads to stochasticity in the d_i statistic, which may explain the fact that even the longest d_i segments still do not pass a 5% FDR. When even denser surveys of SNP variation (e.g. from whole genome sequencing) are available, a more promising approach could be to identify selective sweeps using reductions in heterozygosity, and identify potential causative variants within these sweeps by their elevated F_{ST} (see e.g. [45]).

In addition to aiding in the dissection of the genetic components of phenotypic variation in dog breeds, we anticipate that our fine-scale map of genomic regions of extreme population differentiation and fixation of extended haplotypes will find utility for identification of disease causing variants. Firstly, regardless of whether they are caused by selection or drift, regions with reduced heterozygosity in a particular breed are problematic to interrogate with GWAS and may harbor disease-causing variants that are not tagged on a SNP array. Secondly, genetic variants responsible for breed characteristics may have pleiotropic effects that increase incidence of disease in that breed. Thirdly, disease-causing mutations may have risen in frequency in regions under selection by genetic hitchhiking on haplotypes bearing variants under artificial selection. These considerations suggest that our candidate selection regions warrant additional scrutiny in disease mapping studies. An example of the second effect has recently been highlighted in the Shar Pei breed, where strong artificial selection for genetic variants that likely affect expression of the HAS2 gene is responsible for both the characteristic wrinkled skin of the breed and an increased predisposition to periodic fever syndrome [16].

Our analysis of single-SNP F_{ST} across breeds identified a number of extended genomic regions of extreme population differentiation between dog breeds, which harbor variants responsible for commonly varying traits between dog breeds. Genetic variation in some of these regions correlates with multiple traits that vary between dog breeds, in some cases including both morphological and behavioral differences. There are several possible reasons for these multiple associations. One possibility is that these regions harbor multiple variants that each has an effect on different traits. Alternatively the associations could be the result of single mutations with pleiotropic effects that affect multiple traits. It is also possible that traits may correlate with each other for other reasons. For example, there may have been coordinated selection for more than one trait in a subset of breeds, or a subset of breeds may share a trait simply by chance. We have comprehensively surveyed genetic variation in a region of extreme population differentiation on chromosome 10, where genetic variation correlates with body size, drop ears and boldness. As boldness shows strong correspondence with drop ears it is unclear whether this trait is affected by an independent variant in this region. A more detailed analyses of the allele frequencies of SNPs associated with body size and drop ears is consistent with a hypothesis that these traits are controlled by two linked SNPs, which in combination produce three observed haplotypes associated with distinct phenotypes. It is therefore possible that additional regions of extreme population differentiation also harbor multiple variants affecting different traits. Careful genetic dissection of each region is necessary to identify all functional variants and the traits they affect. As extensive LD is found in these regions, it is difficult to determine how many functional variants are present and their precise location. Such analysis would therefore be aided by the use of multiple breeds or populations with less extensive LD in order to narrow down the associated intervals.

In its most extreme form, a selective sweep is characterized by the rapid fixation of a new mutation under selection along with linked genetic variants (a hard sweep). However, less extreme selective episodes (soft sweeps), such as incomplete selective sweeps or selection on standing variation may also be common [46,47]. It has been argued that polygenic adaptation, where subtle changes in allele frequencies occur at many loci, is the dominant form of phenotypic evolution in natural populations [48]. This type of evolution is likely when variation in a trait of interest is controlled by a large number of loci with small effect, which is now known to be the case with a number of highly heritable quantitative metabolic and morphological traits in humans. A long-term selection experiment in *Drosophila melanogaster* also uncovers evidence for this kind of adaptation [49]. Artificial selection in dogs appears to have caused genetic variants with much larger phenotypic effects to segregate at high frequencies, resulting in the simplification of the genetic architecture of phenotypic variation. In some cases, breed-defining characteristics such as chondrodysplasia, skin wrinkling and brachycephaly are likely to result from hard sweeps at breed creation. However, many variants with large phenotypic effects appear to show continuous variation between breeds that correlates with particular traits, including genetic

variants that associate with body size in the IGF1 locus on chromosome 15 and with drop ear on chromosome 10, suggesting that selection by attenuation of allele frequencies is also common. Hence, although hard sweeps are likely to be a more common form of selection in domestic compared with wild species, it is likely that more minor changes in allele frequencies across many loci also contribute to phenotypic evolution.

The huge phenotypic diversity present in dogs raises the question as to whether levels of functional genetic variation in the ancestral dog population were elevated, adding to the raw material that artificial selection could act on. Relatively higher levels of replacement amino acid changes are found in dogs compared with wolves, possibly indicating a relaxation of selective constraint [50,51]. There are also a large number of loci in the dog genome polymorphic for the active SINEC_Cf elements [52], which may also contribute to functional genetic variation, although it is not known whether functional variation due to these elements is increased in dogs compared with wolves. It has also been suggested (and disputed) that the dog genome has a high intrinsic mutation rate [53,54]. There is also great interest in looking for "domestication genes" by identifying loci under selection in domestic species compared to wild ancestors. Investigation of these processes that occurred in the ancestral dog population requires detailed comparisons of patterns of genetic variation in dogs and wolves. As the majority (>98%) of SNPs on the CanineHD array were discovered by comparisons of dog breeds, they are biased against fixed differences between dogs and wolves and wolf-specific SNPs. Additional SNP discovery in wolves is therefore necessary to unravel the evolutionary processes involved in early dog domestication. Whole genome resequencing of both dogs and wolves will be important for a more detailed understanding of these processes.

It is likely that artificial selection in dogs (and other domestic animals) has led to the proliferation of mutations with large effects. This has contributed to the success of the dog as a model for genetic dissection of phenotypic traits. Such variants are likely to be maladaptive in the wild, and may also increase susceptibility to disease. Hence examining regions under selection in breeds may aid in identification of genetic risk factors affecting susceptibility to disease. Studying the extreme variation in forms produced by artificial selection also gives us a window into studying the effects of selection in natural populations, as first realized by Darwin [55]. Understanding the effects of selection on the genomes of domestic animals should give us insight into understanding its effects on nondomestic species, including our own.

Materials and Methods

Ethics statement

Blood samples were taken from dogs by trained veterinarians according to relevant national and international guidelines.

SNP discovery

We scanned the existing list of 2.8 million high quality SNPs and identified 1,555 regions >40 kb (gaps) with no known SNPs in non-repetitive DNA (588 of these are on chromosome X). Gaps >100 kb were divided into a series of shorter ones resulting in a set of 2,375 genomic segments with no known SNPs of average length 50kb. We designed a Roche NimbleGen sequence capture array containing probes matching on average 2.1 kb within each segment, giving a total of 5 Mb. This array was used to enrich pools of DNA from Belgian Shepherds, Irish Wolfhounds, West Highland White Terrier, Shar-Pei and wolves. The samples were then sequenced using an Illumina Genome Analyzer and aligned to the CanFam2 dog reference sequence using MAQ. We identified 4,353 novel SNPs (973 on chromosome X). After updating the canine SNP map with these variants the number of gaps >40 kb was reduced to 714 (392 on chrX).

Design of CanineHD array

We selected SNPs from initial list of 2.8 million augmented by the resequencing to be included in the Illumina CanineHD array. We selected SNPs by scanning the genome using non-overlapping windows of length 11,500 bp (this length was calculated to return the desired number of SNPs). Every SNP in each window was scored and ranked according to a number of different criteria in order to maximize quality, coverage of the genome and a number of other factors according to a scoring criteria (Table S10). The main criteria considered for each SNP were Illumina design score, presence on a list of SNPs known to be informative for studies of canid phylogeny and presence on lists of previous dog SNP arrays (Affymetrix and Illumina). SNPs in repetitive DNA or those that required two bead types on the Illumina array were disfavored. We also included 13 Y chromosome specific SNPs presented in ref [56]. The resultant list was analyzed to identify possible duplicates or incompatibilities between primers. The problematic SNPs were removed, and the final SNP list was edited manually to produce a list of 200k bead types by removing SNPs with the smallest distance to other SNPs.

Genotyping

Genotyping was performed by Illumina Inc., USA (Gentrain dataset) and Centre National de Genotypage, France (LUPA dataset). All data is available at: http://dogs.genouest.org/SWEEP.dir/Supplemental.html.

GWAS

For each trait we performed a GWAS with plink (http://pngu.mgh.harvard.edu/~purcell/plink/), using a breed permutation procedure to determine genome-wide significance implemented using a perl script. Each sample within a breed was first assigned a phenotype corresponding to the breed-specific value of a trait. Traits were either coded as dichotomous or quantitative depending on how they were measured (see below). An association study was performed for each trait followed by a permutation procedure, where the phenotypes of each breed were randomized, always assigning an identical phenotype value to each sample within the same breed. For each GWAS, 1000 permutations were performed, and the real significance values at each SNP were compared to the maximum permuted values across all SNPs in order to calculate genomewide significance.

We used the full LUPA dataset of 46 breeds to perform breed GWAS. The phenotypic values used are shown in Table S2. Personality traits and size were considered as quantitative traits. Other traits were considered dichotomous, and breeds were divided as follows (breed abbreviations in Table 1):

Furnishing association. 4 breeds with furnishings (BoT, IrW, StP, TYo) compared with all other breeds. Drop ear association: drop ear breeds (Bgl, BMD, BrS, CKC, CoS, Dac, Dal, ECS, ESS, ESt, GoS, LMu, NFd, ShP, Wei) compared with pricked ear breeds (BeT, Chi, CWD, Elk, Eur, FSp, Gsh, GSl, Hus, Sam, Sar, Sci). Curly tail association: straight tail breed (BoT, DaC, Dal, EBT, EBD, Gry, LRe) compared with curly tail breeds (Elk, Eur, FSp, GSD, Mop, Sar, Sci, Scn, ShP, Hus, Sam, TYo). Boldness association: bold breeds (ASh, BeT, BMD, BoC, BoT, Box, Dal, Dob, Elk, GSh, Hus, JRT, Rtw, Sam, Sci, Scn, ShP, TYo) compared with non-bold breeds (Bgl, BrS, Chi, CKC, Dac,

EBD, ECS, ESS, ESt, FcR, GoS, Gry, GRe, IrW, LRe, NFd, NSD, StP, Wei).

The rs numbers corresponding to SNPs mentioned in the text are listed in Table S11.

Phasing and imputation

We phased the genotypes in the reduced dataset containing 471 dogs from breeds with 10 or more samples using fastPHASE [57] version 1 with the default parameters. We analyzed each breed and chromosome separately, dividing the X chromosome into the pseudo-autosomal region (PAR) and nonrecombining portion. Missing genotypes were imputed by the software, and we subsequently removed all SNPs that were not polymorphic or had less than a 100% call rate in all dog samples. In total, 19,176 invariant SNPs were removed: 14,309 on the autosomes and PAR, and 1027 on the nonrecombining X chromosome. An additional 3,840 SNPs were removed due to poor call rate. This dataset was used for subsequent selection scans and coalescent modeling analyses.

Scans for selective sweeps using S_i and d_i statistics

The S_i statistic is a measure of the proportion of SNPs that are variable in a region in a particular breed relative to all other breeds. We first divided the genome into 150 kb sliding windows, overlapping by 25 kb. Each window contained on average 10 SNPs; windows with less than 5 SNPs were not retained in the analysis. The same sliding window coordinates were used for the S_i and d_i analyses. Given a pair of breeds i and j and a given genomic window, we define relative heterozygosity as:

$$\theta_{ij} = \frac{h_i}{h_i + h_j}$$

where h_i is the number of polymorphic SNPs in breed i and h_j is the number of polymorphic SNPs in breed j in a given genomic window.

S_i for a given genomic window in breed i is then calculated as:

$$S_i = \sum_{j \neq i} \frac{\theta_{ij} - E[\theta_{ij}]}{sd[\theta_{ij}]}$$

where $E[\theta_{ij}]$ is the expected value of θ_{ij}, calculated by comparing all of the SNPs between breed i and j, and $sd[\theta_{ij}]$ is the standard deviation of all sliding windows. The S_i statistic was calculated in this manner for all predefined 150 kb sliding windows across the genome, for all 30 breeds in the dataset. The S_i statistic was calculated separately for the autosomal regions (including PAR) and the nonrecombining portion of the X chromosome, and was calculated in exactly the same way outlined above for the coalescent simulated data.

Using the same dataset, we calculated F_{ST} for each pairwise breed combination. To identify regions with elevated F_{ST} calculated the d_i statistic for each SNP (Akey et al), which is a standardized measure of pairwise F_{ST} values involving breed i and all other breeds:

$$d_i = \sum_{j \neq i} \frac{F_{ST}^{ij} - E\left[F_{ST}^{ij}\right]}{sd\left[F_{ST}^{ij}\right]}$$

where $E[F_{ST}^{ij}]$ and $sd[F_{ST}^{ij}]$ represent respectively, the expected value and the standard deviation of F_{ST} between breed i and j computed from all SNPs. For each breed, d_i values were calculated for the 150 kb windows used for the S_i analysis. We retained, for each breed, windows with an average d_i within the top 1% of all d_i values.

For each breed, we retained the top 1% of windows in each breed based on both S_i and d_i statistics. Overlapping windows were then combined to create a set of larger regions for each statistic. We applied this method to both the real and simulated data (see below) after which we compared the distribution of lengths. We then computed a marginal p-value for each region as the proportion of regions defined from the simulated dataset of the same breed that were longer. Finally we corrected the p-values using the Benjamini-Hochberg FDR method [33].

The UCSC graphical display of regions identified by the S_i and d_i statistics as well as d_i values for all SNPs from the CanineHD array are available at the following URL:

http://dogs.genouest.org/SWEEP.dir/Supplemental.html

Shared haplotype analysis

The aim of this analysis was to identify putative regions of Identity By Descent (IBD) within haplotypes inferred to be involved in selective sweeps in multiple breeds in order to narrow down the boundaries of putative sweep regions. This is based on the assumption that the selected variant was present on an ancestral haplotype prior to breed creation and is shared by multiple breeds. We first identified core regions that overlapped S_i sweeps (at the 5% FDR) and were completely homozygous in each breed. Once these fixed regions were defined they were then grouped into clusters of overlapping physical locations between breeds. Where possible, we then identified the region of maximal overlap between all homozygous regions in all of the breeds in a cluster that had been fixed for identical haplotypes.

Cross-population extended haplotype homozygosity (XP-EHH)

In order to calculate XP-EHH for SNPs in our dataset, we first removed SNPs with a minor allele frequency $< 5\%$ in the entire dataset. We calculated the EHH statistic between all SNP pairs across all breeds in the whole dataset. We retained SNP pairs with EHH between 0.03 and 0.05 for the XP-EHH analysis. We calculated normalized log XP-EHH scores between these SNP pairs from iHS scores as described by [18]. However, instead of comparing iHS score between pairs of populations, we compared iHS scores in a given breed and SNP pair to the average of iHS scores in all other breeds. The normalization step was performed for each chromosome in each breed separately. In order to confirm the presence of extended haplotypes in putative sweep regions, we averaged XP-EHH scores across these regions in each breed compared to the genomewide average.

Coalescent simulations

We performed whole genome simulations under a realistic demographic model, using variable regional recombination rates as inferred from the original data. The simulation process consisted of three main steps: (1) recombination rate inference, (2) breed bottleneck modeling and (3) main simulations.

Recombination rate inference. We used interval in the LDhat package [58] to estimate the total population scaled recombination rate ($\rho = 4Ner$) of each dog chromosome, as well as regional recombination rate variation across all chromosomes. For each chromosome, we randomly chose 100 haplotypes from the original data as input for interval. We split the input data into consecutive blocks of 2000 SNPs, each overlapping the previous

block by 200 SNPs. Recombination rate estimates from individual blocks were then concatenated to get chromosome wide rate estimates. Interval was provided a look up table downloaded from http://www.stats.ox.ac.uk/~mcvean/LDhat/instructions.html, which assumes a population mutation rate of 0.01. The PAR was analyzed separately from rest of the X-chromosome. To assess the general performance of interval on dog data, we averaged the regional rate estimates obtained here across 5 Mb windows to make it compatible with a previously published, coarse, linkage map [59]. The concordance between the population genetic map and the linkage map is good (Axelsson et al, unpublished). To convert the population scaled recombination rate estimated for the domesticated dog into breed specific rates we first estimated the effective population size of the domesticated dog ($Ne_{dog\ Autosomes}$ = 7752) by comparing the autosomal part of the population genetic map generated here, with that of the linkage map [59]. Then to take a potential bias in reproductive success between males and females into account we used the same approach to estimate the effective population size using only the X-chromosome ($Ne_{dog\ X}$ = 9134).

Modeling breed bottlenecks. We built on previous dog demographic modeling efforts in setting up a simple simulation scheme to estimate the strength of bottlenecks at breed formation for each breed in our dataset. Our model thus assumed an effective population size of the ancient wolf population (Ne_{wolf}) of 22600 [60]. It furthermore assumed that dog domestication occurred 5000 generations ago, accompanied by an instantaneous decrease in population size to $5560Ne_{dog}$ and that breed formation took place 100 generations ago [60]. The mutation rate was set to 10^{-8} [1] and the generation time was assumed to be 3 years. We then used MaCS [61] to simulate genome wide replicas of our dataset according to the model described above, for 59 breed bottle neck sizes, ranging from $0.001Ne_{wolf}$ - $0.03Ne_{wolf}$ (this corresponds to an increment in bottle neck size of $0.0005Ne_{wolf}$ for every new simulation). We repeated these simulations for each of the sample sizes represented in the original data (ranging from 10 to 52 haplotypes), in total rendering 767 simulated datasets. All simulations were run using regional recombination rates as inferred in the real data. We also corrected simulated data for ascertainment bias in the original data by providing MaCS with allele frequency distributions from the original data. Next, we estimated LD decay, measured as r^2, for markers separated up to a maximum of 500 kb, in the original, as well as all simulated datasets. We subsequently used least squares to fit LD decay curves of real and simulated datasets. The best fitting simulation provided an estimation of bottleneck size for each breed individually.

Main simulation. By implementing the estimated breed bottleneck sizes in the model described above we were then able to simulate a single complete dataset including all breeds in the original dataset. As before, regional recombination rates were inferred from the original data, and ascertainment bias in the original data was corrected for in the simulated data. Finally, we thinned the simulated dataset to match the marker density of the real dataset (one marker every 13,046 bp). The PAR and X-specific parts of the X-chromosome were simulated separately. For the X-specific simulations we adjusted all population sizes according to the difference between $Ne_{dog\ Autosomes}$ and $Ne_{dog\ X}$.

Functional analysis of gene categories

We selected human orthologs with a 1:1 human-dog orthologous relationship to perform GO analyses. Biomart version 0.8 (Ensembl v.62) was used to collect orthologous human protein-coding genes. WebGestalt [62], a web-based gene set analysis toolkit, was used to retrieve GO terms associated with human ensembl gene stable IDs. A hypergeometric test computed the statistical significance of over-representations of GO terms that were compared to a background list of genes selected to control for possible gene length bias as observed in the selected gene set. The background set was composed of human genes selected using biomart with 1:1 human-dog orthologous relationship, longer than 100 kb and with a mean size of 230 kb, similar to the tested set. GO biological processes that were significantly over-represented ($p<0.05$) were considered.

Resequencing of a candidate selection region on chromosome 10

We selected a 3 Mb region on chromosome 10 (9.5-12-5 Mb) for resequencing in 6 breeds. We first prepared pools of DNA containing 5 samples from each of the breeds (Chinese Crested, Lagotto, Schipperke, Finnish Spitz, Leonberger and Bernese Mountain Dog). We next performed sequence capture using a Roche NimbleGen array containing probes designed to hybridize to this region. This was followed by sequencing using the Illumina Genome Analyzer. Reads were mapped to the dog genome reference sequence using bwa (http://bio-bwa.sourceforge.net/) followed by SNP calling using samtools (http://samtools.sourceforge.net/). Mapping and SNP calling was done independently for each breed and custom scripts were used to identify SNPs with certain patterns of segregation. SNPs in conserved elements were identified relative to those defined by the dog genome analysis based on human-dog-mouse-rat alignments, and on identification of phastcons elements within mammals based on alignments of 44 vertebrates, converted from human to dog coordinates by LiftOver (http://genome.ucsc.edu/).

Supporting Information

Figure S1 Coverage of HD array. Number of SNPs in 100 kb windows across the genome contained in the Illumina CanineHD array and the Affymetrix V2 Canine array on A) autosomes and B) X chromosome. For autosomes, there is an average of 9 SNPs per 100 kb window on the HD array and only 5% of windows do not contain SNPs, whereas the majority (>25%) of 100 kb windows do not contain SNPs on the Affymetrix array. For the X chromosome, >75% of windows do not contain SNPs on the Affymetrix array, whereas <5% of windows do not contain SNPs on the HD array.
(PDF)

Figure S2 Decay of linkage disequilibrium in real versus simulated data. Decay of linkage disequilibrium (LD decay) across the autosomes of all breeds included in this study (solid lines) compared with that of simulated datasets (circles). Wolf is included as comparison. LD decay was calculated as r^2 for markers separated by at most 400kb and averaged across bins of 10kb. Simulation were run in MaCS [61] with the following general model; ancient Wolf Ne: 22600 [60], ancient domesticated dog Ne: 5650 [60], dog domestication 5000 generations ago [60], dog breed formation 100 generations ago [60], mutation rate: 1×10^{-8} per site per generation [1], generation time: 3 years. Each breed was then assigned (1) a specific breed bottleneck size (determined by simulation, see Table S1) from the following range: 0.001-0.03 x (ancient Wolf N_e) as well as (2) a sample size to match that of the real dataset (10-52 haplotypes). Furthermore, recombination rates were allowed to vary locally as inferred in real data using LDhat [58]. We corrected for ascertainment bias by supplying MaCS allele frequencies from the real dataset.
(PDF)

Figure S3 Allele frequency of 3 SNPs strongly associated with body size in the dataset across breeds plotted against body size. Data on body size is presented in Table S3.
(PDF)

Figure S4 Signal of association with curly tail of chr1: 95-98 Mb. The y-axis shows the raw p-value and genes in the region are shown below the graph.
(PDF)

Figure S5 Signal of association between drop ear, boldness and size in the region chr10:9.5-12.5 Mb. The y-axis show the raw p-value for each association, and genes in the region are display beneath the graph.
(PDF)

Figure S6 Signal of association with sociability on chrX: 102-110 Mb. The y-axis shows the raw p-value and genes in the region are shown below the graph.
(PDF)

Figure S7 Map of extended segments in the dog genome of significantly reduced S_i that pass the 5% FDR cut-off. Each region is color coded according to the breed in which the significant reduction is observed. The breed codes are shown in Table 1.
(PDF)

Figure S8 Examples of S_i and d_i statistics in top 10 longest S_i regions. Variation in these statistics is shown independently in genomic segments encompassing each region.
(PDF)

Table S1 Total samples in dataset.
(DOCX)

Table S2 Breed bottleneck sizes used in simulations.
(DOCX)

Table S3 Phenotypes used in across-breed GWAs.
(DOCX)

Table S4 Single-SNP F_{ST}. SNPs with minor allele frequency >0.15 and F_{ST}>0.55. Nearby SNPs <500kb are clustered into regions.
(DOCX)

Table S5
(XLSX)

Table S6
(XLS)

Table S7
(XLS)

Table S8
(DOCX)

Table S9
(DOCX)

Table S10
(DOCX)

Table S11
(XLSX)

Acknowledgments

We thank all of the dog owners who contributed samples used in this project. We also thank three anonymous reviewers for valuable comments. For a list of partners in the The LUPA Consortium, please see http://www.eurolupa.org/.

Author Contributions

Conceived and designed the experiments: KL-T CH MTW. Performed the experiments: GRP SS MSTH CTL. Analyzed the data: AV (di, FDR analysis), AR (phasing and imputation, Si, XP-EHH, shared haplotype analysis), TD (browser display), EA (coalescent simulations), KL-T (input on all analyses), CH (GO analysis), MTW (SNP discovery analysis, HD array design, dataset construction, breed relationships, across-breed GWAS, FST, chromosome 10 resequencing, FDR analysis). Contributed reagents/materials/analysis tools: TF EKK EHS DB CV HL FG MF JH ÅH CA KL-T. Wrote the paper: AV AR KL-T CH MTW.

References

1. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803.
2. Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, et al. (1997) Multiple and Ancient Origins of the Domestic Dog. Science 276: 1687–1689.
3. Karlsson EK, Lindblad-Toh K (2008) Leader of the pack: gene mapping in dogs and other model organisms. Nat Rev Genet 9: 713–725.
4. Wayne RK, Ostrander EA (2007) Lessons learned from the dog genome. Trends Genet 23: 557–567.
5. Sutter NB, Eberle MA, Parker HG, Pullar BJ, Kirkness EF, et al. (2004) Extensive and breed-specific linkage disequilibrium in Canis familiaris. Genome Research 14: 2388–2396.
6. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, et al. (2010) A simple genetic architecture underlies morphological variation in dogs. PLoS Biol 8: e1000451. doi:10.1371/journal.pbio.1000451.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
8. Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, et al. (2007) A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. Science 316: 112–115.
9. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, et al. (2009) Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes. Science 326: 150–153.
10. Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, et al. (2007) A -defensin mutation causes black coat color in domestic dogs. Science 318: 1418–1423.
11. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, et al. (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet 39: 1321–1328.
12. Jones P, Chase K, Martin A, Davern P, Ostrander EA, et al. (2008) Single-nucleotide-polymorphism-based association mapping of dog stereotypes. Genetics 179: 1033–1044.
13. Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, et al. (2009) An Expressed Fgf4 Retrogene Is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs. Science 325: 995–998.
14. Bannasch D, Young A, Myers J, Truve K, Dickinson P, et al. (2010) Localization of canine brachycephaly using an across breed mapping approach. PLoS ONE 5: e9632. doi:10.1371/journal.pone.0009632.
15. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, et al. (2010) Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A 107: 1160–1165.
16. Olsson M, Meadows JR, Truve K, Rosengren Pielberg G, Puppo F, et al. (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 7: e1001332. doi:10.1371/journal.pgen.1001332.
17. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
18. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
19. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72. doi:10.1371/journal.pbio.0040072.
20. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
21. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, et al. (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19: 838–849.
22. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PLoS Genet 5: e1000500. doi:10.1371/journal.pgen.1000500.

23. Innan H, Kim Y (2008) Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics 179: 1713–1720.
24. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
25. vonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898–902.
26. Fogle B (2000) The new encyclopedia of the dog. London: Dorling Kindersley.
27. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, et al. (2008) Many sequence variants affecting diversity of adult human height. Nat Genet 40: 609–615.
28. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40: 584–591.
29. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40: 575–583.
30. Svartberg K, Forkman B (2002) Personality traits in the domestic dog (Canis familiaris). Applied Animal Behaviour Science 79: 133–155.
31. Svartberg K, Tapper I, Temrin H, Radesater T, Thorman S (2005) Consistency of personality traits in dogs. Animal Behaviour 69: 283–291.
32. Jackson IJ (1994) Molecular and developmental genetics of mouse coat color. Annu Rev Genet 28: 189–217.
33. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc [Ser B] 57: 289–300.
34. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
35. Roberts MC, Mickelson JR, Patterson EE, Nelson TE, Armstrong PJ, et al. (2001) Autosomal dominant canine malignant hyperthermia is caused by a mutation in the gene encoding the skeletal muscle calcium release channel (RYR1). Anesthesiology 95: 716–725.
36. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, et al. (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4: e1000144. doi:10.1371/journal.pgen.1000144.
37. Obholz KL, Akopyan A, Waymire KG, MacGregor GR (2006) FNDC3A is required for adhesion between spermatids and Sertoli cells. Developmental Biology 298: 498–513.
38. McPherron AC, Lee SJ (1997) Double muscling in cattle due to mutations in the myostatin gene. Proc Natl Acad Sci U S A 94: 12457–12461.
39. Mosher DS, Quignon P, Bustamante CD, Sutter NB, Mellersh CS, et al. (2007) A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. PLoS Genet 3: e79. doi:10.1371/journal.pgen.0030079.
40. Xu W, Gong L, Haddad MM, Bischof O, Campisi J, et al. (2000) Regulation of microphthalmia-associated transcription factor MITF protein levels by association with the ubiquitin-conjugating enzyme hUBC9. Exp Cell Res 255: 135–143.
41. Balikova I, Lehesjoki AE, de Ravel TJ, Thienpont B, Chandler KE, et al. (2009) Deletions in the VPS13B (COH1) gene as a cause of Cohen syndrome. Hum Mutat 30: E845–854.
42. Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, et al. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587–591.
43. Wang B, Zhang YB, Zhang F, Lin H, Wang X, et al. (2011) On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS ONE 6: e17002. doi:10.1371/journal.pone.0017002.
44. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
45. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327: 883–886.
46. Pritchard JK, Pickrell JK, Coop G (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20: R208–R215.
47. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.
48. Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. Nat Rev Genet 11: 665–667.
49. Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, et al. (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467: 587–590.
50. Bjornerfeldt S, Webster MT, Vila C (2006) Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Res 16: 990–994.
51. Cruz F, Vila C, Webster MT (2008) The legacy of domestication: accumulation of deleterious mutations in the dog genome. Mol Biol Evol 25: 2331–2336.
52. Wang W, Kirkness EF (2005) Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res 15: 1798–1808.
53. Shearin AL, Ostrander EA (2010) Canine morphology: hunting for genes and tracking mutations. PLoS Biol 8: e1000310. doi:10.1371/journal.pbio.1000310.
54. Hedrick PW, Andersson L (2011) Are dogs genetically special? Heredity 106: 712–713.
55. Darwin CR (1859) On the Origin of Species. London: John Murray.
56. Natanaelsson C, Oskarsson MC, Angleby H, Lundeberg J, Kirkness E, et al. (2006) Dog Y chromosomal DNA sequence: identification, sequencing and SNP discovery. BMC Genet 7: 45.
57. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
58. Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome Res 17: 1219–1227.
59. Wong AK, Ruhe AL, Dumont BL, Robertson KR, Guerrero G, et al. (2010) A comprehensive linkage map of the dog genome. Genetics 184: 595–605.
60. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, et al. (2009) Linkage disequilibrium and demographic history of wild and domestic canids. Genetics 181: 1493–1505.
61. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Res 19: 136–142.
62. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33: W741–748.

Mamm Genome
DOI 10.1007/s00335-011-9372-0

Annotation of the domestic dog genome sequence: finding the missing genes

Thomas Derrien · Amaury Vaysse · Catherine André · Christophe Hitte

Received: 5 September 2011 / Accepted: 23 October 2011

Abstract There are over 350 genetically distinct breeds of domestic dog that present considerable variation in morphology, physiology, and disease susceptibility. The genome sequence of the domestic dog was assembled and released in 2005, providing an estimated 20,000 protein-coding genes that are a great asset to the scientific community that uses the dog system as a genetic biomedical model and for comparative and evolutionary studies. Although the canine gene set had been predicted using a combination of ab initio methods, homology studies, motif analysis, and similarity-based programs, it still requires a deep annotation of noncoding genes, alternative splicing, pseudogenes, regulatory regions, and gain and loss events. Such analyses could benefit from new sequencing technologies (RNA-Seq) to better exploit the advantages of the canine genetic system in tracking disease genes. Here, we review the catalog of canine protein-coding genes and the search for missing genes, and we propose rationales for an accurate identification of noncoding genes though next-generation sequencing.

Introduction

The assembled sequence of the domestic dog (*Canis familiaris*) genome and a catalog of over two million single nucleotide polymorphisms (SNPs) were completed and published in 2005 (Lindblad-Toh et al. 2005). This provided a useful resource for the scientific community and reinforced the position of the dog as an important model organism for biomedical research of human diseases (Sutter and Ostrander 2004; Galibert and André 2006). The genetic structure of the domestic dog is particularly relevant to the investigation of human diseases, with over 350 isolated dog breeds. Each of these has a reduced genetic variation compared with humans, which simplifies the mapping of simple and complex diseases (Drogemuller et al. 2008; Parker et al. 2009; Abitbol et al. 2010; Beggs et al. 2010; Wilbe et al. 2010; Lequarré et al. 2011; Merveille et al. 2011; Seppala et al. 2011). The scientific community has recognized this potential and has developed molecular tools to successfully investigate this powerful model of human diseases (Guyon et al. 2003; Breen et al. 2004; Hitte et al. 2005; Lindblad-Toh et al. 2005; Hitte et al. 2008). Indeed, inherited canine disorders include monogenic and complex diseases that are homologous to human diseases such as cardiovascular and neurological diseases, inflammatory disorders, immune system failures, metabolic defects, and numerous cancers such as mammary tumors, melanoma, lymphoma (Lequarré et al. 2011). All of these defects are frequently observed in humans with symptoms, pathophysiology, and clinical responses closely related to dog disorders. Substantial efforts have been made to reannotate the canine protein-coding gene set in recent years (Goodstadt and Ponting 2006; Derrien et al. 2007a; Derrien et al. 2009). However, there is still a need for a complete annotation of the dog protein-coding and non-protein-coding gene sets (ncRNAs) to fully exploit the dog as a model for human and veterinary medicine. As shown by the ENCODE sequencing consortium (ENCODE Project Consortium et al. 2007), about half of the mammalian genome is transcribed into RNA molecules, most of which are not translated into proteins. In addition, functional classification of sequence variations identified by genome-wide association studies (GWAS)

T. Derrien · A. Vaysse · C. André · C. Hitte (✉)
Institut de Génétique et Développement de Rennes, CNRS-UMR6061, Université de Rennes 1, 2 av Pr. Léon Bernard, 35043 Rennes, France
e-mail: hitte@univ-rennes1.fr

Published online: 11 November 2011

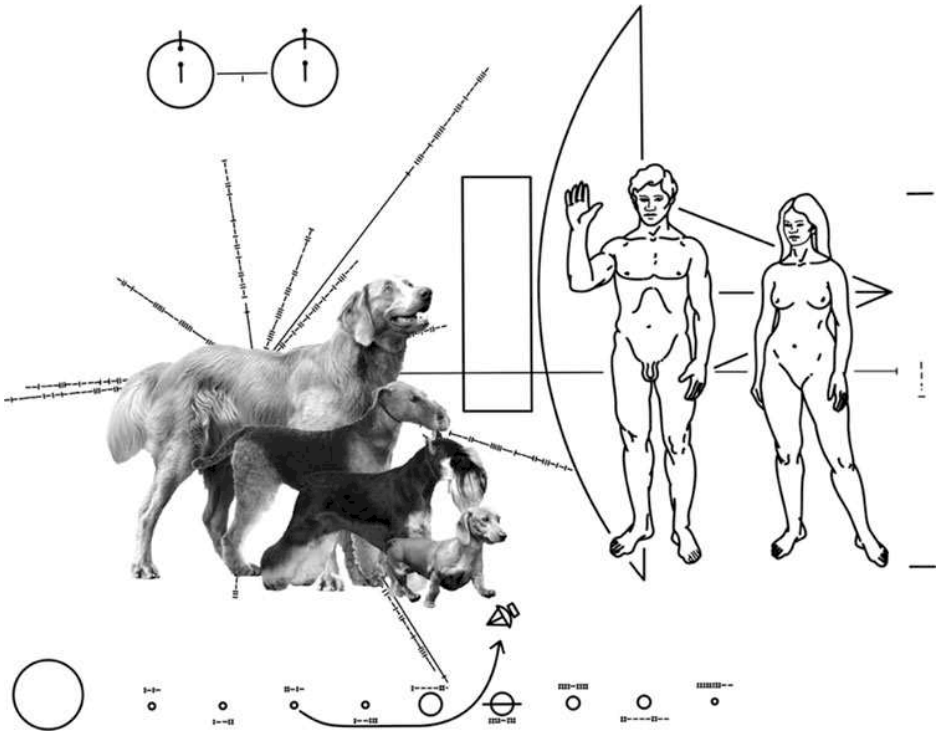


Fig. 1 Dog breed diversity helps to cure humans. The photo of four dog breeds, meant to represent canine diversity, is superimposed on the famous pictorial message sent on the Pioneer 10 spacecraft. The Pioneer plaque shows a man and a woman, a hyperfine transition of neutral hydrogen (*circles top left*), the relative position of the Sun with respect to the center of the galaxy and 14 pulsars (*lines*), the solar system (*circles bottom*), and silhouette of the spacecraft. It aimed to represent how the human species portrays itself and its place in the galaxy

shows that the vast majority of sequence variations are located in noncoding regions (ncRNAs or regulatory regions such as promoters or enhancers) (Hindorff et al. 2009; Manolio et al. 2009).

This review describes the gene predictions of the dog genome sequence, with a specific emphasis on missing gene prediction. It also advocates for Next Generation Sequencing (NGS) approaches to refine coding genes and identify noncoding genes to provide a complete and accurate gene set of the canine genome.

The domestic dog as a powerful genetic and evolutionary model

Isolated and inbred human populations are often investigated to understand the genetics background of rare diseases that have specifically segregated in small geographic, ethnic, or religious human isolates or families. As an alternative genetic approach to rare and complex diseases, the dog is a powerful spontaneous model with numerous naturally occurring genetic diseases and with a population structure that is split into breeds of diverse phenotypes (Fig. 1). Over 350 dog breeds constitute as many genetic isolates maintained by inbreeding, popular sire effects, and repeated matings, aiming at the fixation of particular phenotypic or behavioral traits in each breed. Most canine genetic diseases are breed-specific due to founder effects at the creation of the breed (Parker et al. 2004). These features make it easier to map the causative genes than in a heterogeneous human population (Sutter and Ostrander 2004; Galibert and André 2006). The domestic dog therefore represents a unique system to decipher phenotype/genotype relationships by family-based genetic linkage approaches and by GWAS within or across breed or by searching for the smallest shared haplotypes associated with fixed traits across breeds (Mosher et al. 2007; Parker et al. 2007).

Besides being a powerful model for human diseases, the dog also represents an important genetic resource for identifying morphology and behavior traits that have been stringently selected to conform to particular criteria in different breeds (Jones et al. 2008). Indeed, intense selective breeding has created hundreds of dog breeds (Fig. 1). This large-scale artificial selection has provided an ideal resource that geneticists can use to search for the genetic bases that control these differences (Akey et al. 2010; Boyko et al. 2010; Vaysse et al. 2011). There are two main categories of genetic variants under selection in dogs: first, variants that control variations in common traits such as size and ear carriage, which segregate across many breeds, and second, variants that control rare traits that are present in one or a few breeds, such skin wrinkling in Shar-Pei (Akey et al. 2010; Olsson et al. 2011) or chondrodysplasia in most short-legged breeds (Parker et al. 2009). In addition, much of the variation in traits related to canine morphology appears to be governed by a small number of genetic variants with large effect (Boyko et al. 2010).

In contrast, morphological variations in human are likely controlled by hundreds of loci with small effects, such as height (Manolio et al. 2009). This is likely because new variants with large effects in dog are rapidly fixed or preserved by artificial selection. The identification of targets of artificial selection among dog breeds therefore allows detection of the genetic variants which may be involved in phenotypic variations (Akey et al. 2010; Vaysse et al. 2011).

Performing a single nucleotide polymorphism (SNP) GWAS with DNA from a few individuals across breeds can be used to identify and pinpoint the genetic variants that contribute to traits shared by several breeds. This has been well documented in several recent studies of body size, coat type, and coat color as well as various morphological traits such as height and the shape of the ears, snout, and limbs (Cadieu et al. 2009; Bannasch et al. 2010; Boyko et al. 2010). Genotyping data are also used to search for regions of selective sweeps, characterized by long segments of reduced heterozygosity and indicative of selection. Selective sweeps can be associated with a specific trait within a single breed or within a group of breeds that share a trait (Akey et al. 2010; Boyko et al. 2010; Olsson et al. 2011; Vaysse et al. 2011). Indeed, most dog breeds were created within a few centuries or less and this likely involved rapid fixation of genetic variants and haplotypes under strong artificial selection.

Once a gene involved in a monogenic trait or a susceptibility gene has been discovered and validated in the domestic dog system, it then becomes a good candidate to explain the genetic background of similar diseases in humans, provided there are accurate gene repertoires for dog and human. To increase the impact of the canine model, the precise and exhaustive mapping of coding and noncoding genes, transcribed pseudogenes, regulatory regions, alternative splicing, gain and loss events, and the identification of lowly expressed genes must be achieved.

Canine genes prediction

Gene identification in eukaryotic genomic sequences is difficult due to (1) the small amount of DNA ($\sim$2%), that actually encodes proteins, (2) the complex identification of all alternative spliced isoforms of a gene, and (3) the appreciation of the non-protein-coding part of the genome. Current gene annotation in dog also suffers from unsequenced GC-rich regions that are often located in untranslated regions (UTR) and the 5′ end of gene sequences. In addition, gene identification is challenging in dog because of the low number of expressed sequence tags (EST) that have been produced. In comparison, there are large human EST sequence libraries that have given us the opportunity to identify the number of human genes and examine the overall incidence and frequency of alternative splicing in human genes. There are 20 times fewer ESTs annotated in dbEST (Boguski et al. 1993) for the domestic dog compared with human (as of November 2010). Due to this lack of experimental evidence for gene annotation, most of the 20,000 protein-coding genes of the canine genome (Kirkness et al. 2003; Lindblad-Toh et al. 2005) have been identified using ab initio predictions at the genome-scale level and sequence similarities analyses between species using various computational methods and tools. The canine gene set prediction was based mainly on the Ensembl genebuild pipeline (Flicek et al. 2011) and was augmented by additional, evidence-based genes from the Broad Institute annotation pipeline and from the phylogenetic orthology prediction pipeline (PhyOP) based on synonymous rate estimates (Goodstadt and Ponting 2006). These methods include gene prediction algorithms such as GeneId (Blanco et al. 2007), Genescan (Tiwari et al. 1997), and sequence homology searches based on local alignment tools. The gene prediction algorithms use different types of information, including gene-specific signals in the genomic sequence such as start and stop codons, splice sites, codon usage, and similarity to known genes and proteins in the databases. Some programs such as Genewise and GenomeWise (Birney et al. 2004) combine the protein sequence similarity analysis with ab initio gene-finding algorithms to improve predictions.

Algorithms not based on sequence alignment, including comparative genomics approaches, allow prediction of genes on the basis of patterns across multiple genomes. Indeed, gene order between species is not random (Alekseyev and Pevzner 2007) due to the relatively slow rate of evolutionary rearrangements. Gene order conservation has also been shown to correlate with coexpressed and coregulated genes, suggesting a functional significance (Hurst et al. 2004) such as the functions of bidirectional promoter (Semon and Duret 2006). On the other hand, the conservation of gene order between species is also used to identify protein-coding genes relocated during evolution in nonsyntenic chromosomal regions, as well as to localize retrotransposed genes and pseudogenes inserted in nonsyntenic regions. By taking advantage of whole-genome sequence assemblies, fine-scale comparative maps are an essential tool to identify conserved segments as well as colinearity of gene order between species (Derrien et al. 2007b; Muffato et al. 2010). Gene-order-based studies in dog have refined the protein-coding repertoire through multiple pairwise synteny maps that identify short and targeted orthologous genomic intervals (Fig. 2) (Derrien et al. 2009). The intervals were delimited by one-to-one (1:1) orthologous protein-coding genes. This allowed the identification of

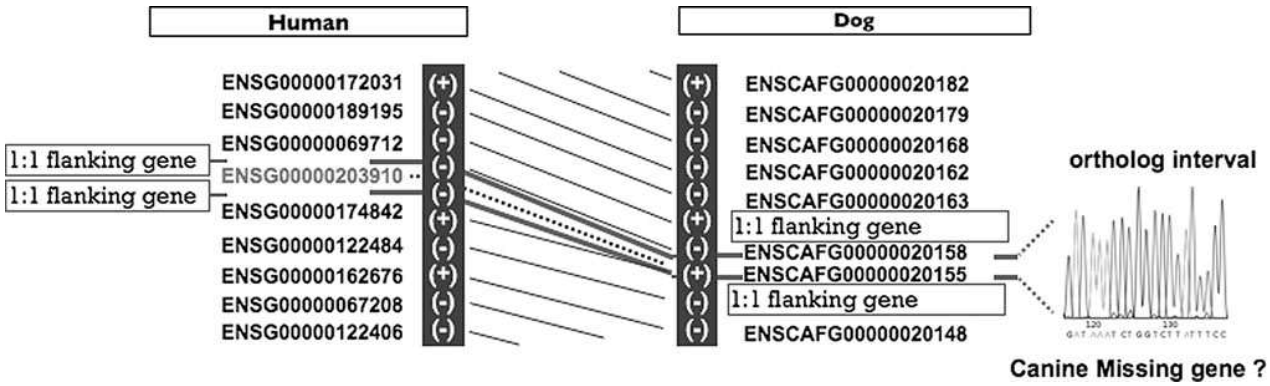


Fig. 2 Inferring missing gene localization through synteny maps: a dog–human example. The figure illustrates the method by which to infer a target interval for gene prediction. In the method, the first step builds a pairwise synteny map as exemplified by a schematic human–dog syntenic map. The 1:1 orthologs are connected through lines. Dog missing gene is positioned in gray on the reference species of the synteny map. The flanking 1:1 orthologs are used to define an orthologous interval on the canine chromosome as indicated by bold lines

short consensus intervals on the canine genome in order to focus on genes that had not been initially annotated (i.e., not detected in the whole-genome assembly of the dog but annotated in four rodents and primate species). The use of short orthologous genomic intervals reduces the cost of detecting false positives as it mostly filters out paralogs and allows balancing toward sequence alignment sensitivity versus accuracy. Alternatively, for more divergent sequences, the alignment criteria may be relaxed in a short predefined space where the background noise is significantly reduced compared with a genome-scale search. Focusing on those missing canine genes, we used (Derrien et al. 2009) the Genewise program (Birney et al. 2004), a sequence similarity-based method that explicitly models the conservation of gene structure, a high degree of conservation, and a probabilistic-pair hidden-Markov model (HMM) that show a weaker dependence on percent identity (Meyer and Durbin 2002). As a result, 232 new genes could be predicted with the delineation of new orthologous relationships with rodents, chimpanzee, and human species.

Selective constraints analysis to assess gene and gene loss prediction

In addition to gene structural analysis, selective constraints acting on protein-coding genes have been investigated through the d_N/d_S index (Yang 1997; Yang and dos Reis 2011). This index measures the ratio of amino acid replacement (nonsynonymous, d_N) to silent substitution (synonymous, d_S) and provides a good proxy of selective pressure acting on a given coding sequence. Overall, mutations in genes causing amino acid replacements with functional consequences are selected against in contrast to mutations occurring in nonfunctional pseudogenes. It is now standard practice to use the distinctive patterns of d_N/d_S ratios to refine genome annotation (Goodstadt and Ponting 2006; Enard et al. 2010). In the study described above, we predicted 232 new genes for which we have used d_N/d_S ratios to assess those compared with their human functional orthologous gene from pairwise transcripts alignments (Derrien et al. 2009). We calculated a median d_N/d_S of 0.19. This was very similar to the benchmark set (composed of all 1:1 dog:human orthologs) value of 0.15, which indicates similar selective constraints on both benchmark and predicted genes sets and reinforces the quality of the new predicted dog protein genes. We also detected 55 new predictions containing open reading frame (ORF)-disrupting mutations leading to pseudogene identification, of which 21 were predicted with accumulated mutations (mean = 4.2; range = 2-11). Independent of the presence of stop codons or frameshifts, we assessed the validity and the selective constraints acting on the 21 pseudogene predictions by calculating the d_N/d_S ratio for each of the candidate pseudogene predictions compared with their human functional orthologous gene from pairwise transcript pair alignments. A median d_N/d_S value of 0.5 was determined for the predicted pseudogene set, indicating a considerable relaxation of selective constraints of the canine predicted pseudogenes in comparison to the 1:1 dog:human orthologs benchmark set (Mann–Whitney test, $p = 5.17e-6$). We also found that the canine predicted pseudogenes show deviations from the expected rate of evolution using a phylogenetic context, including human and mouse gene sequences ($d_N/d_S = 0.41$ for dog, 0.19 for mouse, and 0.26 for human; Kruskal–Wallis test, $p = 1.04e-2$). This set of 21 canine gene predictions with higher d_N/d_S values, as characterized by pairwise and phylogenetic approaches, and high mutation rate was

classified as gene-loss candidates. An additional set of 48 undetected genes or genes detected with insufficient protein identity (average = 21.7%) was also considered dog gene-loss candidates, leading to 69 lost genes in comparison to human, chimpanzee, mouse, and rat. Among the 69 predicted canine gene losses, we found 28 genes that have been functional for over 170 million years, a time period that extends from platypus to human. Functional annotation of the human orthologous genes showed that they are involved in the biological process of response to stimulus, are transcription factors, and are involved in the regulation of various aspects of embryonic development.

Deep sequence annotation of the domestic dog genome

Part of the genetic basis that governs variations in morphology, physiology, and disease susceptibility of the canine breeds cannot be explained by mutations into protein-coding sequences. This makes it difficult to analyze the genetic bases for variations using the current annotation of the dog genome. The identification of all processing of the primary RNA transcripts into messenger RNAs (mRNAs) and non-protein-coding RNAs (ncRNAs) is required to provide an exhaustive annotation of canine genes. So far, ESTs are the principal source used to annotate experiment-based canine genes. However, at present, only 500,000 canine ESTs have been described in dbEST ($\sim$ 20 times fewer than in human). As a result, the NCBI Reference Sequence (RefSeq) contains only 1,208 entries ("NM" entries for the RefSeq curated nucleotide sequence record for mature mRNA) for the canine genome. This registry is a manually curated sequence database of nonredundant, extensively cross-linked, and deeply annotated nucleic acid and protein records. In comparison, 34,102 RefSeq "NM" entries are available for the human genome. On the other hand, entire transcriptome analyses through NGS produce the throughput required for deep RNA sequencing (RNA-Seq), which provides discovery and quantification of the entire RNA repertoire (Denoeud et al. 2008; Cloonan et al. 2009). RNA-Seq technologies therefore help to track the many novel RNA species that carry biological function. In addition, the whole-gene structures can also be discovered using NGS-based technologies for the gene 5′ ends using 5′ RACE-Seq (ENCODE Project Consortium et al. 2011), CAGE (Kodzius et al. 2006), and Paired-End diTags (Ng et al. 2005) to aid in defining the 5′ gene boundaries. RNA-Seq helps to characterize the vast majority of the genome that is transcribed beyond the structure of known genes. This phenomenon is termed "pervasive transcription"(ENCODE Project Consortium et al. 2007; Gingeras 2007). However, pervasive transcription of the genome recently has been refuted by van Bakel et al. (2010) who argued that the majority of the detected low-level transcripts (concerning especially ncRNAs transcripts) are more likely the result of technical artifacts or background biological noise rather than bona-fide functional transcripts. However, Clark et al. (2011) contested van Bakel et al.'s arguments by highlighting several lines of evidence that support the functionality of this "dark matter" transcription of the eukaryotic genomes. First, ncRNAs show high sequence conservation (although it is lower than in protein-coding genes), especially at important gene domains such as promoters, splice junctions, and exons. Second, ncRNAs have been characterized with predicted secondary structures and conserved genomic positions on the genome (Carninci et al. 2005; Mattick et al. 2010). Third, ncRNAs are associated with particular chromatin signatures that are indicative of actively transcribed genes (Guttman et al. 2010). Fourth, ncRNAs exhibit tissue- and cell-specific expression patterns as well as subcellular localization. Finally, and very importantly for the dog as a genetic model for human diseases, ncRNAs show modified expression or splicing patterns in cancer and other diseases (Taft et al. 2010).

Identification, classification, and functional analysis of long noncoding RNAs

The characterization and classification of transcribed regions without coding capacity are not trivial. Noncoding RNAs could be classified according to their genomic locations (intergenic, intronic, antisense, promoter of known dog genes), their size (short, e.g., miRNAs, snRNAs, snoRNAs vs. long ncRNAs), their level of conservation, their syntenic environment, and their expression profiles (high, moderate, or low). To discover and catalog all existing ncRNAs in dog, RNA-Seq experiments need to be performed across multiple cell types, physiological and disease states, breeds and individuals, and specific cellular compartments or subcompartments (nucleus, cytosol, chromatin, polysomes), belonging to different classes (short vs. long) or at different states of processing (polyadenylated vs. nonpolyadenylated or capped vs. uncapped) (Kawaji et al. 2009; Encode-Project-Consortium et al. 2011).

Identification of small RNAs, termed microRNAs, led to a new appreciation of the role of RNA in the regulation of gene expression (Lee et al. 1993; Bartel 2004, 2009). Short ncRNAs seem to constitute only one part of the noncoding transcriptome. Indeed, a new class of ncRNAs, the long noncoding RNAs (lncRNAs), is now emerging as a major component of the noncoding transcriptome with important roles within the cell machinery. LncRNAs are generally defined as being greater than 200 nucleotides in length and are distributed on all human chromosomes. Most of them

are lying within introns or in the antisense of protein-coding genes, but recent investigations have identified and characterized intergenic lncRNAs (also termed lincRNAs for large intervening noncoding RNAs) (Guttman et al. 2010; Orom et al. 2010). Similar to protein-coding genes, lncRNAs do contain introns and poly(A)-tails but lack a functional ORF. Until recently, only a few lncRNAs were identified in mammals, such as those lying in the imprinted loci (Air, H19, Kcnq1ot1) (Nagano et al. 2008; Gabory et al. 2009; Mohammad et al. 2010) and X-inactivation loci (Brockdorff et al. 1992; Ciaudo et al. 2006). However, recent investigations using large-scale cDNA sequencing have revealed the existence of thousands of lncRNAs, and chromatin signatures approaches have uncovered thousands of intergenic lncRNA in mouse and human genomes (Khalil et al. 2009; Guttman et al. 2010). The chromatin signatures reveal over 1,000 highly conserved large noncoding RNAs in mammals, many of which appear to regulate gene expression by targeting chromatin remodeling complexes. The study of lncRNAs is of significant relevance to canine and human biology and diseases. This is because they represent a huge and largely unexplored functional component of the genome (Mattick 2009). There is diverse evidence that lncRNAs are intimately involved in gene networks underlying cancers, e.g., an antisense lncRNA that is overexpressed in leukemia and which represses expression of the p15 tumor suppressor (Yu et al. 2008) by lincp21 or by MEG3. Lincp21 is an intergenic lncRNA that functions as a downstream effector of the p53 tumor suppressor (Huarte et al. 2010) and MEG3 activates p53 through a currently unknown mechanism (Zhou et al. 2007). Evidence is also mounting that many neurological diseases involve components of toxic RNA gain-of-function (particularly trinucleotide repeat disorders) (Daughters et al. 2009), or misregulation of coding genes by antisense RNA. There is also genetic evidence that the loss or mutation of lncRNAs can give rise to DiGeorge syndrome (DGCR5) (Faghihi et al. 2008) or myocardial infarction (MIAT) (Ishii et al. 2006). Given the lack of lncRNA annotation in the canine genome, it is likely that many genomic regions associated with diseases in dog or human are due to lncRNAs. To fully exploit the dog as a genetic model for human diseases, it is therefore important to identify the noncoding transcripts of its genome with particular attention paid to lncRNAs.

Conclusions

The sequence assembly of the domestic dog, for which the annotation process identified 20,000 protein-coding genes, will benefit from future efforts to fully classify coding and noncoding genes, pseudogenes, and genomic alterations. Massive parallel sequencing technology has revolutionized the study of genomes and the capacity to accurately annotate genes (Wang et al. 2009). Several studies have demonstrated the power of RNA-Seq to greatly improve the annotation of an already well-studied genome (Wang et al. 2008; Trapnell et al. 2010). RNA-Seq provides unprecedented resolution, allowing us to not only reconstruct and quantify alternatively spliced transcripts but also to accurately monitor the expression output of all genomic loci. Sequence data generated from RNA-Seq reveals alternatively transcribed isoforms active in a particular tissue or a cell population. It can also be used to focus on genomic alterations such as point mutations and structural variations, including gain and loss of chromosomal regions between dog breeds. Finally, RNA-Seq will help to monitor expression changes of both coding and noncoding RNAs that map to or close to genomic regions that vary in copy numbers, and thus help to unravel the effects of copy number variation (CNV) of DNA segments on expression at the gene rather than locus level. This will provide insight into the regulatory mechanisms underlying the phenotypic diversity that exists between breeds and is important not only for dog but also to transpose this knowledge to human species.

Acknowledgments We acknowledge the Centre National de la Recherche Scientifique, the University of Rennes 1 for funding. TD was supported by the Conseil Régional de Bretagne and AV was supported by the European Commission (FP7-LUPA, GA-201370). We thank Jocelyn Plassais for the dog photographs.

References

Abitbol M, Thibaud JL, Olby NJ, Hitte C, Puech JP, Maurer M, Pilot-Storck F et al (2010) A canine Arylsulfatase G (ARSG) mutation leading to a sulfatase deficiency is associated with neuronal ceroid lipofuscinosis. Proc Natl Acad Sci USA 107:14775–14780

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ et al (2010) Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci USA 107:1160–1165

Alekseyev MA, Pevzner PA (2007) Are there rearrangement hotspots in the human genome? PLoS Comput Biol 3:e209

Bannasch D, Young A, Myers J, Truve K, Dickinson P, Gregg J, Davis R et al (2010) Localization of canine brachycephaly using an across breed mapping approach. PLoS One 5:e9632

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233

Beggs AH, Bohm J, Snead E, Kozlowski M, Maurer M, Minor K, Childers MK et al (2010) MTM1 mutation associated with X-linked myotubular myopathy in Labrador Retrievers. Proc Natl Acad Sci USA 107:14697–14702

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995

Blanco E, Parra G, Guigo R (2007) Using geneid to identify genes. Curr Protoc Bioinformatics Chapter 4:Unit 4.3

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST–database for "expressed sequence tags". Nat Genet 4:332–333

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K et al (2010) A simple genetic architecture underlies morphological variation in dogs. PLoS Biol 8:e1000451

Breen M, Hitte C, Lorentzen TD, Thomas R, Cadieu E, Sabacan L, Scott A et al (2004) An integrated 4249 marker FISH/RH map of the canine genome. BMC Genomics 5:65

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S et al (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71:515–526

Cadieu E, Neff M, Quignon P, Walsh K, Chase K, Parker HG, Vonholdt BM et al (2009) Coat variation in the domestic dog is governed by variants in three genes. Science 326(5949):150–153

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Ciaudo C, Bourdet A, Cohen-Tannoudji M, Dietz HC, Rougeulle C, Avner P (2006) Nuclear mRNA degradation pathway(s) are implicated in Xist regulation and X chromosome inactivation. PLoS Genet 2:e94

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP et al (2011) The reality of pervasive transcription. PLoS Biol 9:e1000625

Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle G, Grimmond SM (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics 25:2615–2616

Daughters RS, Tuttle DL, Gao W, Ikeda Y, Moseley ML, Ebner TJ, Swanson MS et al (2009) RNA gain-of-function in spinocerebellar ataxia type 8. PLoS Genet 5:e1000600

Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M et al (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9:R175

Derrien T, Andre C, Galibert F, Hitte C (2007a) Analysis of the unassembled part of the dog genome sequence: chromosomal localization of 115 genes inferred from multispecies comparative genomics. J Hered 98:461–467

Derrien T, Andre C, Galibert F, Hitte C (2007b) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics 23:498–499

Derrien T, Theze J, Vaysse A, Andre C, Ostrander EA, Galibert F, Hitte C (2009) Revisiting the missing protein-coding gene catalog of the domestic dog. BMC Genomics 10:62

Drogemuller C, Karlsson EK, Hytonen MK, Perloski M, Dolf G, Sainio K, Lohi H et al (2008) A mutation in hairless dogs implicates FOXI3 in ectodermal development. Science 321:1462

Enard D, Depaulis F, Roest Crollius H (2010) Human and non-human primate genomes share hotspots of positive selection. PLoS Genet 6:e1000840

Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE et al (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med 14:723–730

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P et al (2011) Ensembl 2011. Nucleic Acids Res 39:D800–D806

Gabory A, Ripoche MA, Le Digarcher A, Watrin F, Ziyyat A, Forne T, Jammes H et al (2009) H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. Development 136:3413–3421

Galibert F, André C (2006) The dog genome. Genome Dyn 2:46–59

Gingeras TR (2007) Origin of phenotypes: genes and transcripts. Genome Res 17:682–690

Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol 2:e133

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510

Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, Quignon P et al (2003) A 1-Mb resolution radiation hybrid map of the canine genome. Proc Natl Acad Sci USA 100:5296–5301

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D et al (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. Nat Rev Genet 6:643–648

Hitte C, Kirkness EF, Ostrander EA, Galibert F (2008) Survey sequencing and radiation hybrid mapping to construct comparative maps. Methods Mol Biol 422:65–77

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM et al (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419

Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310

Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A, Miyamoto Y et al (2006) Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet 51:1087–1099

Jones P, Chase K, Martin A, Davern P, Ostrander EA, Lark KG (2008) Single-nucleotide-polymorphism-based association mapping of dog stereotypes. Genetics 179:1033–1044

Kawaji H, Severin J, Lizio M, Waterhouse A, Katayama S, Irvine KM, Hume DA et al (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. Genome Biol 10:R40

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K et al (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106:11667–11672

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL et al (2003) The dog genome: survey sequencing and comparative analysis. Science 301:1898–1903

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D et al (2006) CAGE: cap analysis of gene expression. Nat Methods 3:211–222

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854

Lequarré AS, Andersson L, André C, Fredholm M, Hitte C, Leeb T, Lohi H et al (2011) LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. Vet J 189:155–159

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5:e1000459

Mattick JS, Taft RJ, Faulkner GJ (2010) A global view of genomic information–moving beyond the gene and the master regulator. Trends Genet 26:21–28

Merveille AC, Davis EE, Becker-Heck A, Legendre M, Amirav I, Bataille G, Belmont J et al (2011) CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. Nat Genet 43:72–78

Meyer IM, Durbin R (2002) Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 18:1309–1318

Mohammad F, Mondal T, Guseva N, Pandey GK, Kanduri C (2010) Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. Development 137:2493–2499

Mosher DS, Quignon P, Bustamante CD, Sutter NB, Mellersh CS, Parker HG, Ostrander EA (2007) A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. PLoS Genet 3:e79

Muffato M, Louis A, Poisnel CE, Roest Crollius H (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26:1119–1121

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S et al (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2:105–111

Olsson M, Meadows JR, Truve K, Rosengren Pielberg G, Puppo F, Mauceli E, Quilez J et al (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 7:e1001332

Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F et al (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58

Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS et al (2004) Genetic structure of the purebred domestic dog. Science 304:1160–1164

Parker HG, Kukekova AV, Akey DT, Goldstein O, Kirkness EF, Baysac KC, Mosher DS et al (2007) Breed relationships facilitate fine-mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. Genome Res 17:1562–1571

Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC et al (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325:995–998

Project Consortium ENCODE, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816

Project Consortium ENCODE, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE et al (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9:e1001046

Semon M, Duret L (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol 23: 1715–1723

Seppala EH, Jokinen TS, Fukata M, Fukata Y, Webster MT, Karlsson EK, Kilpinen SK et al (2011) LGI2 Truncation causes a remitting focal epilepsy in dogs. PLoS Genet 7:e1002194

Sutter NB, Ostrander EA (2004) Dog star rising: the canine genetic system. Nat Rev Genet 5:900–910

Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS (2010) Noncoding RNAs: regulators of disease. J Pathol 220:126–139

Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R (1997) Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 13:263–270

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol 8:e1000371

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T et al (2011) Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet 7(10):e1002316

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wilbe M, Jokinen P, Truve K, Seppala EH, Karlsson EK, Biagi T, Hughes A et al (2010) Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. Nat Genet 42:250–254

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228

Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui H (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 451:202–206

Zhou Y, Zhong Y, Wang Y, Zhang X, Batista DL, Gejman R, Ansell PJ et al (2007) Activation of p53 by MEG3 non-coding RNA. J Biol Chem 282:24731–24742