L'analyse comparée des génomes : applications à l'identification de nouveaux gènes canins

Thomas Derrien

Thèse n°3627

THÈSE

présentée devant

l'UNIVERSITÉ DE RENNES 1

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention Biologie

Par

THOMAS DERRIEN

Équipe d'accueil : Institut de Génétique et Développement - UM6061 - CNRS - Rennes
École doctorale : Vie - Agronomie - Santé
Composante Universitaire : SVE

L'analyse comparée des génomes : applications à l'identification de nouveaux gènes canins

COMPOSITION DU JURY :

Dr. Laurent DURET	Laboratoire "Biométrie et Biologie Évolutive" - UMR CRNS 5558, Villeurbanne	**Rapporteur**
Dr. Gilles FISCHER	Unité de Génétique Moléculaire des Levures - Institut Pasteur, Paris	**Rapporteur**
Dr. Thomas FARAUT	Laboratoire de Génétique Cellulaire INRA UMR444, Castanet Tolosan	**Examinateur**
Pr. Christian DELAMARCHE	Laboratoire Interactions Cellulaire et Moléculaire UMR6026 - CNRS, Rennes 1	**Examinateur**
Pr. Francis GALIBERT	Institut de Génétique et Développement - UMR6061 - CNRS, Rennes 1	**Examinateur**
Dr. Hélène DAUCHEL	UFR Sciences et Techniques - LITIS EA 4108, Université de Rouen	**Examinatrice**
Dr. Catherine ANDRE	Institut de Génétique et Développement - UMR6061 - CNRS, Rennes 1	**Directrice de thèse**
Dr. Christophe HITTE	Institut de Génétique et Développement - UMR6061 - CNRS, Rennes 1	**Encadrant**

Table des matières

Résumé

Au cours de ces trois dernières années, le génome du chien a bénéficié d'avancées majeures à sa connaissance. Les projets de cartographie et de séquençage de son génome, motivés par le formidable potentiel qu'offre le chien en tant que modèle génétique, ont généré de grandes quantités de données à analyser. Dans ce contexte, mes travaux de thèse se sont d'abord focalisés sur la conception d'outils bioinformatiques d'intégration de plusieurs ressources afin d'évaluer et de comparer les informations issues des projets de cartographie et de séquence du génome du chien.
Avec la disponibilité d'un nombre croissant de génomes séquencés, nous avons développé le programme AutoGRAPH pour formaliser la conservation de l'ordre des gènes orthologues entre les génomes mammifères, automatiser la construction de cartes de synténie entre ces génomes et, enfin, faciliter l'annotation du génome du chien.
Un première application de notre méthode a permis de redéfinir la localisation d'une centaine de gènes préalablement assignés au chromosome canin non-assemblé ou "chromosome Unknown".
Dans un second projet, nous avons combiné notre approche de conservation de l'ordre des gènes entre deux génomes avec des alignements de séquences ciblés afin d'identifier des nouvelles structures de gènes canins codant pour des protéines. À partir d'un ensemble de 412 gènes orthologues entre quatre génomes de référence (homme - chimpanzé - rat - souris) et présumés absents chez le chien, nous identifions 285 nouveaux gènes canins et/ou nouvelles relations d'orthologie avec les génomes de référence. Enfin, différents mécanismes évolutifs sont suggérés mettant en relation la nature des gènes, la présence de famille de gènes et la composition en séquences pour expliquer la perte de gènes chez le chien.

Abstract

Dog is an important model for genetic studies and has benefited from extensive genomic resources for its knowledge for three years. Recent genome mapping and sequencing projects have produced a huge amount of data that needed to be analysed. In this context, my PhD works firstly focused on the design of integrated bioinformatic tools in order to assess and compare mapping and sequencing resources. With the availability of many genome sequences, we extended the use of the program

AutoGRAPH to formalise gene order conservation between mammals genomes, to automate synteny maps construction and finally, to facilitate dog genome annotation.

We first applied our method to refine more than a hundred of gene location annotated on the unassembled part of the dog genome also called “Chromosome Unknown”. Then, we combined our gene order conservation approach with targetted sequence alignments on the dog genome to identify new canine protein coding genes. From a subset of 412 ortholog genes between four reference species (human - chimpanze - mouse - rat) and not annotated in the dog genome, our method showed that 285 new orthology relationships and/or new dog genes could be determined. Finally, we provide evidences for dog gene loss scenarios for more than 90 genes.

Table des Publications

Par ordre de présentation dans ce mémoire

‣ **Publication n°1 :**

AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps.
Derrien T, Andre C, Galibert F, Hitte C.
Bioinformatics. 2007 Feb 15;23(4):498-9. PMID: 17145741

‣ **Publication n°2 :**

Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping.
Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C, Evanno G, Pullar B, Cadieu E, Oza V, Lourgant K, Jaffe DB, Tacher S, Dreano S, Berkova N, Andre C, Deloukas P, Fraser C, Lindblad-Toh K, Ostrander EA, Galibert F.
Nat Rev Genet. 2005 Aug;6(8):643-8. PMID: 16012527

‣ **Publication n°3 :**

Analysis of the unassembled part of the dog genome sequence: chromosomal localization of 115 genes inferred from multispecies comparative genomics.
Derrien T, Andre C, Galibert F, Hitte C.
J Hered. 2007;98(5):461-7. PMID: 17573383

‣ **Publication n°4 :**

Refining the Dog Gene Repertoire using Multispecies Consensus Orthologous IntervaLs
Derrien *et al.*, en préparation

‣ **Publication n°5 :**

Coat colour in dogs: identification of the merle locus in the Australian shepherd breed.
Hedan B, Corre S, Hitte C, Dreano S, Vilboux T, Derrien T, Denis B, Galibert F, Galibert MD, Andre C.
BMC Vet Res. 2006 Feb 27;2:9. PMID: 16504149

* Publication n°6 :

A comparative genome approach to marker ordering.
Faraut T, de Givry S, Chabrier P, Derrien T, Galibert F, Hitte C, Schiex T.
Bioinformatics. 2007 Jan 15;23(2):e50-6. PMID: 17237105

* Publication n°7 :

Building comparative maps using 1.5x sequence coverage: human chromosome 1p and the canine genome.
Guyon R, Kirkness EF, Lorentzen TD, Hitte C, Comstock KE, Quignon P, Derrien T, Andre C, Fraser CM, Galibert F, Ostrander EA.
Cold Spring Harb Symp Quant Biol. 2003;68:171-7. PMID: 15338615

* Publication n°8 :

CRH_Server: an online comparative and radiation hybrid mapping server for the canine genome.
Hitte C, Derrien T, Andre C, Ostrander EA, Galibert F.
Bioinformatics. 2004 Dec 12;20(18):3665-7. PMID: 15256409

* *Publications non présentées dans ce manuscrit.*

Table des illustrations

Liste des figures

Liste des tableaux

Abréviations

ADN	: Acide Désoxyribonucléique
ADNc	: Acide Désoxyribonucléique complémentaire
APR	: Atrophie Progressive de la Rétine
ARN	: Acide : Acide Ribonucléique
ARNr	: Acide Ribonucléique ribosomal
BAC	: Bacterial Artificial Chromosome
BBH	: Bidirectional Best Hit
BGC	: Biased Gene Conversion (Conversion Génique Biaisée)
BLAST	: Basic Local Alignment Search Tool
BUS	: Best Unambigous Subset
CFA	: Canis Familiaris (chromosome)
CMH	: Complexe Majeur d'Histocompatibilité
COG	: Cluster of Orthologous Groups (Groupe de Cluster Orthologues)
COIL	: Consensus Ortholog IntervaL
CPU	: Central Processing Unit
CS	: Conserved Segment (Segments Conservé)
CSO	: Conserved Segment Ordered (Segments Conservé Ordonné)
ENCODE	: ENCyclopedia Of Dna Element
EST	: Expressed Sequence Tag
FISH	: Fluorescent In Situ Hybridisation
Gb, Mb, pb	: Gigabases, Megabases, paire de bases
GFF	: General File Format
GO	: Gene Ontology
GOLD	: Genome Online DataBase
GRIMM	: Genome Rearrangement In Man and Mouse
HAR	: Human Accelerated Region
LIS	: Longest Increasing Subsequence
LTR	: Long Terminal Repeat
MA	: Millions d'Années
MITF	: Microphtalmia associated Transcription Factor gene
MMC	: Modèle de Markov Caché
OR	: Olfactory Receptor
PAML	: Phylogenetic Annalysis by Maximum Likelihood
PCR	: Polymerase Chain Reaction
RBH	: Reciprocal Best Hit
RBM	: Random Breakage Model
RH	: Radiation Hybrid (Hybride d'irradiation)
RSD	: Reciprocal Smallest Distance (Plus petite distance réciproque)
RT-PCR	: Reverse Transcription Polymerase Chain Reaction
SINE	: Short Interspersed Nuclear Element
YGOB	: Yeast Gene Order Browser

Introduction générale

Le 25 avril 1953, la revue “Nature” publiait trois articles qui allaient révolutionner la conception de la molécule d’ADN (Acide désoxyribonucléique) et engendrer l’avènement de l’ère de la biologie moléculaire.

Le premier article de James Watson et Francis Crick modélisait la structure de l’ADN en une double hélice où les quatre bases (Adénine, Thymine, Guanine et Cytosine) étaient liées par des liaisons hydrogènes (Watson *et al.*, 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, par ailleurs, 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 *et al.*, 1953). En 1962, quatre ans après la mort de Rosalind Franklin, Watson, Crick et Wilkins se partagèrent le prix Nobel de médecine “pour leurs découvertes concernant les structures moléculaires des acides nucléiques et leur importance pour le transfert des informations dans les organismes vivants” (http://nobelprize.org/nobel_prizes/medicine/laureates/1962/).

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 developpé par L. Hood dès 1987) et dans l’automatisation des procédures expérimentales (technique du shot-gun, production d’étiquettes de séquences transcrites ou ESTs) vont aboutir au déchiffrage des séquences d’ADN de nombreux organismes dont celui de l’homme en 2003. La communauté scientifique internationale commémore alors le cinquantième anniversaire de la découverte de la double hélice en pleine “ère de la génomique”. En effet, 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 exponentielle. 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 banques de données publiques telles que GenBank (1982). En octobre 2007, plus de 2 500 projets de séquençage (complets ou partiels) sont recensés et près de 700 organismes appartenant aux archébactéries, 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 enchaînement des quatre nucléotides A, T, G et C), celle-ci ne renseigne pas sur les zones fonctionnelles qui la composent (gènes, promoteurs, séquences régulatrices) (Hardison *et al.*, 2003). Cette tâche s'avère particulièrement complexe pour les génomes eucaryotes où la proportion de zones codantes ne représente qu'une faible part de la totalité du génome (<2% du génome humain). Pour interpréter l'information contenue dans une séquence génomique, une approche couramment utilisée est de la comparer aux séquences d'autres génomes. Ces études de génomique comparative reposent sur le principe selon lequel toute séquence d'ADN est soumise à la sélection naturelle. Ainsi, des caractères communs entre deux organismes seront souvent codés par des séquences génomiques conservées entre ces deux espèces (Miller *et al.*, 2004 ; Hardison *et al.*, 2003). À l'inverse, les séquences qui codent pour des fonctions spécifiques à chacune des espèces auront des profils de séquences divergents. Ce processus dit d'annotation des génomes en séquences fonctionnelles, est indispensable à la compréhension des phénomènes biologiques au niveau moléculaire et cellulaire, et a bénéficié largement de l'automatisation des programmes informatiques pour autoriser la comparaison de larges séquences d'ADN et maintenant de génomes entiers.

Au-delà de la comparaison de séquences au niveau très résolutif du nucléotide, la génomique comparative permet l'étude de l'évolution et de la dynamique des génomes. En effet, l'analyse comparée de chromosomes de multiples espèces renseigne sur les segments conservés entre les espèces au cours de l'évolution, l'ordre préservé des séquences fonctionnelles au sein de ces segments et enfin les mécanismes de réarrangements chromosomiques intra ou inter espèces. Ces approches comparatives, initiées dans les années 90 par des études de cartographies comparées, peuvent être affinées et étendues maintenant à un nombre toujours plus grand d'espèces séquencées telles que le chien en décembre 2005.

Le formidable potentiel en génétique qu'offre le modèle 'chien' a amené notre laboratoire à développer des outils de connaissance de son génome. En effet, à la très grande variation phénotypique entre les races de chiens s'oppose une forte homogénéité au sein de chaque race. Ainsi, les races canines peuvent être considérées comme de véritables isolats génétiques à l'instar des populations humaines isolées (Galibert *et al.,* 2004). Cette structuration particulière de la population canine, liée aux pratiques de sélection intensive menées par l'homme, a entraîné une

forte prévalence des maladies génétiques au sein de chaque race. Chez le chien, la déclaration des maladies génétiques est spontanée et la plupart sont homologues aux maladies génétiques humaines. L'utilisation du chien en tant que modèle pour des applications en médecine humaine et vétérinaire repose donc sur une connaissance approfondie de la structure de son génome et des régions fonctionnelles qui le composent. Notre laboratoire a initié en 1995 un travail de cartographie du génome canin par la méthode des hybrides d'irradiation. La dernière carte (Hitte *et al.*, 2005) contient plus de 10 000 marqueurs de gènes balisant le génome du chien. En 2005, le génome du chien devient le cinquième mammifère à bénéficier d'un séquençage complet après l'homme, la souris, le chimpanzé et le rat.

Mon travail de thèse s'est inscrit à un moment important de la connaissance du génome canin, avec la disponibilité d'une cartographie fine et du séquençage initial de son génome.
Les objectifs de mon projet étaient donc, en premier lieu, de concevoir des outils bioinformatiques capables de formaliser la comparaison et l'intégration des deux ressources que sont la cartographie et la séquence. La deuxième partie de mon travail de thèse a consisté à développer ces outils pour la comparaison inter-espèces et à les appliquer à l'annotation du génome canin. J'ai formalisé une approche qui intègre l'analyse comparée de l'ordre des gènes entre plusieurs espèces et la prédiction de structure de gènes par analyses de séquences. L'application de notre approche aux gènes présumés absents chez le chien, mais annotés chez les autres génomes mammifères, a permis, en premier lieu, de déterminer de nouveaux gènes canins puis, d'identifier un ensemble de gènes absents chez le chien pour lesquels nous apportons des arguments impliquant des scénarios de perte de gènes. Une partie de mes travaux a été utilisée dans le cadre de collaborations où l'apport de la génomique comparative s'est avérée complémentaire aux problématiques traitées.
L'ensemble de mon travail de thèse présenté dans ce manuscrit décrit les connaissances actuelles sur la comparaison des génomes et le contexte scientifique de ce projet dans une introduction générale. Les résultats obtenus sont présentés sous forme de publications avec au préalable une description détaillée. Une partie discussion sur l'ensemble des travaux expose les limites, les conclusions et les perspectives de ce projet de thèse.

I. INTRODUCTION

L'identification des régions fonctionnelles codantes et non-codantes au sein des génomes, représente un problème complexe qui a connu des avancées remarquables durant les dix dernières années. En effet, moins de 2% du génome humain, code pour des protéines ce qui a amené à comparer notre génome "à un désert parsemé ici et là de rares oasis génétiques" (Watson, 2003). Alors qu'à la fin des années 1990, la plupart des biologistes prédisaient l'existence d'environ 100 000 gènes chez l'homme, l'achèvement du séquençage du premier chromosome humain (le chromosome 22) fut le premier acte d'une révision à la baisse du catalogue génique de l'homme (Watson, 2003). En effet, le chromosome 22 qui représente ~1,1% du génome humain contient moins de 600 gènes, loin du millier de gènes attendus. De plus, la majorité des séquences non-codantes étaient souvent considérées comme non fonctionnelles et régulièrement référencées sous le terme "d'ADN poubelle". Dans les années 1990, le dogme "un gène, une protéine" prévalait. Or, ces dernières années ont éclairé l'étude des génomes par la complexité des processus moléculaires et l'importance des séquences fonctionnelles non-codantes.

Parallèlement, les avancées technologiques en matière de séquençage des génomes et leurs applications systématiques à de nombreux organismes (virus, bactéries, levures, vertébrés) donnent l'occasion unique de pouvoir répondre à des questions fondamentales en biologie : Combien de gènes possèdent les génomes des métazoaires et comment sont-ils organisés ? Existent-t-il des gènes spécifiques d'espèce, combien sont communs à la plupart des organismes ? Quelles interactions régissent le bon fonctionnement des organismes ?

1. La génomique comparative dans le contexte de la séquence

1.1. Homologie - Orthologie - Paralogie

La comparaison des séquences fonctionnelles entre espèces nécessite de définir trois notions importantes en biologie : homologie, orthologie et paralogie (Fitch *et al.*, 1970 ; Petsko *et al.*, 2003 ; Varshavsky *et al.,* 2004 ; Koonin *et al.,* 2005). Historiquement, le terme "d'homologue" a été utilisé pour la première fois en anatomie pour désigner le même organe chez différents animaux (Owen, 1848). Plus tard, le terme d'homologie a été utilisé pour désigner une origine ancestrale commune (Huxley, 1860). Dans les années 1970, Walter Fitch (Fitch *et al.*, 1970) précisa la notion d'homologie en introduisant (i) 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 (ii) de gènes paralogues faisant référence à des gènes homologues mais séparés par un événement de duplication.

Plus récemment, Sonnhammer et Koonin (Sonnhammer *et al.*, 2002) proposèrent d'affiner le terme de paralogue en fonction de l'ordre des scénarios de duplication et de spéciation. En effet, la duplication d'un gène peut précéder ou suivre un événement de spéciation. Ainsi, le terme 'inparalogue" désignera des paralogues qui ont évolué après un événement de spéciation (la duplication est spécifique de la lignée) alors que le terme de "outparalogue" concernera les gènes appartenant à deux espèces différentes puisque la duplication du gène aura précédé l'événement de spéciation (**Fig.1**).

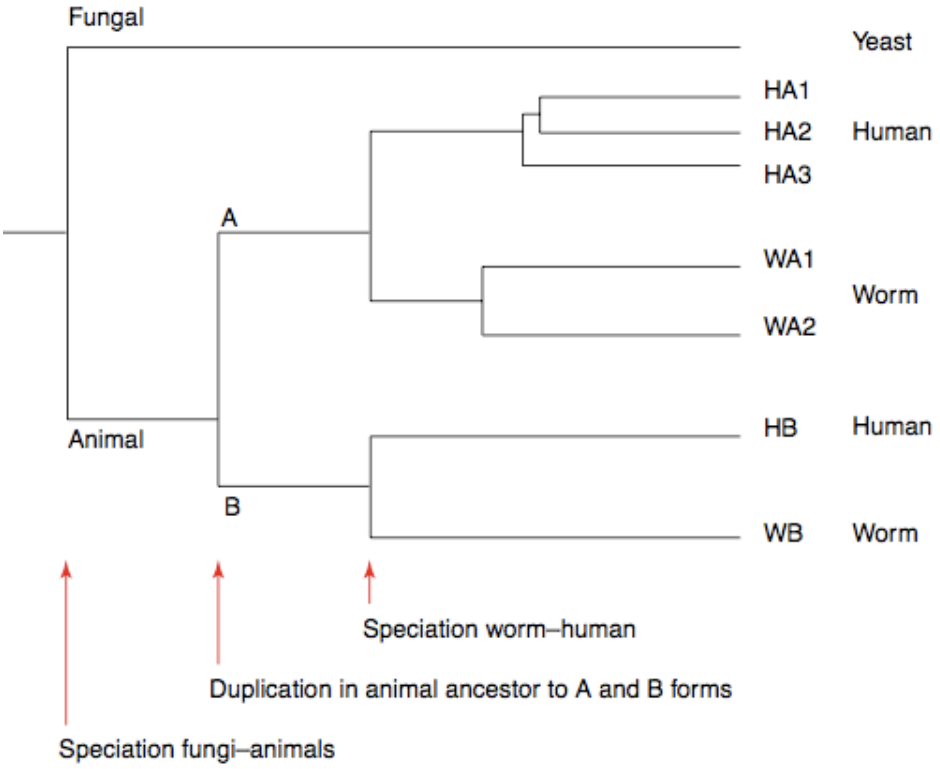


Figure 1 : Distinction entre inparalogues et outparalogues. (Sonnhammer *et al.*, 2002)
Les gènes humains HA* sont tous inparalogues puisque la duplication du HA qui leur a donné naissance est postérieure à la spéciation entre le vers et l'homme (3ème flèche en partant de la gauche). À l'inverse, les gènes HB et HA* sont outparalogues car la duplication du gène H (2ème flèche en partant de la gauche) a précédé l'événement de spéciation.

La recherche de gènes communs entre deux espèces consiste donc à identifier les correspondances de gènes entre deux génomes, ce qui implique de différencier les gènes orthologues des gènes paralogues. Classiquement, on représentera deux gènes orthologues chez

deux espèces par une relation d'orthologie de type un:un ou 1:1. Un patron d'orthologie limité au type 1:1 supposerait une correspondance parfaite du répertoire des gènes entre deux espèces, soit une vision simpliste et incomplète de l'annotation des génomes. Par extension, plusieurs types d'homologie sont à considérer. La relation d'orthologie 1:0 signifie qu'un gène est présent dans une espèce et présumé absent chez une autre (on parle aussi de "gène orphelin d'annotation") (Domazet-Loso *et al.*, 2003). Pour les patrons d'homologie plus complexes de type 1:n, (où $n \geq 2$), on évoque des mécanismes de duplication(s) de gènes dans un génome pour expliquer l'origine de ce type de correspondances de gènes. Enfin, les relations d'homologie particulièrement difficiles à identifier sont représentées par un patron de type m:n qui implique souvent de vastes familles de gènes (ex. la famille des immunoglobulines ou des récepteurs olfactifs)

De par la quantité d'informations à traiter (un génome de mammifère représente en moyenne 2,5 à 3 Gb), l'identification de séquences homologues entre deux génomes ne peut s'effectuer que par l'intermédiaire d'outils bioinformatiques d'alignements de séquences nucléotidiques/protéiques (tels que BLAST) (Altschul *et al.*, 1990). Cette tâche s'avère d'autant plus complexe que certains événements évolutifs entraînent l'apparition ou la disparition de gènes (Cf. chapitre sur les duplications). Le développement constant des ressources matérielles et des méthodes bioinformatiques dédiées aux analyses d'homologie et à l'annotation des génomes a rendu possible l'analyse des données générées à l'échelle du génome. Nous avons recensé un ensemble des principales méthodologies et des améliorations apportées.

1.2. Les méthodes d'identification des relations d'homologie entre génomes

1.2.1. L'alignement réciproque de séquences

Avant l'existence des programmes de séquençage à haut débit, les analyses comparatives dédiées à la recherche d'homologues entre deux espèces se sont largement inspirées des travaux de Fitch (Fitch *et al.*, 1970 ; Fitch *et al.*, 1995). La méthode la plus généralement utilisée consiste à réaliser des alignements de séquences réciproques entre deux génomes (méthode du "Best Bidirectional Hits" (BBH) ou "meilleurs alignements réciproques"). La popularité de cette approche peut d'ailleurs être mesurée par le nombre de termes qui la désignent. On notera par exemple, les termes de Mutual Hits, SymBests (Symetrical Best hits) (Koonin *et al.,* 2005) ou encore RBH (Reciprocal Best Hits) (Hubbard *et al.*, 2007 ; Potter *et al.* 2004). L'idée sous-jacente de la méthode implique que les séquences de gènes orthologues présentent plus de similitudes entre elles que de similitudes

avec une séquence du génome initial. Ainsi, deux gènes A et B, appartenant respectivement aux espèces S_A et S_B, seront considérés orthologues si le meilleur alignement de la séquence du gène A sur le génome S_B correspond au gène B et, réciproquement, si le meilleur alignement de la séquence du gène B sur le génome S_A correspond au gène A (**Fig.2**). Cependant, dans le cas de duplications d'un des gènes (fréquents chez les eucaryotes) ou de transferts horizontaux de gènes (fréquents chez les procaryotes), cette approche symétrique ne permettra pas, par principe, de distinguer la présence additionnelle d'homologues dans le génome. Elle reste limitée à la mise en évidence ou bien à l'absence d'évidence d'une partie des relations d'orthologie (les relations 1:1 ou 1:0). La notion de réciprocité des alignements de séquences a largement servi de support au développement de méthodes plus précises pour la recherche d'homologies.

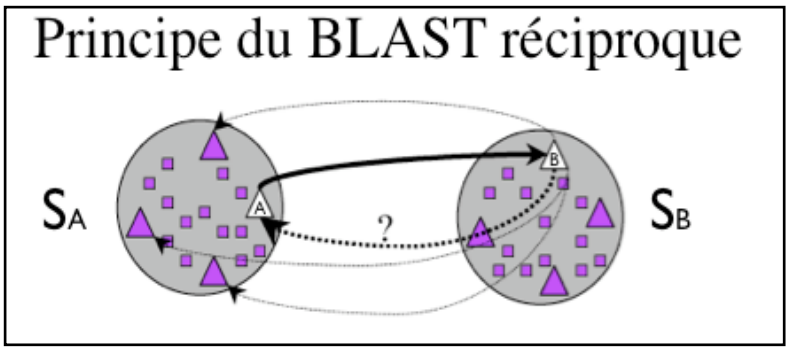


Figure 2 : Le principe du BLAST réciproque entre deux génomes de deux espèces S_A et S_B.
Les gènes orthologues réciproques identifiés par la méthode sont symbolisés par les triangles blancs.

1.2.2. Les méthodes de regroupement

1.2.2.1. Les groupes de gènes orthologues

La méthode des "COGs" pour Cluster of Orthologous Groups" (Tatusov *et al.,* 2001) consiste à identifier, non seulement des paires de gènes orthologues, mais plus généralement, des groupes de gènes orthologues entre plusieurs génomes. Brièvement, toutes les séquences de gènes sont alignées entre les génomes considérés afin d'identifier l'ensemble des relations d'homologie. Puis, la détection des paralogues est réalisée par des alignements de la séquence d'un gène sur son propre génome. Enfin, la dernière étape consiste à rassembler les groupes de gènes (orthologues et paralogues) identifiés dans les deux premières étapes pour former des groupes de gènes homologues. Bien que cette méthode ait permis d'étendre la notion de réciprocité à l'analyse de l'ensemble des relations d'homologie (relations n:m) entre plusieurs génomes, il a été démontré un regroupement artificiel de certains COGs pour des grandes familles de gènes telles que les familles

codant pour les récepteurs olfactifs ou les facteurs de transcription. En effet, dans ces cas particuliers, la grande similitude de séquences entre les gènes de la même famille et la complexité des combinaisons possibles duplication-spéciation entraînent une association artificielle de certains COGs. De plus, la présence de certains gènes codant pour des protéines possédant de multiples domaines fonctionnels entraîne un regroupement biaisé de COGs.

1.2.2.2. La détection des in-paralogues (Inparanoid)

La première étape de la méthode Inparanoid (Remm *et al.*, 2001 ; O'Brien *et al.* 2005) cherche à définir la paire de gènes orthologues (notée A_1 - B_1) correspondant aux meilleurs alignements réciproques de séquences. Une valeur de confiance est assignée à cette paire par une méthode de bootstrap. Cette valeur correspond au nombre de fois où la paire d'orthologue est considérée comme étant la paire la plus fiable après avoir réalisé une succession de ré-échantillonnages de l'ensemble des autres paires d'orthologues possibles. À partir de cette paire initiale identifiée, le regroupement indépendant des paralogues de chacune des espèces avec la paire d'orthologue initiale est réalisé. Le principe de cette méthode repose sur l'hypothèse selon laquelle, après la spéciation, un gène dupliqué (in-paralogue) présentera plus de similitudes avec ses paralogues qu'avec tout autre orthologue d'une autre espèce. Ainsi, si la distance (ou l'inverse du score de l'alignement) entre les gènes orthologues de la paire initiale (A_1 - B_1) est inférieure à la distance entre les deux gènes paralogues du génome initial (A_1 et A_2), alors le paralogue A_2 est considéré

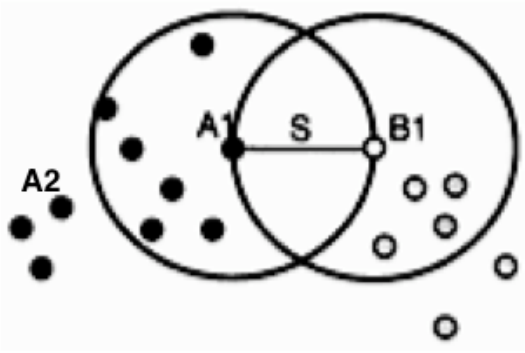


Figure 3 : Principe de l'identification des inparalogues par Inparanoid.
Les points noirs dans le cercle délimité par le rayon de taille S (Score) sont inparalogues du gène A1. À l'inverse, le gène A2 est outparalogue.

comme outparalogue (**Fig.3**). Pour résumer, un groupement d'orthologues par Inparanoid contient

une paire initiale d'orthologues avec une valeur de bootstrap à laquelle est associée une liste d'inparalogue(s) avec un score de confiance. Toutefois, cette méthode ne s'applique qu'à la comparaison de deux génomes et nécessite de connaître, *a priori*, tous les gènes présents chez les deux espèces. En effet, si un gène est absent en raison d'un problème d'annotation, ou d'un défaut de séquence (assemblage), cette approche ne permettra probablement pas d'identifier correctement la relation d'orthologie.

1.2.3. Les méthodes phylogénétiques

1.2.3.1. La plus petite distance réciproque (RSD : Reciprocal Smallest Distance)

Pour distinguer les gènes orthologues entre deux génomes et conserver la notion de réciprocité, Wall *et al.* (Wall *et al.*, 2003) ont développé une méthode qui introduit le calcul d'une distance évolutive entre deux gènes. En pratique, la séquence d'un gène du génome A_1 est alignée sur toutes les séquences des gènes du génome B puis les séquences $B_{(i->j)}$ des meilleurs alignements (dépassant un seuil pré-défini) sont conservées. Une estimation par maximum de vraisemblance du nombre de substitutions séparant chaque paire de séquences protéiques A_1 - $B_{(i->j)}$ est calculée grâce à des programmes dédiés au calcul de taux d'évolution par maximum de vraisemblance tel que le package PAML (Phylogenetic Analysis by Maximun Likelihood) (Yang *et al.*, 1997 ; Yang *et al.*, 2007). Après avoir effectué la même étape des gènes $B_{(i->j)}$ vers A, la paire de séquences présentant la plus petite distance évolutive est alors considérée comme étant orthologue. Même si cette approche réduit le risque d'exclure un orthologue plutôt qu'un paralogue et inclut, par ailleurs, la notion de vitesse évolutive des gènes, elle reste néanmoins toujours limitée à une approche réciproque qui ne permettra pas de mettre en évidence les différents scénarios évolutifs possibles tels que les duplications de gènes.

1.2.3.2. La réconciliation d'arbre phylogénétique

Les méthodes phylogénétiques, construisant l'histoire évolutive d'une famille de gènes, semblent les mieux adaptées pour distinguer les relations d'orthologie/paralogie. Après un alignement multiple des séquences d'une famille de gènes homologues, l'arbre phylogénétique est reconstruit par des méthodes de parcimonie ou de maximum de vraisemblance. La topologie de l'arbre des gènes construit est alors comparée à celle de l'arbre des espèces. Si des incongruités existent entre

les deux arbres, des algorithmes de réconciliation d'arbre ont été développés notamment par Page et Charleston et Dufayard *et al.* (Page *et al.*, 1997 ; Dufayard *et al.*, 2005). L'outil méthodologique développé par Dufayard *et al.* (RAP) intègre à la fois la comparaison de la topologie des arbres mais aussi une analyse de la longueur des branches. En effet, la présence de longues branches au sein des arbres des gènes peut refléter une évolution plus rapide d'un gène ou la présence d'une paralogie sous-jacente. La limite fondamentale de cette approche concerne, dans le cas des procaryotes, l'incertitude de l'arbre des espèces liée à la prépondérance de transferts horizontaux de gènes (chez les eucaryotes, peu d'exemples de transferts horizontaux ont été démontrés). Chez les eucaryotes, cette approche requiert de définir, au préalable, des familles de gènes à analyser car les études sur des génomes entiers sont coûteuses en terme de calcul informatique. De plus, elles sont plus sensibles aux artefacts lors de la reconstruction d'arbre et demeurent dépendantes de la bonne connaissance de l'arbre des espèces.

1.2.3.3. Reconstruction phylogénétique des orthologies et paralogies entre deux génomes

La "phylogénomique" ou l'étude phylogénétique de génomes entiers, présente un double challenge dans la sensibilité d'identification des relations d'homologie entre deux génomes et dans l'amélioration des vitesses de calcul de ces analyses. Très récemment, Goodstadt L. *et al.* ont développé une nouvelle procédure (PhyOP : PHYlogenetic reconstruction of Orthology and Paralogy) basée sur la reconstruction d'arbres phylogénétiques à partir de toutes les séquences de transcrits de deux génomes (Goodstadt *et al.*, 2006). PhyOP a l'avantage de combiner à la fois la reconstruction d'arbres phylogénétiques basée sur l'analyse de tous les transcrits de deux génomes ainsi que le calcul d'une distance évolutive affinée entre les gènes d'espèces différentes. Contrairement au méthodes traditionnelles de détection d'orthologues fondées sur la mesure du taux de substitution en acides aminés, PhyOP utilise le rapport du nombre de substitutions synonymes nucléotidiques sur le nombre de sites synonymes (d_S). En effet, au sein des séquences codantes, les sites synonymes sont, *a priori*, moins contraints que les autres sites codants du fait de la dégénérescence du code génétique. Ils reflètent donc plus efficacement le taux d'évolution neutre des séquences codantes et donc la distance évolutive "vraie" entre deux gènes. On notera, néanmoins, que pour des génomes très distants (> 300 MA), la saturation au niveau des sites synonymes entraînera une saturation du d_S. La précision de la méthode sera donc d'autant plus petite que la distance entre les deux génomes comparés est grande.

1.2.3.4. L'approche "d'ombrage plylogénétique" ou "phylogenetic shadowing"

Si la comparaison de séquences génomiques entre espèces distantes s'est avérée très efficace pour identifier les gènes et les relations d'homologie entre deux génomes, cette stratégie n'est pas aussi pertinente lorsque l'on considère deux génomes très proches au niveau phylogénétique. En effet, l'alignement de séquence peut se révéler peu informatif étant donné la forte similarité de séquence entre deux espèces ayant divergé récemment (Boffelli *et al.*, 2003). Par contre, la considération simultanée de plusieurs espèces proches va permettre d'accumuler les variations, et c'est l'addition des divergences entre de multiples espèces qui va faciliter l'identification de régions conservées par rapport aux régions non-conservées (**Fig.4**). La prise en compte des variations collectives et complémentaires entre espèces proches (comme les primates) va reproduire la situation

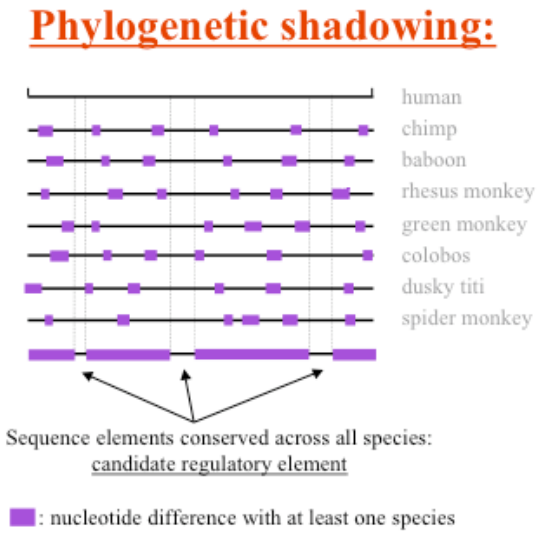


Figure 4 : Illustration de "l'ombrage phylogénétique" entre huit primates.
Mise en évidence des régions variables entre les organismes comparés en violet et déduction des zones très conservées avec l'homme (Boffelli *et al.*, 2003).

phylogénétique d'espèces distantes type homme-souris. C'est sous le nom de "phylogenetic shadowing" que cette approche a été popularisée dès 2003.

4. L'apport de la synténie à l'identification des relations d'orthologie

En 1971, J.H. Renwick introduit le terme de synténie, étymologiquement "sur le même brin", pour désigner des loci localisés sur le même chromosome. En génomique comparative, ce terme décrit des régions chromosomiques conservées entre plusieurs espèces. Pour des espèces relativement proches au niveau phylogénétique, l'organisation spatiale des gènes au sein des ces régions est

souvent conservée en blocs au cours de l'évolution. Si l'ordre comparé des gènes entre deux génomes est linéaire (colinéarité), on parle alors de segments conservés ordonnés (CSO pour Conserved Segment Ordered) (O'Brien *et al.*, 1999) ou segments synténiques (Waterston *et al.*, 2002 ; Gibbs *et al.*, 2004 ; Lindblad-Toh *et al.*, 2005). Plus généralement, des régions contiguës entre deux génomes, où l'ordre des gènes peut être altéré, seront définies par le terme de segments conservés (CS pour Conserved Segment) ou blocs de synténie. Des réarrangements chromosomiques peuvent survenir au cours de l'évolution aboutissant à des altérations de l'ordre des gènes entre deux génomes appelées ruptures de synténie. Il est intéressant de remarquer que la définition de la synténie a été déviée de son contexte initial (Passarge *et al.*, 1999), et certains auteurs préfèrent utiliser la notion de segments orthologues.

Pour la plupart des méthodes d'identification d'homologues présentées précédemment, l'apport de l'information de synténie reste limité à une étape de confirmation *a posteriori* des prédictions des relations d'orthologie entre les génomes. Si l'on prend l'exemple d'un gène A_1 d'une espèce A ayant pour orthologue un gène B_j d'une espèce B, la relation d'orthologie A_1 - B_j sera d'autant plus vraisemblable que les deux gènes sont localisés dans un segment conservé ordonné (exempt de tout

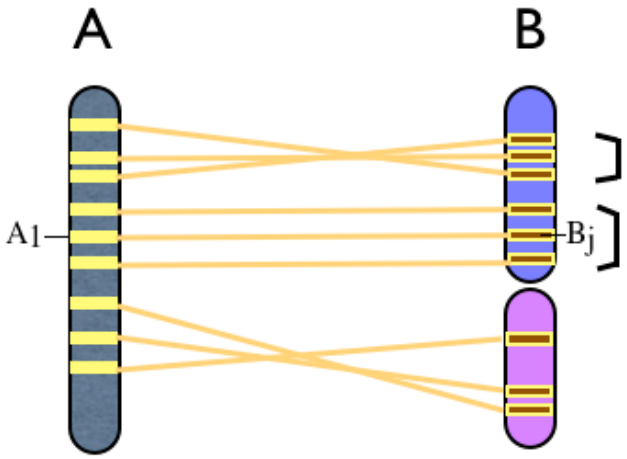


Figure 5 : Représentation de la conservation de synténie entre deux génomes A et B.
Les gènes du chromosome A sont représentés en jaune et les relations d'orthologie avec le génome B sont symbolisées par les lignes horizontales oranges. Le chromosome A possède deux segments conservés avec le génome B (en bleu et mauve) et, au sein du premier segment conservé, il existe deux segments conservés ordonnés symbolisés par les intervalles en noirs. Les gènes orthologues A_1 et B_j appartiennent au deuxième segment conservé ordonné.

remaniement chromosomique depuis la spéciation) (**Fig.5**). Ainsi, PhyOP pour les méthodes phylogénétiques et BUS (Best Unambiguous Subset - méthode similaire à Inparanoid (Kellis *et al.*, 2004)) pour les méthodes de regroupement intègrent la notion de synténie des gènes afin de valider

leurs prédictions d'orthologie. Les duplications en tandem sont très fréquentes chez les eucaryotes et sont parfois à l'origine de la création de grandes familles multigéniques pouvant être groupées en "cluster". Ce type de duplication, entraîne une conservation de l'ordre des gènes et favorise ainsi l'identification des paralogues par des approches qui considèrent la comparaison des profils synténiques comme une étape déterminante (Cf. chapitre sur les duplications). À l'inverse, l'absence d'argument de synténie lors de l'identification d'orthologue peut suggérer une identification douteuse qui correspondrait alors à une forme non-fonctionnelle du gène issu d'un mécanisme de rétrotransposition par exemple.

En 2004, Zheng *et al*. ont augmenté de 20% le nombre de relations d'orthologies identifiées entre l'homme et la souris par une approche basée à la fois sur un alignement réciproque des gènes et sur les synténies conservées entre les deux génomes (Zheng *et al*., 2004). L'alignement des génomes complets humain et murin permet de déterminer de courtes séquences très conservées (exons, séquences régulatrices) utilisées comme ancres de synténie. Puis, la conservation de l'ordre de ces ancres le long des chromosomes de l'homme et de la souris sert de support à la définition de blocs de synténie. À partir des meilleurs alignements réciproques de transcrits homme-souris, un premier ensemble d'orthologues est alors identifié. Une autre étape permettra de considérer de nouvelles paires d'orthologues. En effet, deux gènes chez l'homme et la souris seront considérés orthologues si ces gènes (i) présentent de grandes similitudes de séquences (compris dans les cinq meilleurs alignements réciproques), (ii) partagent des ancres de synténie au sein de leurs exons (une portion ou la totalité des exons sont synténiques) et (iii) sont inclus dans le même bloc de synténie défini dans la première étape. Finalement, plus de 95% des gènes orthologues définis par leur approche partagent le même bloc de synténie, soulignant l'importance de la conservation de patrons d'organisation des génomes pour améliorer la qualité des prédictions. Les limites de cette approche dépendent du choix dans la proximité phylogénétique des génomes à comparer. Si le génome de référence est trop proche du génome testé (ex. Homme-chimpanzé), un grand nombre de pseudogènes sera ancestral et donc non détecté par l'analyse de synténie car la conservation d'ordre sera toujours respectée. Ici, le recours aux méthodes d'alignement de séquences qui analysent les régions introniques seront plus efficaces pour la détection de pseudogènes, car la structure exon-intron ne sera pas retrouvée. Dans la situation où deux génomes sont "très" distants (mammifère-poisson), de véritables gènes orthologues ne seront plus en conservation de synténie car séparés par les nombreux réarrangements chromosomiques ayant eu lieu depuis la séparation de leur ancêtre commun.

1.3. Les ressources disponibles pour les recherches d'homologies

1.3.1. Les bases de données associées aux méthodes de recherche d'homologies

Chacune des approches précédentes a été appliquée à un nombre variable de génomes. L'ensemble des relations d'homologies est alors stocké dans des bases de données spécifiques à la méthode utilisée et mis à disposition sur Internet. La table 1 (**Table 1**) résume les méthodes utilisées, les noms des base de données et leurs adresses web associées.

Méthodes	Bases de données	Adresses web
COG : Cluster of Orthologous Group	COG database	http://www.ncbi.nlm.nih.gov/COG/
Inparanoid	Inparanoid	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
RSD : Reciprocal Smallest Distance	RoundUp	http://rodeo.med.harvard.edu/tools/roundup
Réconciliation d'arbres	HOBACGEN, HOVERGEN	http://pbil.univ-lyon1.fr/databases/hobacgen.html http://pbil.univ-lyon1.fr/databases/hovergen.html
PhyOP : Phylogenetic reconstruction of Orthology and Paralogy	Comparative gene server	http://genserv.anat.ox.ac.uk/

Table 1 : Liste (non exhaustive) des méthodes d'identification des relations d'homologies entre génomes.
Les méthodes utilisées, le nom des bases de données associées ainsi que les adresses web sont référencées.

1.3.2. La base de données Ensembl (2007)

Depuis sa version 39 (v.39 Juin 2006), la base de données Ensembl (http://www.ensembl.org) intègre une nouvelle méthode de prédiction des relations d'homologies au sein des familles de gènes basée sur la construction d'arbres phylogénétiques (Hubbard *et al.*, 2006). Chaque famille de gènes est générée en calculant les "meilleures relations réciproques" entre toutes les séquences traduites des gènes. Ces gènes sont regroupées en famille par une méthode de regroupement dite de liaison simple ("Single-linkage clustering"). Les groupes ou familles de séquences traduites en protéines sont alors alignés grâce au logiciel Muscle (Edgar *et al.*, 2004) et le programme PHYML

(Guindon *et al*., 2003; Guindon *et al*., 2005) va construire l'arbre phylogénétique non-enraciné par le principe du maximum de vraisemblance. Brièvement, le principe du maximum de vraisemblance consiste à sélectionner l'hypothèse (i.e la topologie de l'arbre associée à la longueur de ses branches et au modèle de substitution choisi) qui maximise la probabilité des données (l'alignement multiple des séquences homologues). L'idée principale de PHYML est d'appliquer simultanément plusieurs modifications locales de la topologie de l'arbre tout en ajustant l'ensemble des longueurs des branches. Pour des résultats équivalents aux méthodes de reconstruction d'arbres par maximum de vraisemblance, PHYML améliore très nettement le temps de calcul (12 sec pour des séquences de 500 pb appartenant à 100 taxa vs. plus de 25 min pour fastDNAml) et pave ainsi la voie aux études phylogénétiques sur génomes entiers i.e la phylogénomique. Finalement, l'algorithme RAP (Cf. ci-dessus) va réconcilier l'arbre des gènes avec l'arbre des espèces, déduire les événements de duplications au niveau des noeuds internes de l'arbre et ainsi préciser les relations plus complexes d'homologie (1:n et n:m) pour la famille de gènes considérée. Les relations d'orthologie entre l'homme et quatre génomes mammifères analysées par Ensembl sont présentées dans la table suivante (**Table 2**) :

Type de relation d'orthologie	**Chimpanzé** *(Pan troglodytes)*	**Chien** *(Canis familiaris)*	**Souris** *(Mus Musculus)*	**Rat** *(Rattus Norvegicus)*
1:0	1628	4703	4688	5026
1:1	19 901	14 997	14 923	14 265
1:n	1527	2697	2540	2913
m:n	204	413	659	606

Table 2 : Relations d'orthologie d'Ensembl (v.42) entre les 22 810 gènes humains codant pour des protéines et les quatre espèces : chimpanzé, chien, souris et rat.

Préalablement, les méthodes de prédiction des relations orthologues/paralogues nécessitent de connaître le catalogue exhaustif des gènes pour les organismes comparés. Cette tâche est d'autant plus grande que la taille des génomes des mammifères est importante. Un des défis majeurs de la bioinformatique est d'exploiter au mieux l'information contenue dans les génomes séquencés par

des approches d'identifications et de caractérisations structurales des gènes et, plus généralement, des séquences fonctionnelles (RNA, séquences régulatrices, promotrices). Nous allons donc préciser quelles sont les principales méthodes d'annotation des génomes et quel a été l'apport de la génomique comparative dans ces approches de prédiction de gènes.

1.4. L'annotation des génomes et les apports de la génomique comparée

Chaque génome séquencé constitue une ressource d'intérêt pour les études de génomique comparative. Plusieurs bases de données internationales sont dédiées au stockage, à l'analyse et à la mise à disposition des informations biologiques. De façon non exhaustive, nous pouvons citer quatre serveurs internationaux qui recensent des assemblages de génomes, UCSC (Kuhn *et al.*, 2007), NCBI (Pruitt *et al.*, 2007) et Ensembl (Hubbard *et al.*, 2007) et DDBJ (Sugawara *et al.*, 2007). Chacun de ces serveurs diffère par le nombre et le type de génomes stockés et par les outils employés pour l'analyse et la comparaison de ces génomes. Plus particulièrement, le serveur Ensembl, contrairement à NCBI et UCSC, a développé une procédure automatique d'identification des séquences fonctionnelles et plus précisément, d'annotation en gènes des génomes nouvellement séquencés (Curwen *et al.*, 2004) (Cf. chapitre sur l'annotation des génomes par Ensembl). Parallèlement pour l'homme et la souris, l'identification de l'ensemble du catalogue génique fait l'objet d'une annotation dite "manuelle" au travers de projets tels que Havana en Angleterre (Sanger Center), RefSeq aux Etats-Unis (NCBI) ou encore Uniprot en Suisse. Ces deux types d'annotation ont pour but de disposer d'un répertoire de séquences fonctionnelles le plus complet et précis possible pour des organismes modèles en biologie. Il est à noter, toutefois, que d'autres organismes modèles sont gérés par des bases de données spécifiques au génome étudié (Bee Base pour l'abeille (*Apis mellifera*) ou Worm Base pour les nématodes *Caenorhabditis elegans* et *Caenorhabditis briggsae*). Ainsi, plus de la moitié des génomes stockés et analysés par Ensembl appartient au groupe des mammifères.

1.4.1. Annotation des génomes par ENSEMBL

1.4.1.1. Le processus standard d'annotation d'Ensembl

À la difficulté technique due à la taille des génomes analysés (~3Gb pour l'homme), la prédiction de gènes chez les mammifères est rendue plus complexe par la forte présence de séquences *a priori*

non-fonctionnelles (“junk DNA”) représentant une part importante de la totalité du génome chez l’homme (les dernières estimations étant beaucoup plus nuancées (ENCODE, 2007)). Les prédictions d’Ensembl sont basées sur des preuves expérimentales lorsqu’elles sont disponibles puisque les séquences de gènes inférés sont alignées à partir de séquences nucléiques ou protéiques issues de bases de données telles que RefSeq, Swiss-Prot/UniProt. Étant donné la quantité d’informations à analyser, la stratégie employée par Ensembl est nécessairement informatique et consiste à réduire au maximum l’espace de recherche des “îlots géniques dans l’océan de génome”. Deux étapes complémentaires menées en parallèle vont permettre de définir des “mini-séquences” à partir desquelles le logiciel de prédiction de gène Genewise (Birney *et al.*, 2004) va pouvoir prédire, dans un temps de calcul raisonnable, la structure intron-exon du gène. La première étape va aligner toutes les séquences protéiques et nucléiques (ARNm, ADNc) spécifiques de l’organisme et disponibles dans les bases de données relatives au génome nouvellement séquencé. Puis, une seconde étape se base sur les similarités de séquences avec des gènes d’espèces relativement proches pour retrouver des transcrits non identifiés dans la première étape. Cette deuxième étape est particulièrement importante dans le cas de génomes où la quantité d’information biologique (ARN messager, Expressed Sequence Tag (EST), protéines) spécifique à l’organisme séquencé est faible (c’est le cas pour de nombreux projets de séquençage d’espèce séquencée à 2X telles que celles du “Mammalian Genome project” - http://www.broad.mit.edu/mammals/). Ces deux premières étapes permettent de réduire fortement la taille de la région génomique de 3 Gb à ~1 Mb pour l’homme et ainsi localiser “grossièrement” les gènes dans les génomes. Cependant, les programmes d’alignement utilisés (tels que BLAST) ne permettent pas de définir la structure exon-intron basée sur un modèle d’épissage caractéristique des gènes eucaryotes. Le programme Genewise va alors être utilisé pour prédire le gène dans la région identifiée à partir d’une protéine de référence (ou parente). Le principal écueil de Genewise, néanmoins, est son temps de calcul. En effet, l’alignement de 40 000 protéines sur des régions de 100 kb en moyenne durerait environ 6 mois CPU ce qui exclut toute possibilité de mises à jour rapides des différentes versions de la base de données Ensembl. Une étape finale va alors consister à aligner chacun des exons de la protéine de référence sur la région génomique (traduite dans les six phases de lecture) et concaténer les exons alignés correspondants (plus 200 pb introniques) pour former une “miniséquence” opérationnelle pour Genewise (Birney *et al.*, 2004). En utilisant cette procédure à trois temps (positionnement global du gène - positionnement large des exons - alignement final), toutes les protéines humaines, par exemple, peuvent être alignées sur le génome en quelques heures sur 400 processeurs.

1.4.1.2. La mise à jour du processus d'annotation standard

Le processus standard d'annotation des gènes décrit ci-dessus a été modifié en fonction de la nature des génomes analysés et de la profondeur de leur séquençage. On peut définir deux situations possibles pour les deux objectifs suivants : L'annotation en gènes d'un génome fragmenté et l'annotation en gènes pour les génomes ne disposant que de peu de données propres à l'organisme.

(i) Annotation des génomes dont l'assemblage est très fragmenté car issu des projets de séquençage partiels (couverture de 2X)

Le National Institute of Health (NIH) a financé un projet de séquençage léger (2X) de 16 génomes mammifères : le "mammalian genome project". Neuf génomes sont d'ores et déjà intégrés et analysés par Ensembl : éléphant (*Loxodonta africana*), lapin (*Oryctolagus cuniculus*), tatou (*Dasypus novemcinctus*), tenrec (*Echinops telfairi*), chat (*Felis catus*), galago (*Otolemur garnettii*), musaraigne carrelet (*Sorex araneus*), microchiroptère (*Myotis Lucifugus*) et hérisson européen (*Erinaceus europaeus*). Le processus standard (Curwen *et al*., 2004) pour la construction du répertoire de gène ne peut pas être appliqué sur des assemblages aussi fragmentés. L'approche utilisée consiste alors à aligner un génome complet de référence sur ces génomes fragmentés par l'outil d'alignement de séquence Blastz. Les gènes annotés pour le génome de référence sont alors alignés et "projetés" sur les portions de séquence ("contigs") de l'assemblage fragmenté. De nombreuses insertions/délétions sont présentes dans les gènes prédits dues à la nature très morcelée du génome analysé, ce qui aboutit à des ruptures du cadre de lecture. Ensembl corrige ce biais par l'insertion d'un intron artificiel (1 à 2 bp) ("ou frame-shift intron") dans l'exon pour rétablir le cadre de lecture. Un exon est exclu de l'analyse si plus de 50% de l'exon de référence n'est pas aligné sur le gène prédit. Cette procédure a été appliquée au génome de la vache (*Bos taurus*) initialement séquencé en 3X (Hubbard *et al*., 2007). En comparant les résultats obtenus avec l'annotation des génomes bénéficiant d'un séquençage léger, il est intéressant de remarquer que le passage d'un séquençage 3X à 2X double la proportion d'exons non identifiés sur le génome testé par cette méthode.

(ii) Annotation des génomes bénéficiant d'un séquençage profond mais ne disposant que de peu d'informations de transcrits pour leur génome.

Pour chaque génome à analyser, le processus standard d'annotation intègre une étape préalable d'alignement d'ESTs et de ADNc d'espèces de référence sur le génome testé. Cette étape permet la

création d'une "miniséquence" ne contenant que les exons et une partie des introns du gène prédit. Ceci améliore significativement la rapidité de prédiction de la structure finale du gène par le programme GeneWise. Cependant, pour des génomes distants au niveau évolutif tels que l'oppossum ou le poulet, la qualité des alignements initiaux des séquences d'EST et ADNc de référence est très faible et aboutit à des fusions incorrectes de plusieurs transcrits pour des gènes adjacents. Dans ces cas particuliers, l'étape de création initiale d'une "miniséquence" a été écartée au profit d'une amélioration du logiciel GeneWise pour permettre la prédiction d'une structure de gène à partir d'une région génomique plus grande de l'ordre de la dizaine de kb. Néanmoins, il en résulte que le temps de calcul pour la construction du répertoire génique des génomes distants a été multiplié par cinq.

Parallèlement au perfectionnement des annotations des génomes nouvellement séquencés, de nombreux efforts portent sur l'amélioration de la qualité et le développement de nouvelles méthodes de prédictions de gènes pour les génomes humain et murin. Les différents programmes de prédictions de gènes testés peuvent être classer en plusieurs catégories détaillées dans le chapitre suivant (voir Brent *et al*., 2004 et Guigo *et al*., 2006 pour revues).

1.4.2. Les programmes de prédictions des gènes

1.4.2.1. Les méthodes basées sur des alignements de séquences ADNc et protéines

Pour des organismes modèles en biologie, le nombre de séquences d'ESTs et de protéines disponibles dans les bases de données permet d'envisager l'annotation des gènes par de "simples" alignements de ces séquences. C'est le cas pour le processus d'annotation standard d'Ensembl (Curwen *et al*., 2004) dont la première étape ("Targetted build") consiste à aligner tous les ESTs et protéines sur le génome "cible" afin de localiser les gènes. Cependant, ces alignements comme nous l'avons décrit, ne permettent pas de rendre compte de la structure intron-exon du gène mais sont, néanmoins, très sensibles (car basées sur des preuves expérimentales) et complémentaires aux méthodes *ab initio*.

1.4.2.2. Les méthodes de prédictions ab initio

On pourra distinguer deux types d'approches en fonction du nombre de génome(s) analysé(s) :

- Prédictions utilisant un seul génome

Historiquement, ce sont les premiers programmes utilisés et aujourd'hui encore très populaires. Ils sont basés sur des modèles statistiques et, plus particulièrement, le Modèle de Markov Caché (MMC ou HMM pour Hidden Markov Model). Ces modèles sont entraînés sur des structures de gènes connues (1999 gènes pour GENSCAN par exemple!) (Burge *et al.*, 1998) afin de reconnaître dans les génomes les structures particulières des gènes (codon start, codon stop, sites accepteurs et donneurs d'épissage...). Bien que très rapides, ces approches sont 'victimes' de leur sensibilité et définissent un grand nombre de faux-positifs. Ceci est d'autant plus vrai que le génome étudié est peu compact. Aussi, GENSCAN prédit respectivement ~ 70 000 et 50 000 gènes pour les génomes humain et canin. Par contre, pour des génomes plus compacts comme celui de la plante modèle *Arabidopsis Thaliana*, près des deux tiers des gènes prédits le sont correctement. Les programmes GENSCAN, GENEID, FGENESH, HMMGENE et GENEZILLA représentent les principaux programmes de prédictions de gènes n'utilisant qu'un seul génome.

- Prédictions utilisant deux ou plusieurs génomes

Ces programmes ont été développés parallèlement à la mise à disposition des séquences de génomes et illustrent la puissance de la génomique comparative pour la prédiction de gène. Les prédictions reposent sur le principe général selon lequel les régions fonctionnelles d'un génome, et en particulier, les gènes codant pour des protéines, sont plus conservées au cours de l'évolution que les séquences non fonctionnelles. L'alignement avec le génome informant (dit de référence) va permettre de renforcer et renseigner la prédiction de gènes faite par les méthodes standards. Les programmes TWINSCAN, SGP2, GENEWISE et AUGUSTUS-dual, par exemple, sont les principaux programmes de prédictions de gènes se basant sur un ou plusieurs génomes référents. Ces méthodes ont permis d'améliorer très nettement la sensibilité des prédictions de gènes mais posent le problème du choix du génome informant à utiliser (Boffelli *et al.*, 2004). Si un génome de référence est à une distance évolutive trop faible par rapport au génome à annoter, la trop grande similitude de séquence ne permettra pas de distinguer les séquences réellement fonctionnelles. Alors qu'une proportion de 40% du génome humain s'aligne avec celui de la souris, moins de 2% du génome humain correspond à des exons codant pour des protéines. En 2003, Wang *et al.*, concluaient que le génome qui permettrait d'annoter de manière optimale le génome humain devait

être situé au niveau phylogénétique entre celui du rat (*Rattus Norvegicus*) et celui du poulet (*Gallus gallus*) (Wang *et al.*, 2003). Cependant la séquence d'espèces très distantes au niveau phylogénétique peut apporter des informations nouvelles pour d'autres génomes. En 2004, par exemple, la comparaison de la séquence des gènes du poisson *Tetraodon Nigroviridis* avec celui de l'homme a permis de révéler plus de 900 nouveaux gènes humains initialement non annotés (Jaillon *et al.*, 2004).

Les méthodes se basant sur plusieurs génomes référants (N-SCAN, Arumugam *et al.*, 2006; EUGEN'HOM, Foissac *et al.*, 2003) fournissent une approche prometteuse pour la prédiction de gènes. L'objectif est d'affiner la mesure du taux d'évolution à chaque nucléotide en alignant plusieurs génomes dont le génome à annoter. Ceci a abouti à combiner l'approche phylogénétique et l'approche de prédiction de gènes au sein de modèles dits "Phylo-HMM" ou "Evolutionary HMM" (Siepel *et al.*, 2005).

- Les méthodes combinant plusieurs évidences

Les experts biologistes qui annotent manuellement les génomes se basent classiquement sur la combinaison des approches précédemment citées afin de décrire au mieux la structure et la fonction d'un gène. Le programme Exogean (Djebali *et al.*, 2006), par exemple, a pour objectif de modéliser le raisonnement humain par un ensemble de règles de décision afin d'intégrer et de combiner les méthodes *ab initio* et les alignements d'EST et de protéines.

Au printemps 2005, le projet EGASP (human Encode Genome Annotation Assessement Project) (Guigo *et al.*, 2006), en marge du projet ENCODE, a eu pour objectifs de comparer les performances de différents logiciels de prédiction de gènes sur 44 régions génomiques représentant 30 Mb (1% du génome) préalablement annotées par des "annotateurs" experts du groupe Havana au Sanger Institute. Les performances ont été analysées selon des critères de Sensibilité (Sn) représentant la proportion des annotations Havana correctement prédites et de Spécificité (Sp) correspondant à la proportion des prédictions d'une méthode donnée qui identifie correctement les prédictions Havana (**Fig.6**). Les résultats varient en fonction de la méthode utilisée et du niveau d'analyse considéré : niveau du gène, de l'exon ou du nucléotide. Au niveau nucléotidique correspondant à la proportion de nucléotides annotés par les experts et qui sont correctement prédits, les meilleurs programmes atteignent 90% de précision pour les paramètres de sensibilité. Au niveau du gène, les programmes reproduisant les annotations "humaines" sont les plus précis (70% de moyenne ((Sp+Sn)/2)). Par ailleurs, 221 exons ont été sélectionnés parce qu'ils étaient prédits par plusieurs méthodes informatiques en dehors des annotations des experts. Parmi ces

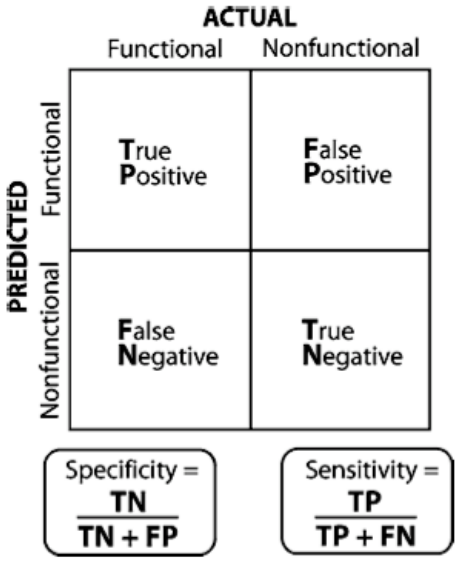


Figure 6 : Calcul de la Spécificité (Sp) et de la Sensibilité (Sn) pour tester la qualité des prédictions de gènes.

exons, seuls 3,2% ont pu être vérifiés expérimentalement par RT-PCR soulignant la fiabilité et l'importance de l'annotation humaine comme référence. Néanmoins, cela reflète aussi les difficultés inhérentes à la validation biologique des prédictions de gènes dont l'expression peut être liée à un stade particulier du développement ou à des conditions environnementales difficilement reproductibles au laboratoire. De plus, le meilleur transcrit annoté par les approches les plus performantes (Ensembl, EXOGEAN) différait, toutefois, des annotations manuelles de référence dans 30% des cas. Les vérifications, *a posteriori*, des erreurs de prédiction ont montré que celles-ci concernaient bien souvent les gènes dupliqués et groupés en "cluster" très denses.

La prédiction de gènes codant pour des protéines a largement bénéficié des approches de génomique comparative incluant deux ou plusieurs génomes. Environ 75% de tous les exons humains connus peuvent être correctement identifiés au niveau de leurs deux sites d'épissage par ces approches comparatives. Les progrès réalisés dans d'identification des pseudogènes, dont le but est d'éliminer le trop grand nombre de faux-positifs, associés avec la prédiction des gènes par comparaison de paire ou de multiples génomes constituent les avancées majeures pour la prédiction de gènes ces dernières années (Zheng *et al*., 2006).

Cependant, la relative proportion de gènes codant pour des protéines (<2%) par rapport à la proportion de séquences fonctionnelles non-codantes jusqu'à 10% selon les dernières études, (Pheasant *et al*., 2007) a amené un grand nombre de laboratoires à développé de nouvelles

méthodes comparatives dont le but est d'affiner l'identification des séquences non-codantes conservées entre espèces au cours de l'évolution et essayer d'inférer leurs rôles fonctionnels.

1.4.3. Approches comparatives pour l'annotation des séquences fonctionnelles non codantes

Une des grandes leçons des projets de séquençage et des annotations initiales des génomes est l'absence de corrélation entre nombre de gènes codant pour des protéines et la complexité d'un organisme. Le génome humain contient aux alentours de 22 000 gènes en fonction des différentes estimations (Lander *et al.*,2003; Lindbald-Toh *et al.*, 2005; Goodstadt *et al.*, 2006), celui récent de l'anémone de mer *Nematostella vectensis* près 18 000 (Putnam *et al.*, 2007) ou encore, le séquençage du génome d'un protiste *Tetrahymena thermophila* a permis d'identifier environ 27 000 gènes (Eisen *et al.*, 2006). Aussi, pour expliquer les différences de complexité biologique d'un organisme, l'annotation des séquences fonctionnelles non-codantes (séquences régulatrices, ARN non codants, sites de fixation à des facteurs de transcription, séquences assurant l'intégrité structurale des chromosomes...) a vu le développement d'un nombre considérable de méthodologie et d'analyses qui nécessitent intrinsèquement de comparer les génomes. Contrairement aux séquences codant pour des protéines très conservées au cours de l'évolution et qui présentent des signatures spécifiques (codons start, codons stop...), les séquences fonctionnelles non codantes et, plus particulièrement, les séquences régulatrices, sont plus courtes, moins préservées et sujettes à un fort renouvellement. La plupart des approches requièrent la réalisation d'un alignement multiple de génomes entiers, la distinction entre régions sous sélection neutre ou positive (favorable à la variabilité génétique) et régions sous sélection négative (conservées au cours de l'évolution) et enfin, la prise en compte des distances phylogénétiques séparant les espèces étudiées (Boffelli *et al.*, 2004). De nombreuses méthodes se basent sur l'analyse du taux de substitution des séquences d'intérêt alignées en rapport avec le taux de substitution de certaines classes de séquences sensées refléter une évolution neutre (Margulies *et al.*, 2005). Ces différentes classes de séquences incluent les sites synonymes (principalement à la troisième base des codons), les séquences en dehors des exons spécifiques d'une lignée et les séquences répétées ancestrales. Cependant, différentes études montrent que ces classes d'éléments n'évoluent peut-être pas de façon neutre et seraient aussi contraintes (Pheasant *et al.*, 2007 ; ENCODE, 2007). Par exemple, le séquençage récent de l'oppossum (*Monodelphis domestica*) a montré que 16% des séquences très conservées au cours de l'évolution des euthériens (placentaires) étaient dérivés de séquences répétées ancestrales (Mikkelsen *et al.*, 2007). De plus, une séquence possédant un taux de substitution différent du taux

de substitution neutre attendu, ne signifie pas obligatoirement que cette séquence est sous sélection. Récemment, Pollard *et al.,* ont identifié près de 50 régions (i) d'une taille de 100 à 500 pb, (ii) très conservées (>96% d'identité) entre de nombreux mammifères (dont le chimpanzé) (iii) et enfin localisées en dehors régions codantes des génomes (Pollard *et al*., 2006). Chez l'homme, ces régions présentent un taux de substitution très supérieur au taux de substitution attendu pour des séquences sous évolution neutre. Les auteurs proposent alors que ces régions dont le taux de substitution est accéléré (HAR Human Accelerated Regions) seraient sous sélection positive (ou adaptative) et pourraient être impliquées dans l'évolution spécifique de l'homme comparée à celle des chimpanzés. Cependant, en étudiant la taille des HARs, leur localisation au niveau de point chaud de recombinaison et enfin leur patron de substitution très orienté vers des substituions enrichies en GC, Galtier *et al*., proposent que la plupart des HARs répondent plus sensiblement au modèle de conversion génique biaisée (BGC Biased Gene Conversion) plutôt qu'à un modèle de sélection positive (Galtier *et al*., 2007).

Au-delà de la comparaison des génomes à l'échelle nucléotidique ou de la séquence fonctionnelle, généraliser cette approche dans un vision plus globale du contexte génomique permet de comprendre l'organisation, la dynamique et l'évolution des gènes entre plusieurs espèces.

2. La comparaison des génomes dans le contexte chromosomique

2.1. Le commencement : la cartographie comparée

2.1.1. Introduction

Projet jusqu'à présent préalable ou complémentaire à l'étape de séquençage d'un génome, la cartographie consiste à positionner et à ordonner des séquences d'ADN ou marqueurs sur les chromosomes d'un génome et à déterminer la distance qui les sépare. On distingue, classiquement, plusieurs types d'approches en cartographie dont principalement, la cartographie cytogénétique, la cartographie physique et la cartographie de liaison. Cette dernière comprend deux méthodologies, la première est la cartographie génétique (Weissenbach *et al*., 1992 ; Rhodes *et al*., 1998) qui basée sur l'analyse de la ségrégation de marqueurs au sein d'une famille et consiste à positionner les marqueurs de façon relative à leur fréquence de co-ségrégation au sein de famille. Puis, la

cartographie par hybride d'irradiation (RH) qui repose sur l'analyse de la fréquence de cassures chromosomiques induite entre marqueurs par l'irradiation du génome (Barrett *et al*., 1992). La cartographie physique cherche à positionner des fragments d'ADN clonés les uns par rapport aux autres et permet de déterminer une distance réelle en nucléotides entre les clones. Il est important de noter que la distance physique entre deux marqueurs est une mesure absolue, contrairement aux cartes de liaisons qui établissent des distances relatives issues de traitements statistiques de la fréquence de cassures ou de recombinaisons entre les marqueurs.

Avant l'avènement des techniques de séquençage à haut-débit et la possibilité de comparer des génomes entiers à l'échelle même du nucléotide, la cartographie comparée a permis de définir les fondements de la comparaison de génomes en cherchant à identifier les régions conservées ou au contraire spécifiques d'un génome par rapport à un autre et ainsi apporter des éclairages sur les mécanismes évolutifs des génomes étudiés. L'unité ponctuelle de comparaison, le marqueur, est un fragment d'une séquence d'ADN (gènes, ESTs, clones, BACs...) qui doit être, par principe, homologue entre les génomes à comparer. En fonction de la nature et de la densité des marqueurs de comparaison entre deux génomes, deux méthodes de cartographie se prêtent à l'étude comparée des génomes : la cartographie cytogénétique et la cartographie par hybride d'irradiation (RH).

2.1.2. La cartographie cytogénétique

L'avènement de techniques d'hybridation fluorescente in situ (FISH), véritable coloriage chromosomique a permis, dès le milieu des années 1990, de mettre en évidence de grandes régions d'orthologie entre des espèces relativement éloignées au niveau phylogénétique (primates, rongeurs, ongulés, cétacés) (Scherthan *et al*., 1993 ; Raudsepp *et al*., 1996). Des études comparatives ont ainsi identifié 30 régions ou blocs conservé(e)s entre les génomes de l'homme et du chat ou bien 47 entre le génome du porc et de l'homme (Menotti-Raymond *et al*., 2003). Toutefois, la faible résolution de cette méthode (~1-2 Mb) ne permet pas de disposer d'un maillage dense et ne renseigne pas sur les éventuelles inversions présentes au sein des blocs d'homologie. (O'Brien *et al*., 1999)

2.1.3. La cartographie par hybrides d'irradiation (RH)

Dans sa phase expérimentale, la cartographie RH revient à établir une collection de clones cellulaires hybrides contenant le génome hôte –souvent issu du hamster- et une fraction aléatoire du

génome testé dont le contenu est différent d'un hybride à l'autre. L'ensemble des hybrides cellulaires, constitue un panel RH (**Fig.7**). Le principe de la cartographie repose sur l'analyse du patron de la distribution des marqueurs sur le panel RH selon le principe suivant : plus deux marqueurs seront proches sur le génome, plus leurs distributions seront semblables. En effet, si deux marqueurs sont proches, la probabilité qu'ils soient séparés par une cassure liée à l'irradiation sera faible (**Fig.7**). Un des avantages de la méthode RH par rapport à une approche cytogénétique réside dans la densité plus élevée de marqueurs qui peuvent être ancrés sur le génome. De plus, contrairement aux cartes de liaison génétique qui nécessitent de distinguer les allèles des marqueurs utilisés, la cartographie RH teste la présence ou l'absence du marqueur dans les différentes lignées du panel (étape de génotypage). Ainsi des marqueurs de gènes, d'ESTs, et plus généralement de toutes les séquences uniques et orthologues entre deux génomes pourront servir de marqueurs afin de déterminer des ancres de comparaison pour établir des relations de synténie entre les cartes des génomes. Les cartes comparées entre espèces possédant une carte RH relativement dense (porc, boeuf, cheval, chien et poulet) et le génome humain ont permis l'identification de 90 blocs de synténie entre les génomes du porc et l'homme et 44 entre celui du cheval et l'homme. En 2003, près de 900 marqueurs ont été cartographiés sur le génome du chat et, par comparaison avec le génome de l'homme, ont permis de définir 110 segments conservés ordonnés (Menotti-Raymond *et al*., 2003).

Cependant, la disparité dans le nombre de segments conservés observés entre deux mêmes génomes est à relativiser car elle est directement proportionnelle à la quantité d'ancres de comparaison utilisées. En effet, un nombre élevé de points de comparaison entre les génomes permet d'obtenir un maillage dense et plus complet pour identifier de manière résolutive les relations de synténie. Or, la disponibilité de marqueurs pour un organisme non séquencé peut être une étape limitante. Des données d'ESTs ou de séquences partielles de gènes peuvent être néanmoins utilisées. Récemment la production de séquençage à faible niveau de couverture (< 2X) a généré une masse considérable de données qui peuvent alors être exploitées en cartographie RH.

2.1.4. Approche mixte séquençage léger - cartographie RH

Une approche efficace pour la quantité d'informations obtenue en fonction du coût d'exploitation consiste à combiner une approche de séquençage léger d'un génome (1-2X) avec la construction d'une carte dense d'hybrides irradiés. Le cas du génome du chien illustre cet exemple puisqu'en 2005, notre laboratoire en collaboration avec deux groupes américains, a réalisé un projet de

cartographie de plus de 10 000 marqueurs répartis sur les 40 chromosomes du génome canin (Hitte *et al*., 2005). Le premier séquençage léger d'un génome mammifère, a été réalisé pour le génome canin avec une couverture de séquence de 1.5X (Kirkness *et al*., 2003). Ce séquençage a produit plus de cinq millions de séquences qui ont été exploitées par des alignements réciproques entre les séquences canines et le catalogue des gènes humains utilisé comme génome de référence. Ainsi plus de 15 000 fragments de gènes canins ont pu être identifiés, à partir desquels nous avons utilisé 10 000 marqueurs de gènes (sélectionnés selon leur disposition régulière sur le génome humain) comme ancres de comparaison en cartographie RH. Après construction de la carte RH canine, les 10 000 marqueurs de gènes se répartissent sur 9 080 positions en raison de la co-localisation de marqueurs sur une même position génomique. Cette approche a permis d'identifier un ensemble de 264 segments conservés supérieurs à 500 kb entre les génomes de l'homme et du chien (Lindblad-Toh *et al*., 2005). Cette cartographie a été utilisée par le consortium international du séquençage pour guider la phase d'assemblage du séquençage dit complet (7.5X) du génome canin finalisé fin 2005. Parmi les enseignements tirés de cette cartographie RH à haute résolution, nous avons retenu trois points essentiels (i) la capacité à identifier la totalité des CS et CSO supérieurs à 500 kb entre paires de génomes, (ii) la bonne colinéarité observée entre le séquençage et carte RH (coefficient de corrélation de Spearman r=0.93) et (iii) la forte conservation de l'ordre des gènes au sein des CSO que l'on peut observer par l'approche RH (Hitte *et al*., 2005).

2.2. La conservation de l'ordre des gènes

Avant même le séquençage de nombreux organismes, la localisation de marqueurs de gènes par cartographie avait montré que leur organisation dans les génomes n'était pas aléatoire. Plus particulièrement, au sein des mammifères, l'ordre des gènes est stable (Chicken genome sequencing Consortium, 2004). Le séquençage très récent du génome de l'anémone de mer *Nematostella* qui a divergé de l'homme il y a ~700 millions d'années, a permis d'identifier environ 18 000 gènes dont une majorité se regroupe dans 98 segments conservés avec le génome humain (Putnam *et al*., 2007). La conservation de l'ordre des gènes qui s'observe entre deux génomes est dépendante du niveau de proximité phylogénétique mais ne constitue pas le seul déterminant. Ainsi, la dynamique d'ordonnancement des gènes corrèle avec leur expression au sein des tissus cellulaires. Par exemple, chez *Drosophilia melanogaster*, 45% des gènes qui sont exprimés dans les testicules sont regroupés sous forme d'une "suite" d'au moins quatre gènes. Différents travaux ont montré qu'il existait une relation entre les regroupements de gènes le long des chromosomes en "clusters" et

leurs expressions ou leurs fonctions associées (Hurst *et al*., 2004 pour revue). Si on extrait les ESTs d'un tissu particulier, les gènes correspondant à ces séquences ont tendance à être regroupés en clusters sur les génomes ou encore, le regroupement de gènes sous forme de "clusters" implique souvent une même voie métabolique (Sonnhammer *et al*., 2001). Ceci est particulièrement vrai chez la levure où 98% des voies métaboliques extraites à partir de la base de données KEGG (Kyoto Encyclopaedia of Genes and Genomes ; Kanehisa *et al*., 2006) sont corrélés à la présence de gènes en cluster au sein de leur génome (Lee *et al*., 2003). De plus, les régions du génome de la levure où sont localisés les gènes impliqués dans des processus essentiels à la survie montrent un faible taux de recombinaison suggérant de forte(s) contrainte(s) à l'interruption de l'ordre des gènes. Même si l'on tient compte du biais engendré par les opérons qui peuvent réguler plusieurs gènes ordonnés en aval (chez *Caenorhabditis elegans,* par exemple) ou aux fréquentes duplications en tandem de gènes donnant naissance à des familles multi-géniques chez les mammifères, la relation entre l'ordre des gènes et leur co-expression a été observée pour de nombreux exemples (cluster de gènes exprimés dans les muscles, gènes de ménage, régions dont l'expression est augmentée dans les tumeurs...). On peut alors se poser la question des contraintes biologiques qui agiraient pour maintenir "le respect de l'ordre" des gènes au sein des clusters. Hurst *et al*., ont suggéré que l'ordre des gènes et donc leur expression serait lié à deux états de conformation de la chromatine (Hurst *et al*., 2004). Une conformation dite ouverte permettant l'expression des gènes et une conformation fermée (telle que dans l'hétérochromatine) empêchant l'expression des gènes. Cependant, une étude plus récente (Sémon *et al*., 2006) montre que la majorité des clusters de gènes co-exprimés renfermerait plutôt un faible nombre de gènes (~ 2). La relation entre les clusters de gènes et leur co-expression serait alors expliquée par la présence de promoteurs bi-directionnels qui réguleraient dans les deux sens l'expression du cluster.

Si l'analyse comparée de l'organisation des génomes permet d'observer une conservation de l'ordre des gènes par blocs, force est de constater qu'au cours de l'évolution des génomes, l'organisation des génomes est interrompue par des réarrangements chromosomiques qui varient selon leurs tailles (micro et macro-réarrangements), leurs natures (translocation, inversion, fission, fusion) et leurs fréquences d'apparition. Comparer l'ordre des gènes entre deux génomes permet alors de repérer les segments conservés ordonnés, les limites de ces segments que sont les ruptures de synténie et d'analyser les forces sélectives sous-jacentes à la dynamique des génomes.

2.3. Les réarrangements génomiques

2.3.1. Nature des réarrangements génomiques

Les principaux réarrangements chromosomiques sont les inversions, les translocations, les fusions-fissions et les duplications-délétions (**Table 3**). Il peuvent être spécifiques d'une lignée (exemple des maladies génétiques humaines liées à des réarrangements chromosomiques) mais sont aussi mis en évidence en comparant l'organisation des chromosomes entre deux espèces relativement proches au niveau phylogénétique.

Type de Réarrangements	Définition	Exemple
Inversion	concerne une région d'un chromosome qui est retournée sans modifications de sa position	HSA19 (~41 Mb) chimpanzé 19
Translocations		
- Translocations réciproques	échanges de "morceaux" d'ADN à partir de deux chromosomes non homologues	HSA22 - HSA9 : le chromosome de Philadelphie
- Translocations de type transposition	échanges d'une région d'ADN vers une autre sans réciprocité	les transposons
Fusion - Fission	concerne la fusion de deux chromosomes pour n'en former qu'un, au contraire de la fission qui va casser un chromosome en 2	HSA2 fusion des deux chromosomes PTR2A et PTR2B du chimpanzé
Duplications	créations d'une copie supplémentaire d'une portion d'ADN (Cf. chap sur les duplications)	duplication en tandem ou duplication inversée
Délétions	Suppressions de régions d'ADN de taille variable	CNV : Copy Number Variation

Table 3 : Définition des principaux réarrangements chromosomiques.

La nature des événements de réarrangements varie en fonction du phylum considéré. Par exemple, plusieurs forces ont modelé le génome des levures *Saccharomyces* telles qu'une duplication entière

de son génome mise en évidence par la présence de double copie des gènes (paralogues) par comparaison avec le génome d'une autre levure *Kluyveromyces*. D'autres mécanismes participent à l'instabilité des génomes de levure. Ainsi, les travaux de Fischer *et al*., ont recensé l'ensemble des micro-inversions présentes au sein des blocs conservés entre 11 lignées de levure *Hemiascomycètes* (Fischer *et al*., 2006). Les auteurs ont pu, entre autre, démontrer que les levures présentant le plus fort taux de d'instabilité correspondaient aux levures pathogènes *Candida albicans* et *Candida glabrata*. Chez les vertébrés (exceptés les poissons osseux téléostéens et quelques amphibiens ayant subi une duplication entière de leur génome), les principaux mécanismes de réarrangements correspondent à des réarrangements de type fusion-fission, inversions et translocations et surtout des expansions très importantes des familles de séquences répétées (Coghlan *et al*., 2005).

Comme pour les levures, le taux de réarrangements varie significativement chez les mammifères. Par exemple, les génomes du chien et du poulet ont subi beaucoup moins de réarrangements que celui des rongeurs depuis la séparation de leur ancêtre commun (~ 300 millions d'années). La taille des génomes variant significativement entre espèces, on peut calculer le taux de réarrangements en divisant le nombre de réarrangements survenus depuis la divergence avec l'ancêtre commun en fonction de la taille des génomes. Par exemple, le taux de réarrangements depuis l'ancêtre commun entre l'homme et les rongeurs est de 3,2-3,5 réarrangements par Mb et par millions d'années pour les rongeurs alors qu'il n'est que de 1,6 par Mb et réarrangements par millions d'années pour l'homme (Coghlan *et al*., 2005). De la même façon, le programme AutoGRAPH, que nous avons développé au laboratoire (Derrien *et al*., 20007), permet de formaliser l'identification et la caractérisation des segments conservés ordonnés qui ont échappé aux événements de réarrangements depuis leur ancêtre commun. Le nombre de segments conservés ordonnés mis en évidence entre l'homme et le chien (220) est inférieur à celui observé entre l'homme et la souris (340) ce qui indique que les génomes humain et canin sont plus "stables" que celui de la souris.

2.3.2. Les outils de visualisation des réarrangements génomiques

Pléthore d'outils bioinformatiques sont dédiés à la mise en évidence des relations de synténie entre génomes (NCBI MapViewer Wheeler *et al*., 2002 ; Tesler *et al*., 2002 ; Iccare Muller *et al*., 2004 ; Synbrowse Pan *et al*., 2005 ; YGOB Byrne *et al*., 2006 ; cinteny Sinha *et al*., 2007 ; Ensembl Synteny View Hubbard *et al*., 2007). Ils diffèrent par les méthodes employées pour l'identification des segments conservés et des segments conservés ordonnés, la nature et le nombre de génomes analysés, les interfaces graphiques proposées en sortie, le niveau de résolution des cartes

synténiques (génomes entiers, chromosomes, gènes, nucléotides...) ou encore la possibilité d'utiliser des données personnelles ou déjà insérées. Les programmes les moins "rigides" permettent de modifier les options d'identification des segments conservées et le nombre minimum d'ancres de comparaison nécessaire à la définition d'un segment conservé. Par exemple, Ensembl Synteny Viewer (**Fig.8**) (Hubbard *et al.*, 2007) utilisent des ancres de comparaison importées du serveur UCSC (Kuhn *et al.*, 2007) et définies à partir de l'alignement de deux génomes entiers par l'outil BLASTZ (Schwartz *et al.*, 2004). Généralement, un segment synténique est identifié si l'ordre des ancres est respecté entre deux chromosomes et si la distance séparant deux ancres ne dépasse pas 100 kb. Par ailleurs, le serveur GRIMM (Tesler *et al.*, 2002 ; Pevzner *et al.*, 2003) sera détaillé dans le paragraphe détaillant l'évolution des chromosomes à la fin de cette partie. Enfin, la nécessité de disposer d'un outil versatile et multi-fonctionnel de comparaison de génomes nous a motivé à développer un nouvel outil baptisé AutoGRAPH (Derrien *et al.*, 2007) qui fera l'objet d'une description détaillée dans le premier chapitre de la partie Résultats.

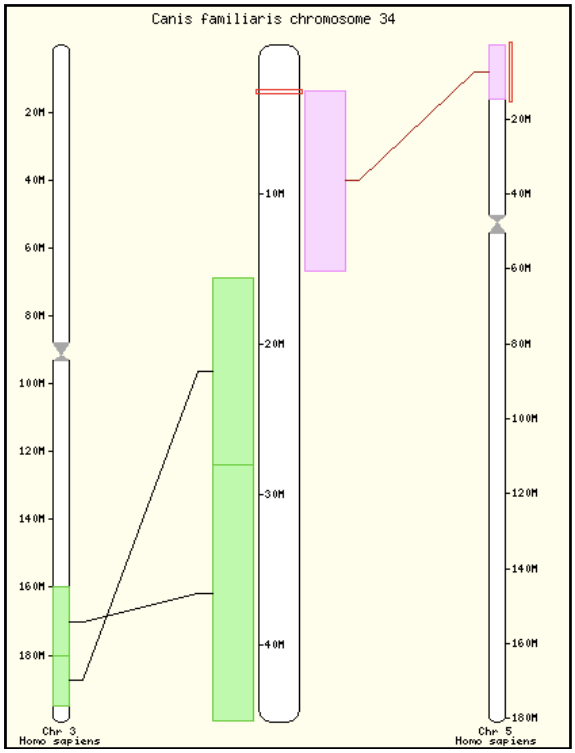


Figure 8 : Ensembl synteny viewer.
Carte de synténie entre le chromosome 34 du chien au centre (en blanc) et le génome humain. Trois segments conservés ordonnés sont identifiés dont deux avec le chromosomes 3 humain (vert) et un avec le chromosome 5 (rose).

2.4. Les duplications : mécanismes majeurs de l'évolution

Les réarrangements de type duplications jouent un rôle majeur au cours de l'évolution des génomes et peuvent être classifiés en fonction de la taille de la région génomique impliquée. On note ainsi des Duplications Globales de Génomes (DGG ou polyploïdie ou encore WGD pour Whole Genome

Duplication) ou des duplications de régions chromosomiques (ou duplications segmentaires). Plusieurs études ont essayé de quantifier le nombre de duplications de génome entier ayant eu lieu dans la lignée des vertébrés. Toutes s'accordent sur l'existence d'au moins une duplication globale du génome il y a environ 500 millions d'années (Blomme *et al*., 2006). Les gènes en double copie issus de duplication globale de génome sont parfois cités sous le terme de "ohnologues" en référence au généticien Susumu Ohno qui fut le premier à postuler que les duplications de gènes jouaient un rôle majeur dans l'évolution des espèces (Ohno, 1970). L'exemple frappant du séquençage du génome de la paramécie (*Paramecium tetraurelia*) a montré, par l'analyse de ses paralogues, que son génome contenait plus de 40 000 gènes résultant de trois duplications successives et globales de son génome (Aury *et al*., 2006).

Contrairement aux duplications de génomes, les duplications segmentaires concernent seulement des portions du génome et peuvent être distinguées si elles sont en tandem (la portion dupliquée est copiée dans le même sens que l'original) ou inverse (la copie est dans le sens opposé à la portion originale). Les duplications de régions géniques entraînent la formation de grande famille de gènes paralogues (gènes ribosomaux, récepteurs olfactifs, cytochromes, globines...) et, en conséquence, compliquent l'identification des relations d'homologie entre espèces. Chez l'homme, environ 200 grandes régions dupliquées (>1Mb) de plus de 90% d'identité de séquence sont recensées représentant près de 5% du génome humain et leur fréquence d'apparition est de ~0,01 par gène et par millions d'années (Bailey *et al*., 2002). À l'instar du taux de réarrangements, le taux de gènes dupliqués est variable entre espèces. Goodstadt *et al*., en étudiant les relations d'orthologie et de paralogie homme/chien, ont récemment montré que le génome humain contient trois fois plus de gènes dupliqués (in-paralogues) que celui du chien (1 550 vs. 540 respectivement) (Goodstadt *et al*., 2006). Plus généralement, les premiers séquençages des génomes de mammifères (homme, souris, chien, oppossum...) ont démontré que les gènes dupliqués spécifiques de lignées appartiennent à des catégories fonctionnelles bien définies (Mouse Genome Sequencing Consortium, 2002 ; Dog genome Consortium, 2005, Mikkelsen *et al*., 2007, Goodstadt *et al*., 2007). Celles-ci incluent la perception de molécules chimiques ou chémosensation (récepteurs olfactifs, gustatifs et récepteurs aux phéromones), la dégradation de toxines (la famille des cytochromes P450), la reproduction (les facteurs de croissance placentaire), l'immunité et la défense de l'hôte (récepteurs des lymphocytes T, immunoglobulines).

2.4.1. L'évolution des gènes dupliqués

La majorité des gènes dupliqués est perdue au cours de l'évolution du fait de la redondance fonctionnelle à priori non-avantageuse qu'ils représentent pour la cellule. En effet, des mutations délétères non-synonymes vont progressivement se fixer sur une des deux copies du gène aboutissant à la formation d'un pseudogène, on parle alors d'un processus de non-fonctionalisation ou de pseudogénéisation (Lynch *et al.*, 2000) (**Fig.9**). Le processus de pseudogénéisation est très fréquent dans les génomes et a abouti à un champ distinct d'étude (le pseudogénome). On recense plus de 19 000 pseudogènes dans le génome humain (www.pseudogene.org), soit une quantité légèrement

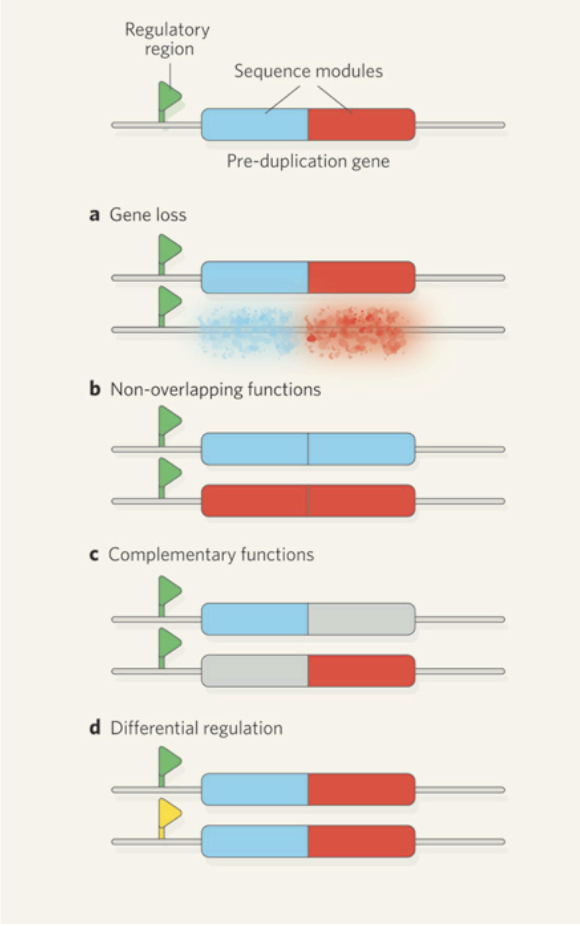


Figure 9 : Les destins des gènes dupliqués (d'après Louis *et al.*, 2007).
Soit un gène avec deux modules fonctionnels (en bleu et rouge).
a) Le plus fréquemment, il y aura une perte d'une des copies.
b) Les deux copies divergent et acquièrent chacune une fonction distincte.
c) Chacun des motifs fonctionnels divergent dans les gènes dupliqués mais les deux gènes assurent une complémentarité des deux fonctions initiales.
d) Enfin, le site de régulation diverge pour une des copies entraînant des modifications de l'expression du gène.

inférieure au nombre de gènes fonctionnels (~22 000).

Cependant, plus rarement, des cas de sous-fonctionalisation des gènes dupliqués ont été observés. Par exemple, la duplication d'un gène ancestral possédant deux domaines fonctionnels distincts peut entraîner une perte réciproque de chacun des domaines ou, autrement dit, la répartition de chacun des domaines fonctionnels dans les deux copies résultantes (Lynch *et al.*, 2000). D'autre cas

de sous-fonctionalisation concerne l'expression différentielle des deux copies du gène ancestral au niveau de deux tissus distincts. Aussi, le gène MITF (microphthalmia-associated transcription factor) est présent en un unique exemplaire chez l'homme et les oiseaux mais est dupliqué chez le poisson téléostéen (MITF-M et MITF-B). Chez le poisson, ces deux paralogues ont des structures exon-intron modifiées et des expressions tissulaires bien distinctes (Altschmied *et al*., 2002).

Enfin, encore plus rarement, une des copies de gène peut acquérir une nouvelle fonction alors que l'autre copie conservera la fonction originelle, on parle de processus de néo-fonctionalisation. Un exemple récent est celui des récepteurs aux acides rétinoïques (Escriva *et al*., 2006). Les mammifères disposent de trois paralogues RAR alpha, beta et gamma. En comparant les séquences des trois paralogues avec le génome du céphalochordé *amphioxus*, les auteurs ont pu montrer que RAR beta présentait le plus de similitudes fonctionnelles et d'expression avec le gène ancestral alors que RAR alpha et gamma avaient évolué par l'acquisition de nouveaux sites de liaison à l'ADN et de nouveaux profils d'expression.

2.4.2. Les deux modèles d'évolution des familles de gènes

Les duplications de gènes aboutissent à la formation de grandes familles multigéniques dont le mode d'évolution, non plus des gènes pris individuellement, mais de la famille prise dans son ensemble a fait l'objet de nombreux travaux. Pour une espèce considérée, certaines familles présentent une forte similitude de séquences des gènes dupliqués qui les composent. Ce modèle d'évolution suggère que des processus de conversion génique (copie d'un brin d'ADN sur un autre) agiraient sur ces familles de gènes. Ce mécanisme "d'évolution concertée" entraîne alors que les membres de chaque famille n'évoluent pas indépendamment les uns des autres (Novozhilov *et al*., 2006). L'exemple classique est celui de la famille des gènes codant pour les ARN ribosomaux pour laquelle les centaines de paralogues qui en sont issus présentent des séquences presque toutes identiques. Les gènes de la famille des ARNr de l'homme sont plus similaires entre eux qu'avec un des gènes de la famille des ARNr chez le chimpanzé. Ce fort degré d'homogénéité est donc favorisé par des mécanismes de conversion génique liés à une recombinaison non réciproque au moment de la méiose (Galtier, 2003).

Cependant, toutes les familles de gènes ne présentent pas ce profil d'évolution et suggèrent que leur histoire évolutive ne réponde pas au modèle d'évolution concertée. Par exemple, les gènes de la famille du CMH (Complexe Majeur d'Histocompatibilité) et des immunoglobulines ne présentent

pas plus de similitudes de séquences entre eux qu'avec un autre membre d'une espèce voisine. Les analyses phylogénétiques réalisées sur ces familles ont montré des profils évolutifs différents et ont amené Nei *et al.*, à proposer un deuxième modèle d'évolution des familles dit de "naissance et mort" de gènes (birth and death model) (Nei *et al*., 1997 ; Nei *et al*., 2005). Dans ce modèle, les gènes paralogues "naissent" par les différents mécanismes de duplications segmentaires, certains deviennent pseudogènes jusqu'à accumuler tellement de mutations qu'ils n'ont plus la possibilité d'être identifiés par des alignements de séquence.

Ainsi, les séquences dupliquées, les gènes et les familles de séquences répétées sont souvent associés à de larges réarrangements chromosomiques et favorisent les mécanismes de recombinaisons homologues non-alléliques dans les régions dupliquées. Chez la levure, les réarrangements chromosomiques sont fréquemment liés à la présence de séquences répétées de type LTR (Long Terminal Repeat) ou d'éléments transposables tels que l'élément Ty (Fischer *et al*., 2000). De la même façon, la couverture complète de la séquence du chromosome 19 humain a permis de mettre en évidence que 10/15 ruptures de synténie avec le génome de la souris se localisaient au niveau de clusters de gènes dupliqués (Eichler *et al*. 2003 ; Coghlan *et al*., 2005). Cependant, la relation de cause à effet n'est pas tout à fait claire et il est encore difficile de déterminer si ces réarrangements chromosomiques sont la cause ou bien la conséquence de la présence de régions dupliquées du génome. Plus généralement, les duplications segmentaires et les ruptures de synténie traduisent une instabilité de la région chromosomique considérée.

Cette instabilité de certaines portions des génomes a amené plusieurs travaux visant à comparer les génomes pour essayer de localiser les régions de cassures chromosomiques et prédire l'évolution des chromosomes.

2.5. Les deux modèles d'évolution des chromosomes

En 1984, les premières comparaisons des génomes de l'homme et de la souris par cartographie génétique avaient permis d'identifier environ 180 segments conservés de tailles variables (Nadeau *et al*., 1984). Les auteurs proposèrent alors que les ruptures de synténie qui déterminent la taille des segments conservés étaient aléatoirement réparties dans les génomes en accord avec le "modèle de

cassure aléatoire des chromosomes" ou "Random Breakage Model" (RBM) proposé par Susumu Ohno dans les années 1970. Ce modèle prévoit que la distribution de la taille des blocs suit une loi exponentielle ayant une fonction de densité $f(x)=1/Le^{-x/L}$ ou L est longueur moyenne de tous les segments. (**Fig.10a**).

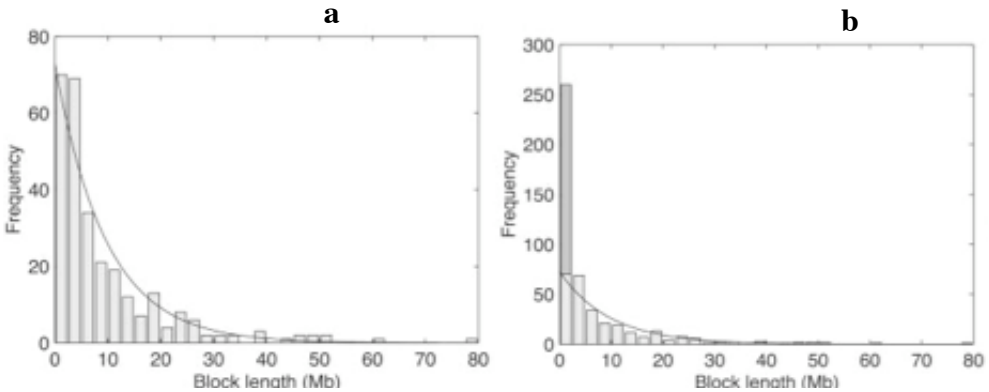


Figure 10 : Distribution des blocs de synténie en fonction de leur taille.
(d'après Pevzner *et al.*, 2003)
a) Dans le modèle de cassure aléatoire des chromosomes, la distribution des blocs en fonction de leur taille suit une loi exponentielle.
b) Le nombre de blocs de petites tailles est beaucoup plus important et la distribution des blocs en fonction de leur taille ne suit plus une loi exponentielle. C'est le modèle des points de cassures fragiles.

Cependant, le séquençage complet des génomes de l'homme et de la souris a ouvert la voie à l'analyse plus fine de l'identification des segments conservés et de leurs limites. Pavel Pevzner et Glenn Tesler (Pevzner *et al.*, 2003) ont ainsi défini près de 560 000 ancres de synténie (séquences uniques et très conservées entre les génomes) en alignant les génomes complets des deux mammifères. Ces ancres ont servi de support à l'identification d'un nombre beaucoup plus important de segments conservés supérieurs à 1Mb (258) dont la plupart correspond à des micro-réarrangements non détectables par cartographie. Ainsi, 40% des segments conservés exempts de tout réarrangement depuis l'ancêtre commun aux deux organismes ont une taille inférieure à 1 Mb. De plus, la distribution de la taille de ces nouveaux blocs ne correspondait plus à une loi exponentielle (**Fig.10b**), un nouveau modèle d'évolution des chromosomes a été proposé : le modèle des points de cassures fragiles ("Fragile Breakage Model"). Ce modèle implique que de nombreux de points de cassure, délimitant les blocs de synténie, ont été réutilisés au cours de l'évolution. Autrement dit, toute transformation d'un chromosome de l'ordre des gènes d'un génome vers l'ordre des gènes d'un second génome requiert un nombre signficatif de réutilisation des points de cassures.

Ces points de cassure peuvent être identifiés quand ils sont partagés par plusieurs espèces. En comparant les génomes de huit espèces distinctes au niveau phylogénétique (homme, cheval, chat, chien, cochon, bovin, rat et souris) appartenant à cinq ordres mammifères, Murphy *et al.*, ont mis en évidence que 20% des points de cassure étaient réutilisés au cours de l'évolution des mammifères (Murphy *et al.*, 2005). Cela suggère que des forces évolutives indépendantes des événements de réarrangements chromosomiques ont concerné des mêmes régions "fragiles" pour différentes lignées mammifères. De plus, les auteurs se sont intéressés à la distribution et la localisation des réarrangements chromosomiques associés à des cancers chez l'homme. Alors qu'aucune cassure chromosomique associée à un cancer n'est située dans les trois plus grands segments conservés humains par exemple, la co-localisation (+/- 0,4Mb) des réarrangements évolutifs avec les aberrations chromosomiques est trois fois plus importante pour les cancers fréquemment associés à des cassures que pour des cancers peu associés à des cassures. Enfin, en calculant la teneur en gènes autour des ruptures de synténie (1Mb centré autour du point de cassure), les auteurs ont observé que ces régions sont significativement plus denses en gènes que des régions prises aléatoirement dans le génome. Cette densification en gènes est à mettre en relation avec les fréquentes duplications segmentaires au niveau des points de cassure chromosomique (exemple du CMH). Par exemple, 95% des ruptures de synténie spécifiques des primates sont associées avec des duplications génomiques.

3. Le génome du chien de la cartographie à l'ère post-génomique

3.1. L'histoire évolutive de l'espèce canine

Comment expliquer les si grandes différences de taille, de poids, de morphologie et de comportement observées entre les presque 400 races de chien ? L'histoire évolutive de l'espèce canine et les méthodes de sélection des races par l'homme peuvent nous renseigner sur l'énorme diversité phénotypique qui caractérise le "meilleur ami de l'homme".

Le chien domestique, *Canis familiaris*, mammifère de l'ordre des carnivores appartient à la famille des *Canidae*. Cette famille regroupe 34 espèces (renard, coyote, dingo, loup, chien...) toutes issues d'un ancêtre commun carnivore il y a 10 millions d'années. Plusieurs travaux ont essayé de dater et de localiser le processus de domestication du chien. En comparant l'ADN mitochondrial de

plusieurs espèces du genre *Canis*, il semblerait que quelques loups gris (femelles) localisés dans l'Est asiatique seraient les ancêtres directs des chiens domestiques (Vila *et al.*, 2002 ; Leonard *et al.*, 2002). Cependant, des travaux plus récents tendraient à nuancer le faible nombre de fondateurs à l'origine du chien. 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.*, (Vila *et al.*, 2005) proposent qu'un grand nombre de croisements (intentionnels ou accidentels) entre les populations sauvages et domestiquées auraient eu lieu. Cette hypothèse permettrait aussi d'expliquer la grande diversité phénotyique observée actuellement chez les espèces domestiquées. Les analyses ADN d'un plus grand nombre d'espèces (coyote, chacal, loup...) pourront nous renseigner de manière plus fiable sur l'origine géographique du chien et la contribution des autres ces espèces à son avènement.

La datation du processus de domestication du chien par des recherches complémentaires en archéologie et en analyses moléculaires de séquences s'oriente sur un consensus autour de la période de 14 000 ans avant J.C (Savolainen *et al.*, 2002 ; Leonard *et al.*, 2002).

Cette domestication a eu un impact considérable sur le mode de vie des chiens, comparé à leurs ancêtres les loups gris. Une étude récente de Björnerfeldt *et al.*, a mis en évidence un mécanisme de relaxation des contraintes sélectives suivant la période de domestication du chien (Björnerfeldt *et al.*, 2006). En effet, les auteurs ont montré que le taux d'accumulation des substitutions non synonymes dans les gènes mitochondriaux était significativement plus rapide chez le chien que chez les loups. Si le relâchement des forces sélectives a aussi agi sur le reste du génome (nucléaire), cela pourrait, en partie, corréler avec la fascinante diversité phénotypique observée chez le chien.

Le processus de domestication correspond au 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'allèle. Cependant, un second goulet d'étranglement est intervenu plus récemment dans l'histoire (XIXème siècle) et coïncide avec la sélection et la création par l'homme de nombreuses races canines modernes (**Fig.11**). Ces pratiques de sélection intensive avaient pour finalité de disposer de

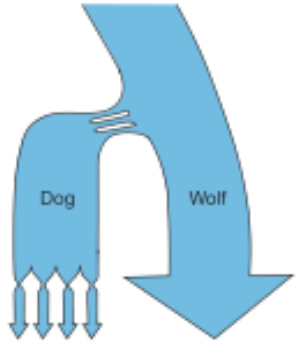


Figure 11 : Représentation schématique des deux goulets d'étranglement de l'histoire du chien.
(d'après Lindblad-Toh *et al.*, 2005)
Le premier goulet correspond à la domestication du loup par l'homme puis, plus récemment, la création des races modernes.

chiens possédant des aptitudes particulières (chiens de chasse, de garde, de traîneaux), des critères physiques, comportementaux ou esthétiques désirés. Cependant, la création des races, à partir d'un nombre limité de fondateurs ("étalons") et donc d'allèles a engendré une diminution de la variabilité génétique, une forte consanguinité et par conséquent, un taux d'homozygotie important au sein de chaque race canine.

C'est ainsi que l'on peut considérer chacune des races canines comme autant d'isolats génétiques à l'instar des populations humaines isolées géographiquement (islandaise par exemple et îlienne en général) ou culturellement (populations Amish ou Ashkénaze). De plus, la forte prévalence des maladies génétiques chez le chien, dont la plupart sont homologues des maladies humaines, et la structuration particulière de sa population font du chien un excellent modèle pour déterminer les relations génotype/phénotype. Ceci a poussé notre groupe à développer des outils de connaissance et d'analyse de son génome depuis plus d'une dizaine d'années afin d'identifier des gènes et des allèles, responsables ou de susceptibilité à des maladies génétiques simples ou complexes. Ces corrélations phénotype/génotype, bien difficiles à faire chez l'homme compte tenu du fort brassage allélique de notre population, sont plus facilement faisables chez le chien, par l'extrême stratification de sa population. Au delà de l'intérêt appliqué en médecine humaine et vétérinaire, les résultats de ces recherches permettent l'annotation fonctionnelle des génomes, par la mise en évidence de gènes nouvellement impliqués dans des processus physiologiques (Fondon *et al* ., 2004 ; Sutter *et al*., 2007) ou pathologiques (narcolepsie : Mignot *et al*., 1994, APR Atrophie Progressive de la Rétine : Guyon *et al*., 2007).

3.2. Les outils de connaissance du génome canin

3.2.1. Les cartes du génome

Dès 1995, des cartes génétiques puis d'hybrides d'irradiation ont cherché à positionner et ordonner des marqueurs sur les 39 paires de chromosomes du génome canin (2n=78) dont la plupart sont de petites tailles et acrocentriques (sauf les chromosomes X et Y). Ces caractéristiques ont empêché la standardisation du caryotype jusqu'à l'utilisation des techniques d'hybridations fluorescentes in situ (FISH) (Breen *et al*., 1999)

À la même époque, deux projets, menés parallèlement, aux États-Unis (par le Dr Elaine Ostrander) et en Europe (consortium DogMap), avaient pour but de développer des cartes génétiques aboutissant à la localisation sur les chromosomes canins de 150 et 94 marqueurs, respectivement.

En juin 2006, la carte génétique du génome du chien (v1.3) s'est considérablement densifiée comportant près de 2 400 marqueurs espacés en moyenne de 1,7 cM (http://www.vgl.ucdavis.edu/research/canine/people/Neff.html).

Dès 1998, notre laboratoire a initié un projet de cartographie du génome canin par la méthode des hybrides d'irradiation permettant de positionner environ 400 marqueurs (Priat *et al.*, 1999). À cette époque, le chien est, après la souris, le deuxième "modèle animal" à disposer d'une carte d'hybrides irradiés. En 2003, une nouvelle carte est produite et positionne plus 3 000 marqueurs correspondant à environ un marqueur tous les mégabases et assurant une couverture de 95% de son génome.
Les étroites collaborations établies entre les différents laboratoires "cartographes" du génome du chien ont permis en 2000 de regrouper les données de carte génétique et RH (Mellersh *et al.*, 2000). De la même façon, Breen *et al.*, ont rassemblé conjointement un ensemble de marqueurs issus des trois méthodes de cartographie afin de disposer d'une carte intégrée du génome du chien contenant 302 marqueurs FISH, 354 marqueurs de liaison génétique et 1 500 marqueurs positionnés en RH (Breen *et al.*, 2001). Par la suite, un sous ensemble de 804 BACs a aussi été cartographié par l'équipe de Matthew Breen *et al.*,. renforçant la complémentarité et l'apport d'informations des cartes intégrées RH/FISH (Breen *et al.*, 2004).
La densité en marqueurs le long des chromosomes du génome du chien a permis, dès 2003, de réaliser les premières études comparées avec le génome humain par la construction de cartes de synténie entre le bras court du chromosome 1 humain (HSA1p) et le génome du chien (travail auquel j'ai participé dans le cadre de mon stage de maîtrise) (Guyon *et al.*, 2003). En 2005, l'utilisation des données de séquençage léger du génome canin dans la réalisation d'une carte RH de haute densité a permis de répondre au trois objectifs du projet, (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) un support à l'assemblage du séquençage profond (7.5X) du génome canin (Hitte *et al.*, 2005). Mon implication dans ce travail sera présentée dans le deuxième chapitre de la partie Résultats.

3.2.2. Le séquençage léger

Dès 2003, le chien est le premier organisme a avoir bénéficié d'un séquençage léger de son génome. Le génome d'un caniche mâle a été séquencé à une profondeur de 1,5X ce qui signifie que chaque nucléotide est séquencé en moyenne 1,5 fois (Kirkness *et al.*, 2003). Les apports de ce type de

projet sont nombreux et ont 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 entre le chien et la souris. 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). Enfin, ce projet a mis en évidence près de 18 500 fragments de gènes orthologues avec les gènes humains.

Afin de ne pas limiter la connaissance du catalogue génique du chien à de courts fragments de gènes, la communauté canine internationale, et plus particulièrement le NIH, lance l'initiative d'un séquençage complet du génome du chien en 2004. Ce projet a été mené par le BROAD Institute de Cambridge (USA) à partir du génome d'un chien femelle de race boxer. En effet, le choix de la race boxer résulte de travaux préalables (Parker *et al.*, 2004) ayant 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.
La séquence du génome du chien présentait, en outre, un intérêt évolutif important compte tenu de sa position phylogénétique dans l'arbre des mammifères. En effet, seuls quatre génomes étaient disponibles à cette époque (homme, chimpanzé, souris et rat), tous regroupés sous une des quatre clades (Euarchontoglires) qui composent les mammifères. Le chien, appartenant à la clade des Laurasiatheria, représentait le meilleur "groupe frère" pour analyser les Euarchontoglires.

3.2.3. Le séquençage profond (7.5X) et caractéristiques du génome canin

3.2.3.1. La séquence

Le séquençage du génome du chien a engendré 36 millions de séquences qui ont, ensuite, été assemblées par le programme ARACHNE (Batzoglou *et al.*, 2002). 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 vides de séquence ("gaps").Le processus d'assemblage a été facilité par la mise à disposition

des cartes du génomes. En effet, 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.

3.2.3.2. Les éléments répétés

La taille du génome canin est sensiblement inférieure à celui de l'homme. Ceci s'explique, en partie, par la plus faible présence de séquences répétées spécifiques à la lignée canine (334 Mb vs 609 Mb pour l'homme).

Cependant, il existe un cas particulier d'éléments répétés lié à la famille des SINEC_Cf (Short INtersperced Element). Cette famille de SINEs, correspondant à des éléments rétrotransposables d'environ 200 pb, est spécifiquement présente dans le génome du chien. Wang et Kirkness (2005) ont estimé que 50% des gènes canins annotés contiennent au moins un SINEC_Cf dans leurs séquences introniques (5' ou 3') et leur comparaison avec la séquence du caniche a mis en évidence plus de 10 000 locus polymorphes (absent/présent) pour cette famille de séquences répétées (Wang *et al.*, 2005). À titre de comparaison, le nombre d'insertions bimorphiques de SINE chez l'homme est inférieure à 1 000 (les séquences *alu* représentant la grande famille de séquences répétées humaines). L'insertion d'un SINE soit dans un exon, soit dans une séquence régulatrice peut considérablement modifier le transcrit ou la régulation du (ou des) gène(s), respectivement. De la même façon, l'insertion anti-sens d'un SINE dans un intron peut engendrer une modification du mécanisme d'épissage en créant un nouveau site accepteur (Mignot *et al.*, 1994 ; Hédan *et al.*, 2007 ; Clark *et al.*, 2007). Il est alors envisageable que cette famille de séquences répétées ait joué un rôle important dans la plasticité et la dynamique de son génome.

3.2.3.3. Les cartes de synténie

En se basant sur la méthode d'identification des blocs de synténie développée par Pevzner *et al.* (Pevzner *et al.*, 2003), les auteurs ont construit des cartes de synténie multi-espèces comprenant les génomes du chien, de l'homme, de la souris et du rat. Tous les segments conservés (reflétant les réarrangements inter-chromosomiques) et les segments conservés ordonnés (correspondant aux réarrangements intra-chromosomiques) d'une taille supérieure à 0,5 Mb ont été identifiés. Ainsi, 183 CSs et CSOs sont synténiques entre les génomes canin et humain alors que plus de 260 CSs-CSOs sont recensés entre les génomes humains et murins. Ces résultats confirment le fort taux de

réarrangements spécifiques de la lignée des rongeurs et, préalablement, mis en évidence lors du séquençage du génome de la souris (Waterston *et al.*, 2002). De manière intéressante, le nombre de cassures intra-chromosomiques est plus faible chez l'homme (83) comparé aux autres génomes chien, rat et souris (respectivement 100, 161 or 176). Les régions de cassures chromosomiques sont significativement associées à une composition de séquence particulière avec un fort taux de divergence nucléotidique (comparé à des séquences ancestrales) et un fort taux de G+C .

3.2.3.4. Annotation du génome du chien

L'annotation en gènes canins a été, principalement, produite à partir du système standard développé par Ensembl. Le nombre de gènes codant pour des protéines chez le chien est inférieur à celui de l'homme et de la souris (19 314 vs 22 810, 24 442 respectivement) (Ensembl v.42). De plus, le nombre d'ESTs canins (~350 000) recensés dans la base de données DB_EST est très inférieur à l'homme (8 000 000) et à la souris (6 000 000) 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...).

L'annotation des gènes canins a aussi été analysée par le programme PhyOP (Goodstadt *et al.*, 2006) qui a permis d'affiner l'identification de ~14 500 gènes canins en relation d'orthologie 1:1 avec le génome humain. Un sous-ensemble de 13 816 gènes humains, murins et canins en relation d'orthologie 1:1:1 a servi de support à l'étude des forces sélectives agissant sur les répertoires géniques des trois espèces. Classiquement, le dn/ds ou (Ka/Ks) est utilisé pour mesurer les forces sélectives agissant sur les gènes et correspond au rapport du taux de substitutions non synonymes par sites non-synonymes sur le taux de substitution synonymes par sites synonymes. Un dn/ds ~ 1 reflète une évolution neutre, un dn/ds << 1 correspondra à une sélection négative alors qu'un dn/ds > 1 mettra en évidence une sélection positive. La distribution de ce taux varie sensiblement entre les trois espèces et renforce les prédictions des théories de génétique des populations qui suggèrent que plus la taille effective de la population est grande plus les forces de sélection négative sont importantes.

Plusieurs hypothèses peuvent alors être soulevées pour expliquer le faible nombre de gènes canins annotés relativement aux autres génomes mammifères. Est-ce lié à une annotation surestimée des génomes humain et rongeur ou plus simplement un défaut de séquence chez le chien ? Ou bien, les gènes non retrouvés chez le chien auraient-ils tellement divergé au niveau nucléotidique que les seuils de similarité seraient trop stricts pour les retrouver ? Ou enfin, y a-t-il eu des duplications de gènes spécifiques à la lignée rongeurs-primates qui expliqueraient les absences de relation

d'orthologie 1:1 avec le chien ? Le travail présenté dans la partie Résultats s'efforce de répondre à ces questions en exposant les travaux réalisés au cours de ces trois années de thèse.

II. RÉSULTATS

Ces trois dernières années, le modèle génétique canin a bénéficié de l'apport de cartes génomiques de plus en plus denses (Guyon *et al.*, 2003 ; Breen *et al.*, 2004 ; Hitte *et al.*, 2005) et de deux projets de séquençage, un léger (1.5x) (Kirkness *et al.*, 2003) et un profond (7.5x) (Lindblad-Toh *et al.*, 2005) qui ont produit des ressources majeures. Intégrer ces différentes ressources puis les comparer avec les données d'autres organismes sont des objectifs importants pour une exploitation optimale de l'information génétique. Dans ce contexte, les résultats de mon travail de thèse sont présentés en trois parties.

Le premier chapitre expose l'outil AutoGRAPH, un serveur interactif de comparaison et d'intégration de données de plusieurs ressources et/ou espèces, dont nous en détaillerons le développement et le formalisme. Pour illustrer une première application, nous présentons l'utilisation d'AutoGRAPH dans le cadre de la comparaison de la cartographie RH de 10 000 marqueurs de gènes avec la séquence du génome canin.

Dans la deuxième partie des résultats, nous présentons l'apport d'AutoGRAPH dans le projet d'annotation des gènes orphelins du répertoire canin. Notre approche a consisté à formaliser l'identification complète des segments conservés entre plusieurs génomes, puis à utiliser la conservation de l'ordre des gènes au sein des segments conservés pour déterminer des intervalles orthologues et enfin, à réaliser des analyses de prédiction de gènes par alignements de séquences ciblés sur les intervalles orthologues canins. Dans cette deuxième partie, deux projets distincts sont détaillés : l'annotation du chromosome canin non assemblé ("chromosome Unknown") et la recherche de gènes à priori absents chez le chien mais présents chez plusieurs espèces de référence (homme-chimpanzé-souris-rat).

La troisième partie des résultats résume un travail de collaboration entrepris au cours de ce projet de thèse.

1. De la carte à la séquence du chien : l'apport de la génomique comparée

1.1. AutoGRAPH un serveur versatile de comparaison de génomes

(Article 1 : Derrien et al., 2007)

Depuis le séquençage du génome de la bactérie *Heamophilus Influenzae* en 1995, (Fleischmann *et al*., 1995) le nombre d'organismes cellulaires ayant bénéficié d'un séquençage complet a connu une croissance exponentielle pour atteindre près de 664 en octobre 2007 (**Fig.12**) (www.genomesonline.org). Par ailleurs, un même organisme dispose souvent de plusieurs ressources (cartes génétiques, RH, physique, séquences partielles ou séquençage complet...) qui viennent initier voire compléter la connaissance et la caractérisation de son génome. Généralement, ces ressources sont comparées à celles "d'espèces de références" afin de déduire des informations sur l'organisme d'intérêt, d'identifier les relations de synténie entre plusieurs génomes et d'inférer les mécanismes évolutifs et l'organisation fonctionnelle du génome d'intérêt. Au sein d'une espèce, la démarche de cartographie comparée va chercher à intégrer les différentes ressources du génome d'intérêt en comparant, par exemple, la carte cytogénétique avec la carte RH ou bien la carte RH avec la (ou les) séquence(s) disponibles. L'intégration des données intra-espèces permet d'évaluer la robustesse et la qualité des ressources pour un génome donné. Tous ces projets génèrent des quantités de données importantes nécessitant la mise en oeuvre de moyens informatiques et statistiques pour les analyser.

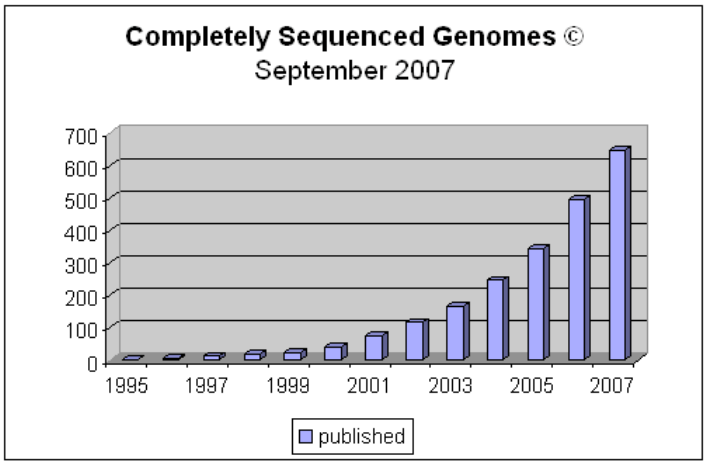


Figure 12 : Croissance exponentielle du nombre d'organismes séquencés dans les douze dernières années.

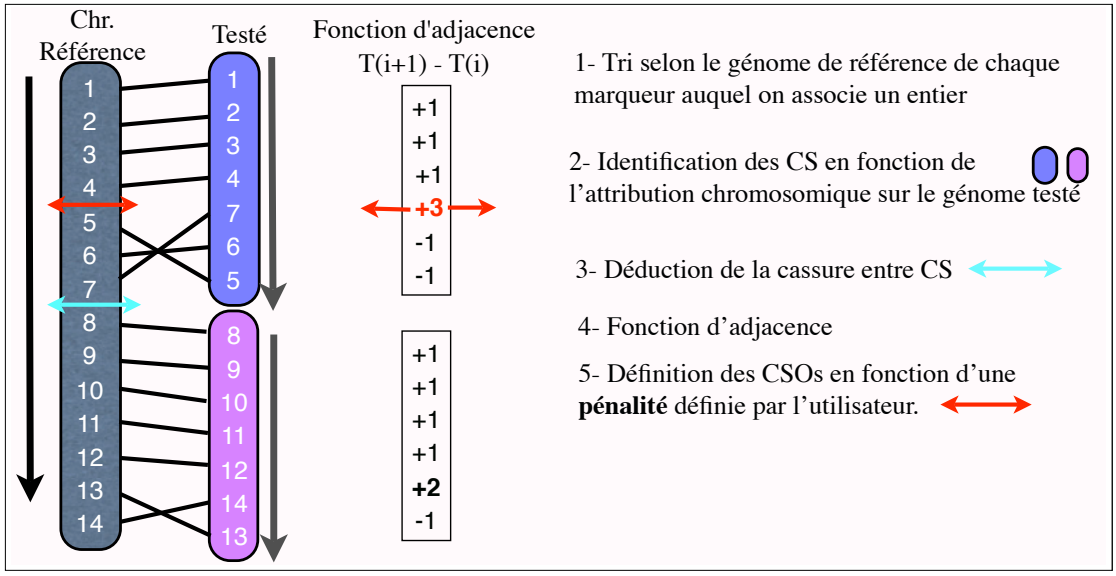
Nous avons développé un programme interactif, AutoGRAPH (Derrien *et al*., 2007), dont le triple objectif était de formaliser et d'automatiser la construction de cartes de synténie entre deux ou trois espèces, d'identifier les segments conservés, les segments conservés ordonnés et les ruptures de

synténie entre espèces et enfin de disposer d'une certaine interactivité permettant à l'utilisateur de comparer son propre jeu de données (cartes de synténie et/ou cartes intégrant plusieurs ressources).

Comparer des ensembles de données nécessite de disposer de points de comparaison (ancres) pour établir des correspondances entre ces données. Ainsi un même marqueur cartographié par deux techniques différentes constitue une ancre de comparaison pour l'intégration des deux ressources. De la même façon, les gènes en relation d'orthologie 1:1 représentent des ancres de choix pour la construction de cartes de synténie entre espèces.

Le programme AutoGRAPH dispose de deux modules, l'un utilisant des données pré-insérées issues de l'annotation par Ensembl (v.42) (Curwen *et al.*, 2004) de six espèces mammifères (homme, chimpanzé, chien, vache, souris et rat) et un second module permet à l'utilisateur d'insérer un jeu de données personnel qui sera comparé et analysé. Le format de données utilisable par AutoGRAPH nécessite que chaque ancre de comparaison soit définie par trois paramètres : un identifiant (nom du gène, du marqueur...), une attribution chromosomique (HSA_3, scaffold 1234) et une position (en paire de bases, centiRay, centiMorgan...). Dans le cas de l'insertion d'un jeu de données fourni par l'utilisateur, le format de fichier GFF (General File Format) (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml) est géré par le programme.

La méthodologie générale du serveur est résumée dans la figure 12 (**Fig.13**) . La troisième étape, du



programme, correspond à la phase d'identification des segments conservés et à l'évaluation de la colinéarité entre les jeux de données. Cette étape est résumée dans la figure 14 (**Fig.14**).

Figure 14 : Formalisation algorithmique de la définition des CSs et CSOs.

Étape d'identification des segments conservés (CS) en fonction de leur assignation chromosomique sur le génome testé. Les segments conservés ordonnés (CSO) sont identifiés par une fonction d'adjacence qui va tester la colinéarité des ancres de comparaison au sein de chaque CS.

Premièrement, un tri des marqueurs du jeu de référence est effectué en fonction de leur position sur le chromosome de référence (gris). Les coordonnées génomiques du génome de référence sont transposées en série de valeurs croissantes permettant d'associer à chaque valeur son rang. Tous les marqueurs, identifiés par leurs rangs, sont alors groupés selon leur appartenance à un chromosome du génome testé. Cette première étape permet de définir les segments conservés (CS) et identifie les ruptures de synténie inter-chromosomiques. Dans un deuxième temps, les marqueurs sont triés au sein des CSs selon leurs valeurs de rangs. Une fonction d'adjacence va évaluer la colinéarité des marqueurs entre les génomes au sein de chaque segment conservé en comparant la contiguïté des valeurs de rangs entre un marqueur (i) et le marqueur suivant (i+1). Le résultat de cette fonction peut être :

- Test_Rang(i+1) - Test_Rang(i) = 1 : signifiant une contiguïté entre deux marqueurs pour les deux génomes.
- Test_Rang(i+1) - Test_Rang(i) = -1 : signifiant une contiguïté inversée entre deux marqueurs.
- Test_Rang(i+1) - Test_Rang(i) > 1 ou <-1 : signifiant une rupture de la colinéarité entre deux marqueurs. Cette rupture de synténie délimitera alors deux segments conservés ordonnés (CSOs).

La conception du programme AutoGRAPH permet de modifier les paramètres d'affichage des résultats (ordre croissant ou décroissant des données, fonctionnalité de zoom sur l'image) et de paramètres de définition des CSs/CSOs. Ainsi, la taille minimale d'un CSO (exprimée en nombre d'ancres) et la contrainte de colinéarité au sein des CS (pénalité de "gap") peuvent être fixées par l'utilisateur. Dans ce dernier cas, plus la pénalité de "gap" sera grande, plus l'utilisateur autorisera des micro-réarrangements au sein des CSs. Par exemple, une pénalité de valeur '4' permettra à un marqueur d'être en décalage de colinéarite de quatre marqueurs au maximum au sein de son CS. À l'inverse, une pénalité de "gap" de 2 (minimum autorisé) impliquera que toute inversion entre deux marqueurs entraîne la définition d'un CSO. En outre, les singletons, correspondant aux segments conservés représentés par un seul marqueur, sont mis en évidence par une couleur bleue.

Une fonctionnalité importante du programme consiste en la possibilité de choisir les types de relations d'orthologie à afficher avec le ou les génomes testés. Ainsi, l'utilisateur peut sélectionner les relations 1:0 ou 1:1 pour chacun des génomes testés. Dans le cas d'un affichage des relations 1:0 (présence du gène sur le chromosome référent et absence sur le génome testé), un intervalle

synténique orthologue est alors proposé en fonction de la colinéarité au sein des CSOs. Cette propriété du programme sera illustrée dans la deuxième partie des Résultats.

Les résultats sont présentés sous forme graphique exportable sous différents formats (pdf, png, jpeg...) ainsi que sous forme tabulaire texte. Le programme retourne un tableau de résultats pour chaque CS/CSO qui liste les informations de localisations sur les génomes référent et testé(s), le nombre et la densité de marqueurs au sein de CS/CSOs ainsi que la taille, la localisation et la densité en marqueurs au niveau des ruptures de synténie.
La sortie graphique est présentée dans la figure 15 (**Fig.15**) et illustre la construction d'une carte de synténie entre le chromosome 19 du chien (CFA19) et le génome humain.

Article n°1

AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps.

Derrien T, Andre C, Galibert F, Hitte C.
Bioinformatics 2007

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 4 2007, pages 498–499
doi:10.1093/bioinformatics/btl618

Genome analysis

AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps

Thomas Derrien, Catherine André, Francis Galibert and Christophe Hitte*

CNRS UMR6061 Génétique et Développement, Université de Rennes1, IFR140, 2 Av du Pr. Léon Bernard, CS 34317, 35043, France

Received on September 20, 2006; revised and accepted on November 29, 2006
Advance Access publication December 4, 2006
Associate Editor: Alfonso Valencia

ABSTRACT

AutoGRAPH is an interactive web server for automatic multi-species comparative genomics analyses based on personal datasets or pre-inserted public datasets. This program automatically identifies conserved segments (CS) and breakpoint regions, assesses the conservation of marker/gene order between organisms, constructs synteny maps for two to three species and generates high-quality, interactive displays facilitating the identification of chromosomal rearrangements. AutoGRAPH can also be used for the integration and comparison of several types of genomic resources (meiotic maps, radiation hybrid maps and genome sequences) for a single species, making AutoGRAPH a versatile tool for comparative genomics analysis.

Availability: http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/

Contact: hitte@univ-rennes1.fr

Supplementary information: A description of the algorithm and additional information are available at http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/Tutorial.php

1 INTRODUCTION

Many large-scale mapping and sequencing projects have been completed in the last 10 years, making it possible to compare the genomes of many species, to study evolutionary changes (Murphy *et al*., 2005), and to improve genome annotation (Chatterji and Pachter, 2006). Comparative genomics is based on the identification of unique, unambiguous orthologous sequences in different species. These sequences, known as comparative anchors (O'Brien *et al*., 1993), are used to determine the contiguity of anchors between genomes from different species and thus to investigate the correspondence of genomic segments and to define their limits.

Synteny maps can be used to identify conserved segments (CS), corresponding to chromosomal segments containing the same list of markers in all the species studied, and CS ordered (CSO), in which not only are the same markers present in all species, but they are also in the same order (Fig. 1). The genomic regions delimiting CS and/or CSO, the breakpoints, can also be identified by constructing synteny maps. Several tools have been developed for identifying CS and CSO (Pan *et al*., 2005; Pavesi *et al*., 2004; Halling-Brown *et al*., 2004; Clamp *et al*., 2003; Tesler, 2002). These tools allow accurate comparative genomics analyses, but are often dedicated to a specific genome, are not compatible with the use of large sets of personal data, and are not always available online. We have therefore developed AutoGRAPH, a versatile, online multi-species server for the automatic identification of CS, CSO, breakpoints, marker/gene order conservation, the integration of multiple resources for a given species (Hitte *et al*., 2005) (Supplementary Figure 2), and the generation of an interactive graphical display of comparative maps. AutoGRAPH frees the user from the tedious task of plotting comparative maps and determining marker order correspondence manually for personal data, enabling the user to focus on interpreting synteny or integrated maps.

*To whom correspondence should be addressed.

2 PRINCIPLES

Using the AutoGRAPH server for constructing synteny maps between species

Public datasets: information for the protein-coding genes (Ensembl v39) of six reference mammalian genomes (human, chimpanzee, cow, dog, mouse and rat) have been inserted into the program database. Orthologous gene pairs with one-to-one relationships (Ensembl v39) are used to construct pairwise or three-way synteny maps. Synteny maps can be built for selected chromosomes or for all the chromosomes of the reference genome. CS, CSO and breakpoints are automatically identified with the fast adjacency function. This function assigns relative integers to the coordinates provided and determines the sequence of integers in the same order for common anchors of the reference and tested genomes (a detailed description is provided in the Supplementary material). Various options are available and can be combined interactively to specify the way in which the synteny map should be constructed. For example, it is possible to specify the nature of the orthology between markers (one-to-one/one-to-zero) and the number of anchors defining CS and CSO. An adjacency penalty value, corresponding to the number of marker that account for an interruption of colinearity can be set by users to specify the marker order conservation criteria.

The program output consists of

- A figure showing gene order relationships by means of connecting lines within each CS and CSO, across two to three species, and identifying the evolutionary breakpoint regions delimiting CS and/or CSO (Fig. 1A). Ouput figure can be exported in several formats.
- Two modes, making it possible to construct comparative maps with or without one-to-zero relationship anchors, to switch the orientation of each map and to scale images by a factor of 0.5 to 2 (Fig. 1B). When mode 1:0 is selected, genes in the reference genome with no ortholog in the tested genome are placed on the synteny map. Their putative interval location in

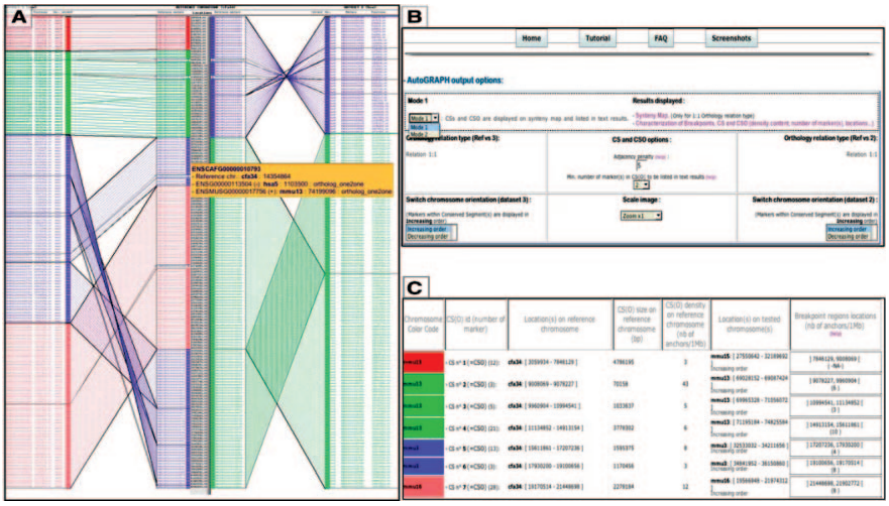


Fig. 1. Examples of AutoGRAPH output. **(A)** Graphical display of a three-way synteny map. Chromosome 34 from dog, the reference species, is represented in the middle of the figure. The bar on the right corresponds to the human genome and the bar on the left to the mouse genome. For each species, markers are identified by their ID and genomic coordinates. A dynamic link (orange box) provides a link to the Ensembl database for a complete gene description. The colored lines connecting orthologous genes show the conservation of gene order. Black lines between colored segments represent breakpoints between CS and/or CSO (see tutorial on the web server site). **(B)** This panel shows the options that can be set by users for the interactive definition of parameters for comparative genomics analysis. **(C)** For all analyses submitted, CS, CSO and breakpoints are identified and characterized in terms of their chromosomal limits, size and marker content, ID and density. All information can be downloaded in text file format.

the tested genome is inferred from the gene-order conservation rule and is automatically output as a tabulated text file.

- A table listing all CS, CSO and breakpoints, with their size, number of genes, chromosome coordinates and gene density (Fig. 1C).

Personal datasets: two formats of input datasets can be uploaded. GFF format and tabular plain-text columns comprising unique marker IDs, chromosome number and genomic coordinates for each dataset to be studied. Input datasets may contain different sets of markers, with missing or duplicated markers or genes. Personal datasets can be entered or uploaded via web form and example inputs are provided by the web interface.

For comparative studies between species, synteny maps can be built for entire chromosomes, specific parts of chromosome or the whole genome, as specified by the user. Datasets for a single species can also be used. Any type of coordinate-base pairs (bp) kilobases (kb) or megabases (Mb)-mapping data units, centimorgans (cM) or centirays (cR) can be used to indicate location in the genome. For example, (Supplementary Figure 2), the server can integrate high-density genetic/radiation hybrid/genome sequence maps for a single species, facilitating map construction and improving sequence assembly. Various options can be combined to specify comparative analyses. The output displays are the same as for public datasets.

3 SYSTEMS

AutoGRAPH runs on a Apache web server and uses a MySQL relational database. MySQL specific queries and Perl functions are used to apply options and the adjacency function. The web interface was developed in PHP language and the graphical display uses a GD graphic library written in C (http://www.boutell.com/gd/).

ACKNOWLEDGEMENTS

The authors would like to thank Denis Larkin and Simon de Givry for testing AutoGRAPH and providing useful suggestions. The authors also thank the GenOuest Bioinformatics Platform for hosting the web server, the French Centre National de la Recherche Scientifique (CNRS) for supporting this work and the Conseil Regional de Bretagne for supporting T.D. with a fellowship.

Conflict of Interest: none declared.

REFERENCES

Chatterji,S. and Pachter,L. (2006) Reference based annotation with GeneMapper. *Genome Biol.*, **7**, R29.

Clamp,M. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.

Halling-Brown,M. *et al.* (2004) A Fugu-Human Genome Synteny Viewer: web software for graphical display and annotation reports of synteny between *Fugu* genomic sequence and human genes. *Nucleic Acids Res.*, **32**, 2618–2622.

Hitte,C. *et al.* (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat. Rev. Genet.*, **8**, 643–648.

Murphy,W.J. *et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**, 613–617.

O'Brien,S.J. *et al.* (1993) Anchored reference loci for comparative genome mapping in mammals. *Nat. Genet.*, **2**, 103–112.

Pan,X. *et al.* (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461–3468.

Pavesi.G. *et al.* (2004) GeneSyn: a tool for detecting conserved gene order across genomes. *Bioinformatics*, **9**, 1472–1474.

Tesler,G. (2002) GRIMM: genome rearrangements web server. *Bioinformatics*, **18**, 492–493.

1.2. L'apport de la cartographie RH dans le séquençage léger d'un organisme

(Article 2 : Hitte et al., 2005)

En 2004, un consortium international, impliquant quatre laboratoires des Drs Elaine Ostrander - National Institute of Health (US), Ewen F Kirkness - The Institute for Genome Research (TIGR) (US), Panos Deloukas - Sanger Center (Uk) et Francis Galibert - UMR6061 (Fr), entreprend un projet de cartographie de 10 000 gènes canins par la méthode des hybrides d'irradiation. Ce projet s'est basé sur l'expertise du laboratoire dans la construction de cartes RH et fait suite au séquençage léger du génome du chien initié par le TIGR deux ans plus tôt. Ce séquençage, d'une couverture de 75% du génome, avait permis d'identifier plus de 18 000 fragments de gènes canins par alignements réciproques avec le génome humain. Néanmoins, l'information issue d'un tel projet de séquençage reste très fragmentaire (recouvrement de 1-2X) et l'étude de l'ordre des gènes et les analyses de synténie sont difficilement exploitables.

1.2.1. Objectifs - Résumé

Le positionnement de manière unique de 10 000 fragments de gènes sur le génome canin constituait l'objectif principal de ce projet. De plus, l'optimisation du séquençage léger, par la possibilité de construire des cartes de synténie avec le génome humain, représentait le second volet de l'étude. Enfin, parallèlement à ce projet de cartographie, un projet de séquençage profond (7,5X) du génome canin a été initié au Broad Institute de Cambridge (US) (Lindblad-Toh *et al.*, 2005). Aussi, la construction d'une carte RH dense permettait de fournir une trame pour la phase d'assemblage du séquençage profond du chien.

Dans ce projet, notre groupe à Rennes a réalisé (i) le panel d'hybride cellulaire irradié à 9 000 rads ayant la capacité de positionner plus de 10 000 marqueurs, (ii) une importante partie du génotypage (~5000 marqueurs), et enfin (iii) la phase de construction des 39 cartes RH des chromosomes canins et les analyses de synténie avec le génome humain.

Mon rôle dans ce projet s'est focalisé sur la conception et l'utilisation d'AutoGRAPH pour l'étude des relations de synténie entre les génomes canin et humain et dans le cadre des analyses des correspondances entre les différentes ressources (cartographie et séquences).

1.2.2. Résultats

- Cartes de synténie

Nous avons dérivé les 39 cartes de synténie chien/homme et identifié l'ensemble des segments conservés (~250) dont la taille varie de 2 à 332 marqueurs (**Fig.16**). La figure 16 représente la carte de synténie construite par AutoGRAPH entre le chromosome 1 canin (CFA 1) et le génome humain.

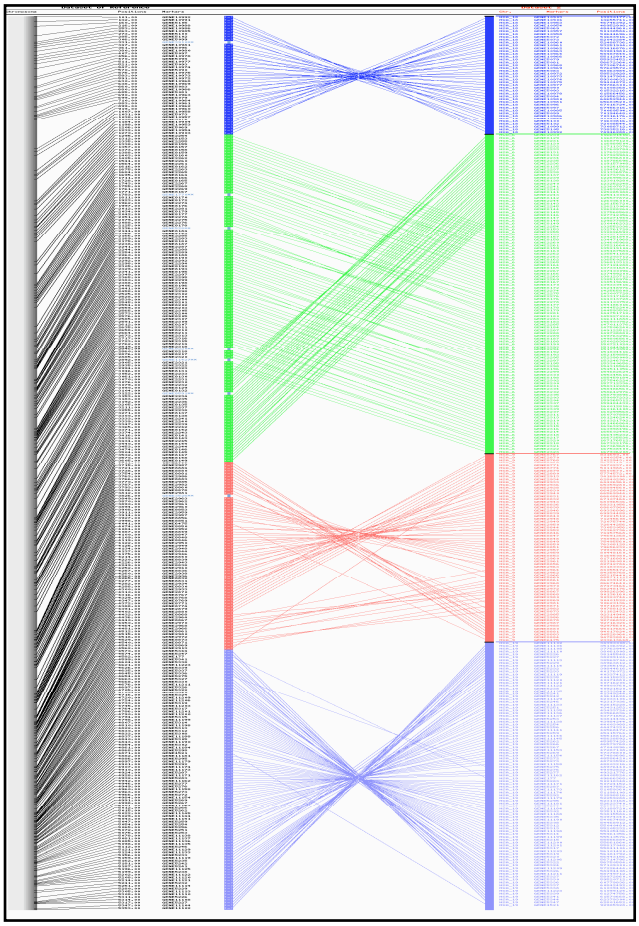


Fig.16 : Carte de synténie construite par le programme AutoGRAPH entre le chromosome 1 canin et le génome humain.
Les marqueurs ordonnés par la carte RH sont positionnés sur la gauche et les relations d'orthologie avec les séquences humains sont représentées par des traits horizontaux colorés en fonction de l'appartenance aux chromosomes humains. Les quatre segments conservés appartiennent aux chromosomes HSA_18 (bleu foncé), HSA_6 (vert), HSA_9 (rouge) et HSA_19 (bleu clair).

Six CSOs sont représentés appartenant à quatre CSs. Les marqueurs canins cartographiés en RH (listés à gauche et en bleu ciel) correspondent à des singletons c'est-à-dire des segments conservés identifiés par un seul marqueur. Ils sont donc à prendre avec précaution tant par l'analyse RH que par leur définition préalable à partir de l'alignement réciproque sur le génome humain.
La construction de ces cartes de synténie denses en gènes nécessite l'ordonnancement des 10 000 marqueurs le long des chromosomes canins par une approche RH et fournit une optimisation de l'information génétique apportée par le séquençage léger seul.

- Cartes inter-ressources avec le séquençage profond

Le serveur AutoGRAPH a été utilisé ici pour évaluer le niveau de correspondance entre plusieurs ressources. Cette fonctionnalité s'est avérée particulièrement intéressante dans le cadre de la comparaison de l'ordre des marqueurs entre les méthodes de cartographie et le séquençage 7.5x du génome canin. Sur la figure 17 est représentée la carte intégrée du CFA11 pour trois ressources (**Fig.17**).

Ces analyses ont permis de souligner les inversions entre carte(s) et séquence(s). Seules les inversions supérieures à 600 kb (>3 marqueurs) ont été examinées. Sur les 39 cartes "inter-ressources" canines construites, huit inversions ont été déterminées entre la cartographie RH et la séquence. Dans tous les cas, des expériences de cytogénétique ont été réalisées sur les régions problématiques et quatre (50%) ont été résolues en accord avec la cartographie RH alors que quatre autres correspondaient à des problèmes de construction de cartes RH. La résolution de ces artefacts de cartographie et d'assemblage de séquence a abouti a des versions finalisées de la carte RH (Hitte *et al*) et d'une version mise à jour de l'assemblage du génome canin (CanFam2, Lindblad-Toh *et al*., 2005).

Article n°2

Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping.

Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C, Evanno G, Pullar B, Cadieu E, Oza V, Lourgant K, Jaffe DB, Tacher S, Dreano S, Berkova N, Andre C, Deloukas P, Fraser C, Lindblad-Toh K, Ostrander EA, Galibert F.

Nat Rev Genet. 2005

Nature Reviews Genetics | AOP, published online 12 July 2005; doi:10.1038/nrg1658

PERSPECTIVES

OPINION

Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping

Christophe Hitte, Jennifer Madeoy*, Ewen F. Kirkness*, Catherine Priat, Travis D. Lorentzen, Fabrice Senger, Dan Thomas, Thomas Derrien, Christina Ramirez, Carol Scott, Gwenaelle Evanno, Barbara Pullar, Edouard Cadieu, Vinay Oza, Kristelle Lourgant, David B. Jaffe, Sandrine Tacher, Stéphane Dréano, Nadia Berkova, Catherine André, Panagiotis Deloukas, Claire Fraser, Kerstin Lindblad-Toh, Elaine A. Ostrander and Francis Galibert*

Abstract | Accurate and comprehensive sequence coverage for large genomes has been restricted to only a few species of specific interest. Lower sequence coverage (survey sequencing) of related species can yield a wealth of information about gene content and putative regulatory elements. But survey sequences lack long-range continuity and provide only a fragmented view of a genome. Here we show the usefulness of combining survey sequencing with dense radiation-hybrid (RH) maps for extracting maximum comparative genome information from model organisms. Based on results from the canine system, we propose that from now on all low-pass sequencing projects should be accompanied by a dense, gene-based RH map-construction effort to extract maximum information from the genome with a marginal extra cost.

Almost all genome sequencing involves the assembly of overlapping sequence reads that are generated from random fragments of a genome. This approach requires that most bases are sequenced redundantly. The accuracy, continuity and completeness of the resulting assembly are largely dependent on the level of this redundancy. Among mammals, the reference human and mouse genomes have been sequenced most extensively, with each base represented by an average of >8 reads (that is, >8x coverage). For the human genome, it is estimated that the assembled genome sequence covers ~99% of the euchromatic DNA, with an accuracy of >99.99% (REF. 1). A small group of mammals that are of particular importance for biomedical, agricultural or evolutionary research have also been targeted for extensive genome sequencing (6x–8x coverage). These species include the rat, dog, macaque, cow, chimpanzee, opossum and platypus (see the National Human Genome Research Institute Genome Sequencing Proposals web page). However, despite continual cost reductions, it remains prohibitively expensive to generate 8x sequence coverage for a typical mammalian genome using current technologies (~US$25,000,000). Consequently, the genomes of species that comprise large sections of the mammalian phylogenetic tree have remained largely unexplored.

But is it necessary to sequence all genomes to 6x–8x coverage? Survey sequencing (1x–2x coverage) generates 60–80% of a genome sequence, but in a highly fragmented form. However, given the existence of the reference human and mouse genome sequences, from which annotation can be obtained, survey-sequence data for a mammalian genome can provide a substantial amount of information[2]. This has been tested by an analysis of the dog genome after 1.5x sequence coverage[3]. The study showed that 1.5x coverage provides reliable estimates for several global parameters of the dog genome, such as its size, repeat content and neutral mutation rate. The coverage included partial sequence data for dog orthologues of ~75% of annotated human genes. An obvious limitation, relative to a 7.5x assembly (K.L.-T., unpublished observations), was that few dog genes were sequenced completely; most consisted of multiple fragments and a small fraction were likely to have been missed entirely. Nonetheless, these gene fragments have provided a valuable resource to rapidly develop short tandem repeat (STR)-based, or SNP-based assays for resolving linkage between candidate genes and specific phenotypes. The study also revealed that >4% of intergenic sequence is conserved between the dog, human and mouse. Each new genome that is sequenced (or randomly sampled) will further resolve those regions that are of crucial functional importance[4].

Survey sequencing addresses two of the principal motives for sequencing mammalian genomes. First, it provides a genomic infrastructure for mapping studies that aim to identify the genetic basis of specific traits and diseases. Second, through the comparative analysis of multiple genome sequences, it can reveal conserved genomic elements that are likely to be of general functional importance.

Currently, the value of survey sequencing is being tested further in an extensive project that has focused on nine additional mammalian genomes (see the National Human Genome Research Institute Genome

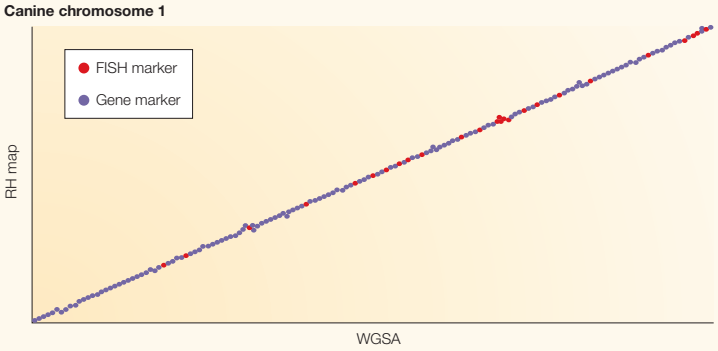


Figure 1 | **Marker order for canine chromosome 1, as determined by radiation-hybrid mapping and sequence assembly.** Sequences for markers on canine chromosome 1 (CFA1) that are localized on the radiation-hybrid (RH) map were aligned with the whole-genome sequence assembly (WGSA; CanFam1.0) by BLAST. CanFam1.0 coordinates for all markers were determined (*x* axis) and plotted against the RH-map marker position (*y* axis). The plot is linear, illustrating the near-perfect agreement of the order of the markers between the two datasets. On all chromosomes, subsets of markers ordered on the RH map were also localized by fluorescence *in situ* hybridization (FISH). Red dots indicate markers that were specifically localized on CFA1 by both FISH and RH mapping, blue dots indicate markers that were localized by RH mapping only. There is perfect concordance between the draft sequence and RH map position for all FISH mapped markers. See also supplementary information S10 (figure).

Sequencing Proposals web page)[4]. For these studies, approximately 2x sequence coverage will be generated from the genomes of the elephant, hedgehog, shrew, guinea pig, tenrec, armadillo, rabbit, wallaby and domestic cat. The resulting data will be particularly valuable for identifying genomic elements that are conserved among widely divergent clades of the mammalian phylogeny. In addition, the expanding datasets of genome survey sequences are bound to provide challenges and promote new approaches for genome characterization.

Perhaps the most serious limitation of survey sequencing, relative to more complete (6x–10x) coverage, is the fragmentary nature of the product. Assembly of the dog genome after 1.5x coverage yielded 1.9 million short CONTIGS, with an average length of 1.0 kb, that could be ordered across only short regions of the genome (a mean of 8.6 kb (REF. 3)). By comparison, assembly of a distinct dog genome sequence after 7.5x sequence coverage yielded a product that covers 99% of the genome and fragment ordering that extends over complete chromosomes (K.L.-T., unpublished observations). The absence of long-range continuity after survey sequencing severely restricts the detection of conserved SYNTENY with the genomes of related species. Consequently, the detailed annotation of reference genomes cannot be readily exploited to predict the order of genes and markers in a surveyed genome. In an attempt to address this issue, we have used survey-sequence data to construct a dense radiation-hybrid (RH) map of the dog genome, and compared the resulting map with those derived from the 7.5x sequence assembly. Could this approach circumvent the lack of high-coverage sequences in organisms that are destined only for survey sequencing?

The added value of RH mapping

Our interest in the dog system derives from its high level of phenotypic variation[5–8]. Through rigorous selection of phenotypic traits, domestic dogs have been divided into more than 350 genetically closed populations, termed breeds. Each corresponds to a genetic isolate and is defined by a breed standard that specifies every conceivable aspect of morphology, including coat colour, body size, leg length, head shape and a range of behaviours[9]. The outcome of such restrictions is a comparatively high incidence of genetic disease in purebred dogs. Collectively, studies of the domestic dog offer the opportunity to identify genes that are important in mammalian morphology, disease susceptibility and behaviour[7,8,10,11].

For the past decade we have focused on developing the resources needed to navigate the canine genome[12–19]. Among the latest developments was the production of a 1.5x survey sequence from a standard poodle genome[3]. To test our hypothesis that RH mapping can maximize survey sequencing, we derived more than 10,000 gene markers from the survey sequence and localized them on the canine genome using a 9,000-rad RH panel. Irradiation doses are measured in rads. A 9,000-rad RH panel is one that was constructed by fusing cells that have been γ-irradiated with 9,000 rads. These results, summarizing both the order of the mapped genes as well as identifying many evolutionarily conserved segments, were compared with data from a 7.5x draft sequence, which was derived from a purebred boxer. More than 90% of the markers are located within the 18,201 genes (and 5-kb flanks), derived from automated annotation of the 7.5x assembly (Ensembl release 30.1b).

RH mapping relies on the distribution of markers within a panel of hybrid cells, each of them harbouring a subset of the genome of interest, usually 15 to 40%. Markers that are physically close on the chromosome will experience less chromosome breakage during the irradiation process, and will therefore tend, on average, to be present more often within the same hybrid cells of a panel than more distantly positioned markers[20]. The RH-mapping process is divided into three steps. First, a panel of hybrid cell lines is constructed. These are obtained by fusing a rodent carrier cell line to cells of interest, following irradiation to break the test genome into fragments of a size suitable for the desired resolution[21]. Next, each cell line is tested by PCR amplification, and markers that are present or absent in each cell line are recorded. Typically, data are scored using a 1/0 system to indicate the presence or absence, respectively, of a given marker in a particular cell line. Subsequently, the resulting dataset is used to determine the order and spacing of markers along each chromosome. The entire data summary for each marker is termed a 'vector'. The final stage of RH map-construction computes the data from all the vectors using dedicated software programs including TSP/Concorde[22,23] and MultiMap[24,25], combined with other approaches[25–27].

We have combined RH-mapping technology with the 1.5x survey sequence of the dog genome to construct a map of more than 10,000 gene-based markers. In theory, the survey sequence covers 78% of the dog genome. Comparison with the human genome identified fragments of 18,473 annotated genes[3]. We have selected markers from the dog genome that were unambiguously orthologous to annotated human genes, and were evenly spaced throughout the genome, based on the reference human genome sequence.

The canine RH map was constructed using a high-resolution 9,000-rad dog–hamster panel, which has a theoretical resolution of 200 kb (12,000 unique positions). For more details, see supplementary information S1 (box), the Dog Genetics and Whole Genome

Mapping web site and the rh_tsp_map2 package[22,23,28] (supplemetary information S2–S4 (boxes and table)). This approach delivered a comprehensive map in which 89.8% of the markers were ordered with strong statistical support. We checked each chromosome map for collinearity with its corresponding canine cytogenetic map and previously published 5,000-rad RH map[12]. In addition, we genotyped a set of 545 BAC-ends on both the newly created 9,000-rad panel (F. Senger, personal communication) and the 5,000-rad panel[21] that was used to construct previous versions of the canine map, allowing easy navigation between maps[12,13,15,16,18]. Based on the resolution of the RH 9,000-rad panel, the 10,348 gene-based markers occupy 9,081 unique positions that are uniformly spaced across all canine chromosomes (CFAs), except the Y chromosome.

We also ordered on the 9,000-rad RH map an extra set of 500 canine BACs that had been previously localized by FLUORESCENCE *IN SITU* HYBRIDIZATION (FISH). These were useful for orienting RH groups relative to one another and the centromere, and for assigning RH groups to specific chromosomes. Such data are not required for RH map-building, or for maximal utilization of the sequence, but are clearly advantageous for investigating evolutionary comparisons between organisms.

Besides RH mapping, BAC mapping can be considered as an alternative approach for obtaining comparative mapping data. However, the workload, cost and objectives associated with each method are not the same. BAC mapping is done by fingerprinting hundreds of thousand of clones, which requires digesting DNA isolated from each BAC clone into appropriately sized pieces. Comparing DNA-fingerprint patterns amongst BAC clones facilitates the identification of overlapping clones. However, to navigate between a BAC-fingerprint map and a reference sequence, both ends of several thousand clones need to have been sequenced. In addition, BAC contigs, assembled by fingerprinting, must be localized onto the surveyed genome. For example, fingerprinting of mouse and rat BACs yielded 7,500 (REF. 29) and 11,300 (REFS 30,31) contigs, respectively, that themselves had to be anchored onto chromosomes by FISH or RH mapping. Therefore, although a BAC physical map, which contains both the sequenced and fingerprinted BAC end, is ideal for assembly of a deep 7x–10x SHOTGUN-SEQUENCING effort[29,31], a dense RH map, which organizes sequence data at a minimal cost and in a rapid time frame, is more appropriate for survey-sequence efforts. The markers of the RH map can connect contigs and, in doing so, confidently define blocks of conserved synteny between the survey-sequenced and reference genome.

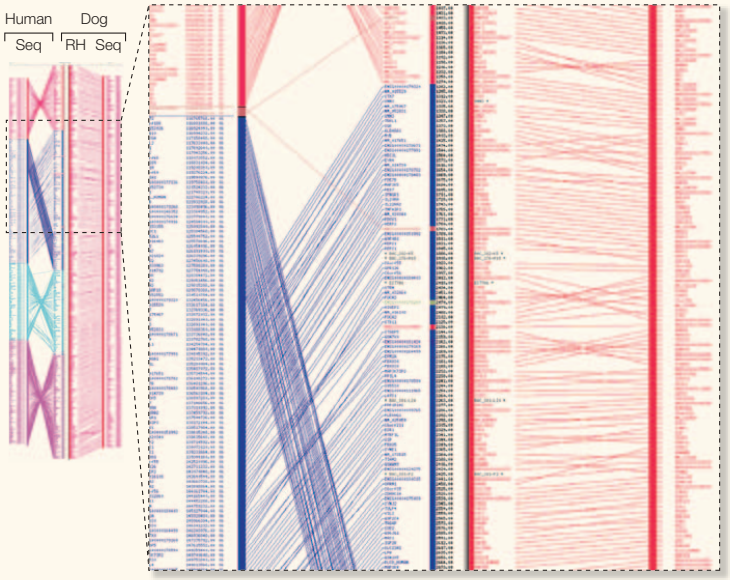


Figure 2 | **Integration of the canine radiation-hybrid map with the dog sequence assembly and human data: the example of canine chromosome 1.** The left panel shows a full view of the integration of the radiation-hybrid (RH) map and sequence assembly (Seq) for canine chromosome 1 (CFA1). The right panel focuses on a portion of CFA1. The RH map of CFA1 is represented in the middle of each panel. Marker names are on the left and right of the bars, and RH coordinates are in the middle. The bar on the far right of the right panel summarizes the sequence assembly (CanFam1.0) for CFA1. Lines connect the RH map and the CanFam1.0 sequence, demonstrating collinearity. Coloured boxes on the left side of the panel indicate the human counterparts, which were identified through mutual-best dog–human BLAST alignment. For CFA1, four human conserved segments, human chromosomes HSA18, HSA6, HSA9 and HSA19, were identified. The connecting lines represent dog–human rearrangements within the conserved segments. Although CFA1 contains four human conserved segments, a high level of intra-chromosomal rearrangement is evident within each. See also supplementary information S9 (figure).

RH mapping and genome organization

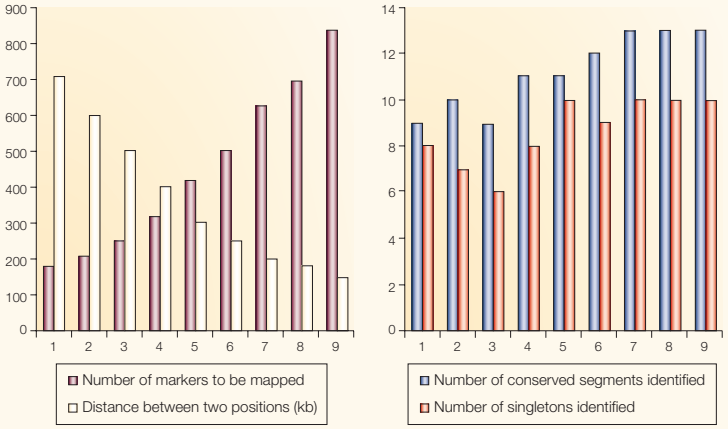
In the case of the dog, we selected gene-based sequences to obtain a well-spaced set of 9,850 markers corresponding to evenly spaced human genes, which were identified by their MUTUAL-BEST BLAST MATCHES[3] (see also the BLAST information web page). Mutual-BLAST alignments identify the best target (human) sequence that aligns with a query (dog) sequence and ensure that the target sequence has no better match on the query genome[32]. As such, the mapping allowed the precise positioning of 9,850 gene fragments and their related contigs and SCAFFOLDS into their constituent dog chromosomes. This facilitated the easy identification of points of conservation between the two genomes. Two hundred and sixty-four conserved segments (CSs) that were >500 kb and comprised 2 to 332 markers were identified by comparing the canine map and the human genome sequence (NCBI Build 34). A total of 243 CSs of >500 kb were identified by the CanFam1.0 7.5x assembly (K.L.-T., unpublished observations), indicating that a dense RH map resolves the conserved segments, as well as a 7x–10x whole-genome shotgun sequence (see supplementary information S5–S7 (tables)). The extra syntenic fragments identified by the RH map could be explained by the localization of about 20 SINGLETONS. These might represent true minor rearrangements that were very close to the size cutoff of 500 kb, or might be errors in the RH-map assembly. Analysis of closely positioned adjacent sequences is needed to determine the significance of these isolated cases. Detailed comparison of the latest canine 7.5x draft (CanFam2.0) (K.L.-T., unpublished observations) to the 9,000-rad RH map shows full agreement between the RH map and the draft sequence (FIGS 1,2; see also supplementary information S6,S8–S10 (table and figures)).

PERSPECTIVES

In addition to comparison with the human sequence that was used to guide the selection of canine gene markers, we compared the dog survey sequence/RH-map resource with the mouse and rat assemblies[30,33,34], identifying 394 and 414 CSs that were ≥500-kb long, respectively, in 8,422 and 8,158 BLAST alignments. These numbers are in agreement with the previously noted higher level of genome rearrangement between dog and rodent chromosomes, as compared with dog and human chromosomes[34].

Optimization, cost and perspectives

The construction of genome maps has been optimized for cost, accuracy and resolution. The RH-mapping methodology allows investigators to map all types of genomic marker, regardless of their sequence content. Of particular interest to geneticists are highly variable MICROSATELLITE sequences that have proved useful for tracking the inheritance of segments of chromosomes. As such, they have become the mainstay of linkage-mapping studies in nearly all the mammalian systems for which they have been developed. In the course of constructing the 10,000-gene dog map, we observed that 10% of the mapped gene markers were associated with a microsatellite sequence. If we had preferentially selected markers based on this criterion, rather than on position and spacing, that number would have been considerably higher. Other mapping projects, for which the primary biological interest in the sequence of their organism is to provide a resource for mapping and cloning studies, might decide to assign more importance to microsatellite proximity than we did in order to develop a resource for genetic linkage mapping.



Position on the *x* axis	Number of markers to be mapped	Distance between two positions (kb)	Number of conserved sequences identified	Number of singletons identified
1	179	700	9	8
2	208	600	10	7
3	250	500	9	6
4	313	400	11	8
5	417	300	11	10
6	500	250	12	9
7	625	200	13	10
8	694	180	13	10
9	833	150	13	10

Figure 3 | **Simulation of the number of radiation-hybrid markers required for optimal comparative genome-mapping studies.** The simulation was done using genes on canine chromosome 1, extracted from Ensembl (release30.1b). Among the 1,059 annotated dog genes, 833 have a single orthologue in the human sequence, as defined by a BLAST threshold. Sets of markers comprising 179 to 833 canine genes were selected to represent an average distribution of 1 gene every 150 kb to 1 gene every 700 kb. The table and histograms show that the optimal number of radiation-hybrid markers needed to identify all canine–human fragments (Ensembl release30.1b of the 7.5x draft assembly) is 625 (1 every 200 kb). This refers to set 7, with the maximum number of conserved segments and singletons identified for the fewest markers. A total of 417 markers (1 every 300 kb) would be needed to identify >90% of canine–human segments. The last two columns of the table relate the number of human–canine conserved segments and singletons (canine–human fragment defined by one locus) that can be identified when considering nine different sets of canine genes, with an inter-gene distance ranging from 150 to 700 kb.

Glossary

CLADE
A monophyletic group of related organisms that share a common ancestor, and therefore share similar features.

CONTIG
An abbreviation for contiguous sequence; used to indicate a set of DNA segments that overlap.

FLUORESCENCE *IN SITU* HYBRIDIZATION
A cytogenetic technique used to analyse specific chromosomal regions. Typically, large flourescently labelled pieces of DNA are used as probes in hybridization experiments to determine the presence, absence or orientation of chromosomal material in a region of interest.

MICROSATELLITE
Di-, tri-, or tetranucleotide repeat sequences that are composed of large numbers of tandem repeats. Microsatellites are widespread throughout mammalian genomes. Large numbers of alleles are generally associated with each microsatellite within most populations. Therefore, they are frequently used as markers for carrying out family-based linkage analysis.

MUTUAL-BEST BLAST MATCHES
A method for carrying out sequence comparisons. The method uses the BLAST algorithm to identify the best match for a given query sequence. The mutual-best BLAST method is frequently used to identify orthologous genes from different species.

REFSEQ
The Reference Sequence project. A database of annotated human genes. The overall goal of the project is to produce a reference sequence for all naturally occurring molecules.

SHOTGUN SEQUENCING
A genomic sequencing strategy that involves random fragmentation of large DNA segments. The fragments are sequenced, and programs with highly refined algorithms are used to reassemble the original DNA sequence.

SINGLETON
A single gene or marker that identifies the comparable region in a target genome, but does not cluster with other genes or markers to form a segment.

SYNTENY
The propery of being located on the same chromosome. Conserved synteny is revealed by the localization of orthologous genes between species.

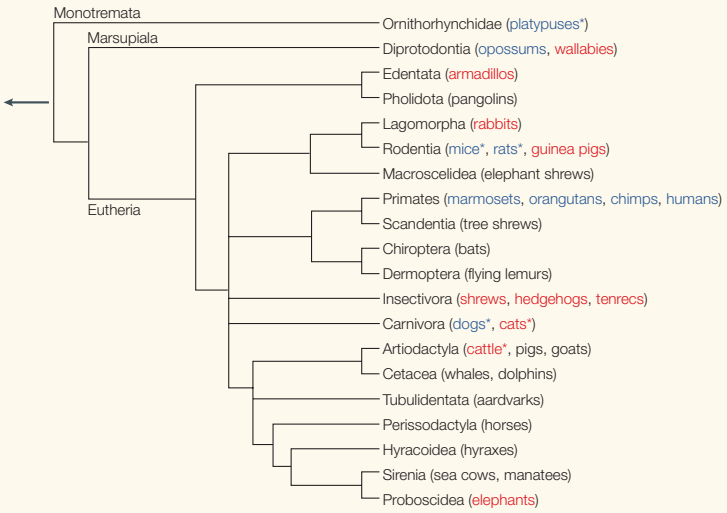


Figure 4 | **Mammals in the genome-sequencing pipeline.** The tree represents the primary orders of mammals as classified by the Tree of Life web project. The names shown in blue represent organisms that are scheduled for a 4x or higher coverage draft sequence; those in red are scheduled for survey sequencing (2x). *Organisms for which the genome sequence is started or already completed. Each of the three groups (Eutheria, Marsupiala and Monotremata) have a representative in the sequencing pipeline. In addition, among the 18 subgroups of the eutherian branch, 8 species have already been sequenced or are scheduled for sequencing (5 subgroups are at 2x coverage).

Importantly, the 18,473 gene fragments available in the dog 1.5x survey sequence[3] did not limit our selection of 10,000 gene sequences for RH mapping. In fact, we suggest that an even lower level of sequence coverage would generate sufficient data for building a high-quality comparative map. Assuming a gene count of 23,269 from humans (the number currently included in REFSEQ), we estimate that 14,500–16,500 dog gene fragments of at least 100 bp could be derived from a 0.5x–1x survey sequence (FIG. 3; see also supplementary information S8 (box)). Although the assignment of a mutual-best hit becomes less reliable with lower coverage, these calculations indicate that a 0.5x–1x survey sequence is indeed sufficient when combined with dense RH mapping to generate a high-quality comparative map resource.

Although 1x–2x survey sequencing cannot provide a contiguous picture of a genome, completion of the canine 10,000-gene RH map shows how a collection of fragmentary survey-sequence data can be transformed into an accurate, high-resolution map of a mammalian genome. The proposed strategy can be accomplished at a cost that is 5–6 times lower than that required for a high-quality draft (7x–10x) sequence. Indeed, in the case of the dog, 1.5x sequence coverage costs approximately $4,000,000, and the 10,000-gene RH map costs less than $1,000,000.

Considering an even lower level of 0.5x–1x sequence coverage, optimized marker selection and an RH map with a resolution of 200–300-kb, we can meet the goal of providing extended and reliable comparative genome information to those interested in the biology of model organisms for about 10% of the cost of a 6x–8x whole-genome shotgun assembly. Nevertheless, a combination of ~2x sequence aligned to the reference genome and an RH map to anchor the ordered and orientated genome sequence would be preferable under some circumstances, for example, when a more complete picture of the genome structure, a gene content to synteny comparison or gene content information and a better motif-sequence search capability[4] is required.

Each of the nine mammalian genome projects planned for low-coverage sequencing represents a specific taxon (FIG. 4). However, to extract maximal information from the dense sequencing efforts that have taken place so far, particularly for the human genome, many more species need to be sequenced. The many applications of the combined survey sequence/RH-map strategy proposed here are evident.

Finally, we note that high-throughput sequencing and RH mapping are by nature complementary methodologies. As such, they suggest intensive and fruitful collaborative projects within the scientific community. Even low-coverage sequencing generates a heavy workload that can be accommodated by only a few specialized centres. RH mapping does not require specialized equipment and, once the initial cell-line panel is made and characterized, can be undertaken in most molecular biology laboratories. Therefore, one result of using this approach might be closer ties between specialized sequencing centres and smaller academic laboratories that are focused on the underlying biological questions of interest. The result is certain to enrich both types of laboratory, providing the smaller academic laboratories with an opportunity to understand better the process of sequence analysis, while offering genome centres a more central role in driving the biological questions that have made specific model organisms stand out from the multitude of choices. It is to be hoped that the strategies and examples presented here will bring us one step closer to the ultimate goal of ultra-low-cost, efficient genome sequencing.

Christophe Hitte, Catherine Priat, Fabrice Senger, Thomas Derrien, Gwenaelle Evanno, Edouard Cadieu, Kristelle Lourgant, Sandrine Tacher, Stéphane Dréano, Nadia Berkova, Catherine André and Francis Galibert are at the CNRS, UMR 6061, Génétique et développement, Faculte de Médecine, 35043 Rennes Cedex, France.*

Jennifer Madeoy, Travis D. Lorentzen, Christina Ramirez, Barbara Pullar, Vinay Oza and Elaine A. Ostrander are at the Divisions of Human Biology and Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024, USA.*

Elaine A. Ostrander is also at the NHGRI/NIH, 50 South Drive, MSC 8000, Building 50, Room 5351, Bethesda, Maryland 20892-8000, USA.

Ewen F. Kirkness and Claire Fraser are at The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA.*

Dan Thomas, Carol Scott and Panagiotis Deloukas are at the Wellcome Trust Sanger Institute, Hixton, Cambridgeshire, UK.

David B. Jaffe and Kerstin Lindblad-Toh are at the Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA.

**These authors and their laboratories contributed equally to this work.*

Correspondence to E.A.O. at the NHGRI/NIH and F.G. at the CNRS
e-mails: eostrand@mail.nih.gov; francis.galibert@univ-rennes1.fr

doi:10.1038/nrg1658
Published online 12 July 2005

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 915–916 (2004).
2. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. & Gibbs, R. A. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8**, 1074–1084 (1998).
3. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
4. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
5. Ostrander, E. A. & Giniger, E. *Semper fidelis*: what man's best friend can teach us about human biology and disease. *Am. J. Hum. Genet.* **61**, 475–480 (1997).
6. Galibert, F. *et al.* The importance of the canine model in medical genetics. *Bull. Acad. Natl Med.* **182**, 811–821 (1998).
7. Galibert, F., Andre, C. & Hitte, C. [Dog as a mammalian genetic model]. *Med. Sci. (Paris)* **20**, 761–766 (2004) (in French).
8. Sutter, N. B. & Ostrander, E. A. Dog star rising: the canine genetic system. *Nature Rev. Genet.* **5**, 900–910 (2004).
9. American Kennel Club Staff. *The Complete Dog Book* (eds Crowley, J. & Adelman, B.) (Howell Book House, New York, 1998).
10. Chase, K. *et al.* Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl Acad. Sci. USA* **99**, 9930–9935 (2002).
11. Patterson, D. Companion animal medicine in the age of medical genetics. *J. Vet. Internal Med.* **14**, 1–9 (2000).
12. Breen, M. *et al.* An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5**, 1–11 (2004).
13. Guyon, R. *et al.* Building comparative maps using 1.5x sequence coverage: human chromosome 1p and the canine genome. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 171–177 (2003).
14. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **14**, 2388–2396 (2004).
15. Guyon, R. *et al.* A 1-Mb resolution radiation hybrid map of the canine genome. *Proc. Natl Acad. Sci. USA* **100**, 5296–5301 (2003).
16. Breen, M. *et al.* 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 (2001).
17. Mellersh, C. S. *et al.* An integrated linkage-radiation hybrid map of the canine genome. *Mamm. Genome* **11**, 120–130 (2000).
18. Priat, C. *et al.* A whole-genome radiation hybrid map of the dog genome. *Genomics* **54**, 361–378 (1998).
19. Mellersh, C. S. *et al.* A linkage map of the canine genome. *Genomics* **46**, 326–336 (1997).
20. Cox, D. R. Radiation hybrid mapping. *Cytogenet. Cell Genet.* **59**, 80–81 (1992).
21. Vignaux, F. *et al.* Construction and optimization of a dog whole-genome radiation hybrid panel. *Mamm. Genome* **10**, 888–894 (1999).
22. Agarwala, R., Applegate, D. L., Maglott, D., Schuler, G. D. & Schaffer, A. A. A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res.* **10**, 350–364 (2000).
23. Hitte, C. *et al.* Comparison of the MultiMap and TSP/CONCORDE packages for constructing radiation hybrid maps. *J. Heredity* **94**, 9–13 (2003).
24. Matise, T. C., Perlin, M. & Chakravarti, A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature Genet.* **6**, 384–390 (1994).
25. Boehnke, M., Lange, K. & Cox, D. R. Statistical methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.* **49**, 1174–1188 (1991).
26. Schiex, T. & Gaspin, C. CARTHAGENE: constructing and joining maximum likelihood genetic maps. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 258–267 (1997).
27. Ben-Dor, A. & Chor, B. On constructing radiation hybrid maps. *J. Comput. Biol.* **4**, 517–533 (1997).
28. Applegate, D., Bixby, R., Chvatal, V. & Cook, W. On the solution of traveling salesman problems. *Documenta Mathematica — Extra volume, ICM* **3**, 645–656 (1998).
29. Gregory, S. G. *et al.* A physical map of the mouse genome. *Nature* **418**, 743–750 (2002).
30. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
31. Gibbs, R. A. & Weinstock, G. M. Evolving methods for the assembly of large genomes. *Cold Spring Harb. Symp. Quant. Biol. B* **68**, 189–194 (2003).
32. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
33. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
34. Zhao, S. *et al.* Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res.* **14**, 1851–1860 (2004).

Acknowledgements

We acknowledge the American Kennel Club Canine Health Foundation (E.A.O. and F.G.) and the US National Institutes of Health (E.A.O., E.K., F.G. and K.L.-T.). E.A.O. is the recipient of a Burroughs Wellcome Award in Functional Genomics; F.G. is supported by the Centre National Recherche Scientifique, the Université de Rennes1 and Conseil Régional de Bretagne (France). The canine high-quality draft sequence was supported by the National Human Genome Research Institute. D.T., C.S. and P.D. were supported by the Wellcome Trust.

Competing interests statement

The authors declare no competing financial interests.

Online links

FURTHER INFORMATION

BLAST information: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
Dog Genetics and Whole Genome Mapping: http://idefix.univ-rennes1.fr:8080/Dogs/RH10K-SOM-web.html
Ensembl: http://www.ensembl.org
National Human Genome Research Institute Genome Sequencing Proposals web page: http://www.genome.gov/10002154
Tree of Life web project: http://tolweb.org/tree

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box) | S3 (box) | S4 (table) | S5 (table) | S6 (table) | S7 (table) | S8 (box) | S9 (figure) | S10 (figure)
Access to this links box is available online.

OPINION

Silence of the genes — mechanisms of long-term repression

Laura Lande-Diner and Howard Cedar

Abstract | A large fraction of genes in the mammalian genome is repressed in every cell throughout development. Here, we propose that this long-term silencing is carried out by distinct molecular mechanisms that operate in a global manner and, once established, can be maintained autonomously through DNA replication. Both individually and in combination these mechanisms bring about repression, mainly by lowering gene accessibility through closed chromatin structures.

In unicellular organisms, most of the genes in the genome are in a perpetual state of activity, with only a small number being specifically recognized as targets for repression. By contrast, repression is a dominant theme in the regulation of gene expression in animal cells, with more than 50% of the genome being silenced in any particular cell type. From a biological point of view, there seem to be two principal categories of long-term repression. On the one hand, genes that are specifically turned on during early embryogenesis are subsequently silenced and then remain in this state throughout development in almost all somatic cell types[1]. On the other hand, most tissue-specific genes are set up to be repressed at an early stage of development and they too are maintained in this form in almost all cell types, only undergoing reactivation in their tissues of expression.

Although much is known about the dynamic regulation of gene expression, there have been few efforts aimed at understanding long-term repression. The fact that gene silencing is highly stable throughout development and seems to work in a global manner indicates that the mechanisms that bring it about must be able to autonomously preserve the correct repression pattern even though the genome is copied at every cell division. Here, we propose that although repression itself is mediated mainly through effects on the local transcription machinery or chromatin structure, the maintenance of long-term silencing is carried out by three fundamental mechanisms that involve sequence-dependent repression factors, DNA methylation and late replication timing. These mechanisms operate either individually or in combination to generate layers of repression that not only ensure correct developmental expression patterns but might also lower transcriptional noise[2].

Layers of long-term silencing

Sequence-dependent repression. The simplest mechanism of gene silencing is mediated by

ONLINE ONLY

Online biographies

Elaine A. Ostrander is a senior investigator and the Branch Chief of cancer genetics at the US National Institutes of Health National Human Genome Research Institute. She also remains an affiliated member of the Divisions of Clinical Research and Human Biology at the Fred Hutchinson Cancer Research Center, Washington, USA, where she was located from 1993 until recently. She received her Ph.D. from the Oregon Health Sciences University, and did her post-doctoral training at Harvard, Massachusetts, USA. She then went to the University of California, Berkeley, and Lawrence Berkeley National Laboratories where she began her work on the genome project. Ostrander's laboratory works on the genetics of cancer, as well as the development of the canine system for the genetic analysis of complex traits. She has co-authored more than 140 papers on these subjects.

Francis Galibert is Emeritus Professor at the University of Rennes1, France, where he works as a senior scientist in the Centre National de la Recherche Scientifique unit, Genetics and Development. He graduated from the Faculty of Pharmaceutical Sciences at the University of Paris, France, in 1958 and received his Ph.D from the University of Paris in 1968 for his work on the identification and characterization of the eukaryotic 5S RNA. He did post-doctoral training in the Laboratory of Molecular Biology, Medical Research Council unit, Protein and Nucleic Acid Section (with Frederick Sanger, at Cambridge, UK) where he began his work on DNA analysis. This led to his eventual sequencing of several genomes, including the hepatitis B virus genome, which in turn facilitated production of the first recombinant vaccine. In 1996, he started a group to focus on development of the canine genome map, hereditary diseases and olfaction. In 1987, he was elected a member of the European Molecular Biology Organisation.

Online links

BLAST information:
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

Dog Genetics and Whole Genome Mapping:
http://idefix.univ-rennes1.fr:8080/Dogs/RH10K-SOM-web.html

Ensembl:
http://www.ensembl.org

National Human Genome Research Institute Genome Sequencing Proposals web page:
http://www.genome.gov/10002154

Tree of Life web project:
http://tolweb.org/tree

2. Vers une meilleure annotation du génome du chien

2.1. L'analyse du chromosome "Unknown" canin

(Article 3 : Derrien et al., 2007)

Depuis 2005 et le séquençage profond de son génome, le chien est devenu le cinquième mammifère à entrer dans "l'ère post-génomique". L'analyse des données générées et, en particulier, l'annotation complète de son génome en séquences fonctionnelles (codantes et non codantes) constitue une étape importante dans l'exploitation de ces données de séquençage. Ainsi, l'identification exhaustive de l'ensemble des gènes codant pour des protéines permettra d'une part, de caractériser l'espèce canine par rapport aux autres mammifères et, d'autre part, facilitera les études de gènes candidats impliqués dans les nombreux traits, phénotypes ou maladies génétiques spécifiques de race et fortement homologues de maladies humaines.

La première version de l'assemblage de la séquence canine (CanFam 1.0) a engendré la création d'un ensemble de données regroupé dans un lot appelé chromosome "Unknown". Ce "chromosome", qui correspond aux contigs non assemblés sur les 39 chromosomes du chien, ne possède qu'un intérêt limité pour la compréhension de l'organisation son génome. Néanmoins, sa taille de 92,8 Mb le classe parmi les six plus grands "chromosomes" canins et le système d'annotation standard d'Ensembl (Curwen *et al.*, 2004) référence plus de 500 gènes disposés sur ce chromosome.

2.1.1. Objectifs - Résumé

Dans cette étude, nous nous sommes intéressés à un sous-ensemble de 132 gènes canins localisés sur le chromosome Unknown et disposant d'un orthologue humain en relation d'orthologie 1:1. Par la construction de cartes de synténie entre plusieurs génomes référents et le chien, nous montrons que l'analyse multi-espèces de la conservation de l'ordre des gènes permet de définir de nouveaux intervalles de localisation pour 115/132 gènes canins préalablement assignés au chromosome Unknown.

2.1.2. Résultats

• Caractéristiques du chromosome Unknown canin (noté CFA_Unk à présent)

Parmi les 506 gènes canins localisés, sur le CFA_Unk, seuls 132 disposaient d'un orthologue humain en relation d'orthologie 1:1 (Ensembl v.38). Cette proportion (26%) est très inférieure à la proportion observée pour les autres chromosomes canins (82%) et reflète, probablement, qu'à l'instar des difficultés d'assemblage des séquences du CFA_Unk, s'ajoutent des difficultés de prédiction des relations d'orthologie avec les autres organismes. La présence d'un fort taux de séquences répétées (46% pour le CFA_Unk comparé à 35% en moyenne pour les autres chromosomes) pourrait engendrer des difficultés d'alignements de séquences soit lors du processus d'assemblage, soit lors de la prédiction des relations d'orthologie.

• Construction des cartes de synténie Homme/Chien

À partir de l'outil Biomart (http://www.biomart.org/ - Durinck *et al.*, 2005), nous avons récupéré 14 456 gènes annotés par Ensembl (v.39) en relation d'orthologie 1:1 entre l'homme et le chien. Ces ancres de comparaison ont ensuite été utilisées par le programme AutoGRAPH pour la construction des 23 cartes de synténie entre l'homme (génome référent) et le chien (génome testé). Cette approche comparative a mis en évidence plus de 200 segments conservés ordonnés (CSOs) d'une taille de ~12 Mb entre les deux génomes.

Nous avons ensuite estimé la colinéarité au sein des CSOs en calculant le nombre de gènes possédant un ordre conservé entre les deux génomes divisé par le nombre total de gènes présents dans le CSO. Plus de 94% des gènes montrait un respect de l'ordre des gènes entre les deux génomes. Les 6% restant sont probablement liés à des micro-réarrangements de type inversion ou à des transpositions de gènes.

• Prédiction d'intervalles orthologues canins

Les 132 gènes humains disposant d'un orthologue canin sur le CFA_Unk ont ensuite été positionnés sur les cartes de synténie homme/chien. Comme le montre la figure (**Fig.18**), étant donné la forte colinéarité des gènes au sein des CSOs, il est possible de définir un intervalle canin délimité par les deux gènes "voisins" les plus proches en relation 1:1 encadrant le gène d'intérêt. Au total, 110/132 nouveaux intervalles ont pu être définis. Pour les 22 autres, leurs localisations entre deux CSOs au niveau d'une rupture de synténie, ne nous permettaient pas d'identifier un intervalle unique sur le génome du chien.

De façon intéressante, la caractérisation de ces intervalles canins a révélé la présence d'un fort taux de "vide" (gap) de séquence canine. Nous avons calculé la proportion de "gap" en fonction de la taille de l'intervalle canin. Alors que cette proportion est de ~1% pour 1 000 intervalles sélectionnés au hasard sur le génome du chien, nous avons observé qu'elle était de 9% en moyenne pour les 110 intervalles définis. De plus, près d'un tiers des intervalles canins ont révélé la présence de grandes régions de "gap" (arbitrairement fixées à 1kb par les projets de séquençage) et correspondant, de fait, à des régions entre deux supercontigs non reliés.

Par ailleurs, nous nous sommes intéressés à la distribution des 110 intervalles sur les chromosomes canins. Aussi, 66/110 (60%) des intervalles étaient localisés en partie télomérique des chromosomes. Les difficultés à assembler les régions télomériques des chromosomes corrèlent avec la présence d'un fort taux de "gap" dans ces régions.

Ces observations ont renforcé la possibilité d'une localisation de séquences canines, préalablement assignées au chromosome Unknown, au niveau de ces intervalles canins enrichis en "gap".

- Approche d'intégration de plusieurs espèces

Les ancres de comparaison ont été utilisées pour trois autres génomes référents : le chimpanzé, la souris et le rat correspondant à un total de 12 204, 14 171 et 13 309 gènes en relation 1:1, respectivement. Parmi les 132 gènes canins localisés sur le CFA_Unk et disposant d'un orthologue humain, 126 (95%) possédaient au moins un orthologue dans les trois autres génomes référents. Nous avons construit l'ensemble des cartes de synténie entre les paires de génomes chimpanzé/chien, souris/chien et rat/chien puis positionné les 126 orthologues sur chacune des cartes comme préalablement décrit. Cinq nouveaux intervalles canins (115/132) ont pu être déterminés quand ces derniers se trouvaient confirmés par au moins deux espèces de référence. En effet, tandis que pour l'homme, 22 gènes étaient localisés au niveau d'une rupture de synténie, l'addition de nouveaux génomes référents a permis de localiser cinq orthologues à l'intérieur de segments conservés entre les nouveaux génomes référents et le chien (**Fig.19**).

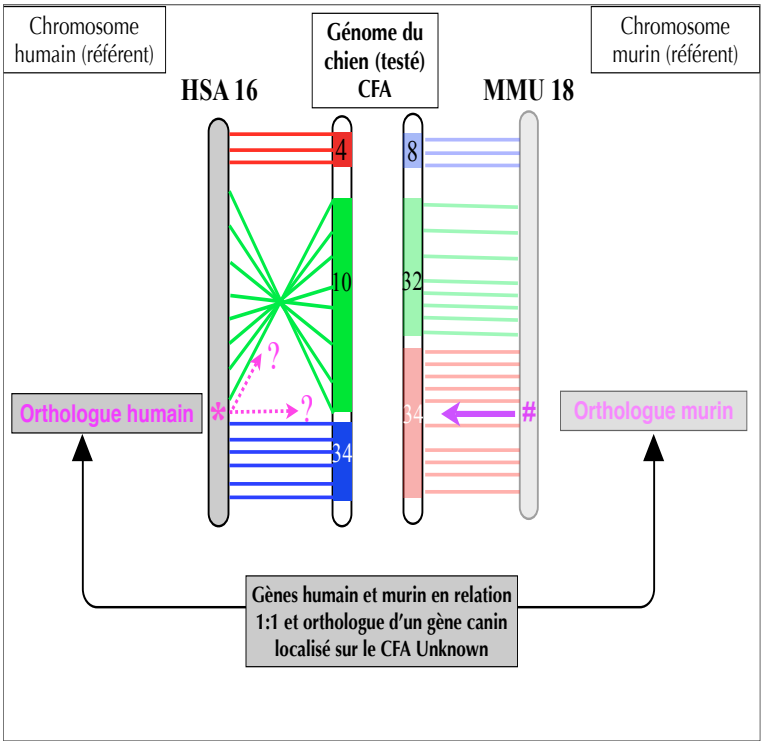


Figure 19 : Apport de l'ajout d'un autre génome de référence sur la possibilité de définir un nouvel intervalle orthologue canin.
Si l'orthologue humain est positionné sur son chromosome en rupture de synténie (entre deux CSOs) avec le génome du chien (astérisque rose), alors il n'est pas possible de définir un intervalle unique.
À l'inverse, si l'orthologue souris est localisé à l'intérieur d'un CSO avec le génome du chien (dièse rose), ce dernier est plus informatif et permet de définir l'intervalle orthologue canin.

De plus, les intervalles canins sont définis par rapport aux gènes voisins en relation 1:1 (référent/chien) encadrant le gène référent d'intérêt positionné sur la carte de synténie. Si le gène humain H_1 le plus proche du gène d'intérêt (en aval, par exemple) est en relation 1:0 avec le génome canin, alors la définition de l'intervalle canin sera "décalée" à la relation d'orthologie 1:1 homme/chien encore plus en aval (H_2:D_2). Cependant, l'ajout d'un autre génome de référence (souris) peut amener une nouvelle relation d'orthologie 1:1 (M_1:D_1) et ainsi permettre d'affiner la localisation de l'intervalle canin. En effet, pour huit intervalles, l'ajout de nouvelles ancres de synténie provenant d'autres espèces (essentiellement souris) a permis de réduire la taille de l'intervalle canin prédit. Ces gènes présents dans les lignées murine et canine ont probablement été perdus ou mal annotés chez l'homme. Les gènes codant pour des récepteurs olfactifs (OR) en sont une illustration puisque un grand nombre de gènes sont orthologues 1:1 entre le chien et la souris alors que beaucoup sont devenus des pseudogènes chez l'homme (Quignon *et al.*, 2005).

Ces 115 nouveaux intervalles canins ont été déterminés à partir de l'annotation Ensembl réalisée sur la première version de l'assemblage canin (CanFam1.0). Toutefois, la seconde version de l'assemblage (CanFam 2.0) était disponible au moment de notre étude mais sans être annotée par Ensembl. Nous avons réalisé des alignements de séquence des 115 gènes canins du CFA_Unk ainsi que des gènes encadrants sur l'assemblage du génome du chien mis à jour. Au total, 78/132 gènes du chromosome Unk (CanFam1.0) avaient été mis à jour et assignés à des chromosomes canins (CanFam2.0). Près de 95% (74/78) étaient en accord avec les prédictions d'intervalles canins que nous avions établies alors que 20/132 restaient annotés sur le chromosome Unk.

Par la construction de cartes de synténie et la mise en évidence de la colinéarité des ancres de comparaison entre un ou plusieurs génomes référents et le génome du chien, nous avons pu identifier 115 nouvelles localisations de gènes canins préalablement annotés sur le chromosome non assemblé. Le cas du chien n'est pas une exception puisque, exceptés les génomes humain et murin, les génomes séquencés possèdent des régions non assemblées regroupées dans des chromosomes unknown. Pour les génomes du chimpanzé et du poulet, par exemple, en plus du chromosome Unknown, on note la présence de chromosomes "virtuels" appelés "chromosomeX_random" ou X correspond au numéro du chromosome sur lequel le chromosome virtuel est ancré. Chez le chimpanzé, 107 Mb de séquence ont été placées dans ces chromosomes "random" et le chromosome Unknown a une taille de 50 Mb. La méthode employée dans cette étude pourrait rapidement être testée sur ces espèces afin d'améliorer leur annotation et d'affiner les études comparatives impliquant ces organismes. Enfin, cette approche comparative nécessite d'utiliser des génomes de référence relativement proche au niveau phylogénétique dans le but de d'observer le moins possible de réarrangements entre le génome référent et le génome testé. Cependant, cette limitation peut être compensée par l'utilisation d'un plus grand nombre de génomes référents qui maximiserait la probabilité d'observer au moins un gène appartenant à un segment conservé ordonné et non à des ruptures de synténie partagées entre toutes les espèces de référence.

Article n°3

Analysis of the unassembled part of the dog genome sequence: chromosomal localization of 115 genes inferred from multispecies comparative genomics.

Derrien T, Andre C, Galibert F, Hitte C.

J Hered. 2007

Journal of Heredity Advance Access published June 15, 2007

Journal of Heredity
doi:10.1093/jhered/esm027

Analysis of the Unassembled Part of the Dog Genome Sequence: Chromosomal Localization of 115 Genes Inferred from Multispecies Comparative Genomics

THOMAS DERRIEN, CATHERINE ANDRÉ, FRANCIS GALIBERT, AND CHRISTOPHE HITTE

From the CNRS UMR6061 Génétique et Développement, Université de Rennes 1, IFR140, 2 Av du Pr Léon Bernard, CS 34317, 35043, Rennes, France.

Address correspondence to C. Hitte at the address above, or e-mail: hitte@univ-rennes1.fr.

Abstract

The identification of dog genes and their accurate localization to chromosomes remain a major challenge in the postgenomics era. The 132 annotated canine genes with human orthologs remaining in the unassembled part (chrUnknown) of the dog sequence assembly (CanFam1) are of limited use for candidate gene approaches or comparative mapping studies. We used a two-step comparative analysis to infer a canine chromosomal interval for localization of the chrUn genes. We first constructed a human–dog synteny map, using 14 456 gene-based comparative anchors. We then mapped the 132 chrUn genes onto the reference (human) synteny map and identified the corresponding, orthologous segment on the canine map, based on conserved gene order. Our results show that 110 chrUn genes could be localized to short intervals on 18 dog chromosomes, whereas 22 genes remained assigned to 2 possible intervals. We extended this comparative analysis to multiple species, using the chimpanzee, mouse, and rat genome sequences. This made it possible to narrow down the intervals concerned and to increase the number of canine chrUn genes with an inferred chromosome location to 115. This study demonstrates that dog chromosomal intervals for chrUn genes can be rapidly inferred, using a reference species, and indicates that comparative strategies based on larger numbers of species may be even more effective.

Many large-scale mapping and sequencing projects have been completed in the last 10 years (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Hitte et al. 2005; Lindblad-Toh et al. 2005). This has made it possible to compare the genomes of different species and to study evolutionary changes (Hardison 2003; Murphy et al. 2005). The emerging field of comparative genomics has already yielded outstanding results in domains such as speciation and evolutionary studies (Hillier et al. 2004; Jaillon et al. 2004), genome annotation (Ashurst et al. 2005), and the identification of new sets of functional elements within annotated genomes (Dujon et al. 2004). Ongoing low-coverage sequencing projects will also provide new additional resources for many model organisms commonly used as human surrogates for research (Margulies et al. 2005, http://www.genome.gov/10002154).

Synteny maps are generated by identifying unambiguous orthologous sequence pairs across species (Pan et al. 2005; Liang and Dandekar 2006). These pairs, known as comparative anchors (Chen et al. 1999), are connected to show regions of conserved synteny and break point regions. Region of conserved synteny are composed of conserved segments (CS) and conserved ordered segments (CSO) (O'Brien et al. 1993). CS are segments shared by 2 or more species that contain orthologous anchor markers with no notion of order. CSO are conserved segments that run continuously, with the same orientation and the same order of genes, reflecting intrachromosomal rearrangements occurring during evolution (Pevzner and Tesler 2003). The extent of gene-order conservation within CSO depends both on the phylogenetic distance between organisms and on the frequency of species-specific rearrangements since divergence from the last common ancestor (Kirkness et al. 2003). Analyses of CS and CSO gene content across species are commonly carried out in studies of gene family expansion or contraction and to facilitate the identification and annotation of orthologous genes (Fischer et al. 2001; Hardison 2003; Zheng et al. 2005).

We report here a strategy making use of the large CSO revealed by synteny maps and gene adjacency information for the canine and human genomes to refine the localization of canine gene repertoire. We applied this method to

the unassembled part of the canine genome sequence (CanFam1) and were able to localize 115 canine chrUn genes to short chromosomal segments. We extended this approach to a multispecies comparative analysis including chimpanzee, rat, and mouse to refine chrUn gene localization.

Materials and Methods

Gene Data Sets

Orthologous gene data were downloaded from Ensembl v39 with the biomart tool (http://www.ensembl.org/Multi/martview). Data sets for human–dog, chimpanzee–dog, mouse–dog, and rat–dog pairs were successively downloaded. Ensembl describes several categories of orthology: one-to-one, one-to-many, and many-to-many. We used this classification to extract one-to-one orthologs. Data sets were stored in a mySQL database. Synteny maps were constructed, using one-to-one orthologs as comparative anchors. Synteny maps were built and drawn using the AutoGRAPH program (http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/).

Colinearity Rate

The colinearity rate was calculated as the proportion of genes from the target species in the same order as the genes of the reference species. It was determined for CSOs containing at least 3 genes. If gene order is perfectly conserved, then the colinearity rate is 100%. The algorithm used by AutoGRAPH makes it possible to relax constraints on colinearity rate. We set a gap penalty threshold of 5, making it possible to include a gene in a CSO, even if gene order is not conserved within a range of 5 positions (http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/Tutorial.php).

Inference of Canine Intervals Serving as Candidates for chrUn Gene Localization

Genomic intervals containing a human gene with an ortholog in the canine chrUn pool were defined based on the closest human flanking genes with one-to-one orthologs. Only flanking orthologous pairs in conserved order were used for the inference of canine ortholog intervals as candidates for chrUn gene localization. The same method was applied to 3 other reference species: chimpanzee, mouse, and rat.

Repeat Content

Repeat sequence content was calculated as the ratio of cumulative repeat sequence size to the total size of the chromosome sequence (the UCSC repeat content table can be downloaded from http://genome.ucsc.edu/cgi-bin/hgTables).

CanFam2 Analysis

Canine chrUn gene sequences and their flanking ortholog sequences obtained from the comparative study were aligned with the CanFam2 assembly, using Blat version 33. The results of chromosomal assignment studies and the order of the 3 sequence alignments were used for comparison with the comparative genomics analysis.

Results and Discussion

Genes from the Unassembled Part of the Canine Sequence Assembly (chrUn)

The unassembled part of a genome sequence assembly is usually placed in a "chromosome unknown" pool (chrUn), corresponding to sequences that cannot be effectively localized to particular chromosomal positions or assembled into sizeable contigs. The 132 protein-coding genes with a single human ortholog extracted from the canine chUn pool account for only 26% of the 506 genes present in this pool. This proportion is significantly smaller than that for Ensembl (Birney et al. 2006) one-to-one orthologs found for all other CFA (82%). The precise reason for this lack of orthology remains unclear, but the sequencing problems encountered in sequence assembly may be also involved in ortholog identification. The high repetitive sequence content of the chrUn—43% repeat elements versus only 35% in the assembled CFA (Karolchik et al. 2004)—may account for problems with the identification of real orthologs of chrUn genes by sequence alignment or phylogenetic studies. The human orthologs of the 132 chrUn genes are widely dispersed over the entire set of human chromosomes (except HSA4), 4 of which (HSA 7, 10, 17, and 21) contain 45% of the genes mapping in clusters at the ends of chromosomes (Table 1).

Selection of Human/Dog Comparative Anchors and Synteny Map Construction

We downloaded canine protein-coding genes from Ensembl v39 (EnsMart tool, Kasprzyk et al. 2004) and selected 14 456 such genes that were annotated and had a one-to-one orthologous relationship with a human gene. One-to-many and many-to-many orthologous relationships were excluded from the analysis to maintain data reliability and to prevent the occurrence of uncertainties in subsequent synteny analysis. We used AutoGRAPH—a web server developed in our laboratory (Derrien et al. 2007, Materials and Methods) that formalizes the construction of comparative maps, making it possible to relax marker-order conservation criteria through an adjacency penalty value—to build a human–dog synteny map. Based on this synteny map, we were able to identify 222 CSOs, with a mean length of 11.8 Mb, containing a mean of 91 genes (range 3–493) (Figure 1). Break points—regions separating CSOs—were identified and characterized on the basis of size (mean size 806 kb), as defined by the immediately flanking one-to-one ortholog pairs. Gene-order conservation within CSOs containing more than 3 genes was evaluated by determining the colinearity rate, corresponding to the percentage of genes in the same order in the target (dog) and reference (human) species. Mean colinearity rate was 94%, indicating a high level of gene adjacency

Table 1. chrUn genes distribution The Oxford grid is a schematic representation of the localization of the chrUn genes on canine (CFA) and human (HSA) chromosomes. Purple squares indicate the chromosomal attribution of chrUn genes. Number of chrUn genes is indicated in the square. The total of chrUn genes is reported in the last line and in the last column.

CFA \ HSA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	
1																			4					4
2																								0
3															2									2
4																								0
5																								0
6							13									5								18
7	1																	2						3
8														1										1
9																								0
10												1												1
11					5				1															6
12																								0
13								3																3
14																								0
15														1										1
16																								0
17		2																						2
18											1						12							13
19																								0
20			1																1					2
21											8													8
22																								0
23																								0
24																				2				2
25																								0
26																								0
27																								0
28										22														22
29																								0
30																								0
31																					15			15
32																								0
33																								0
34																								0
35						3																		3
36																								0
37																								0
38																								0
X																							4	4
	1	2	1	0	5	3	13	3	1	22	9	1	0	2	2	5	12	2	5	2	15	0	4	110

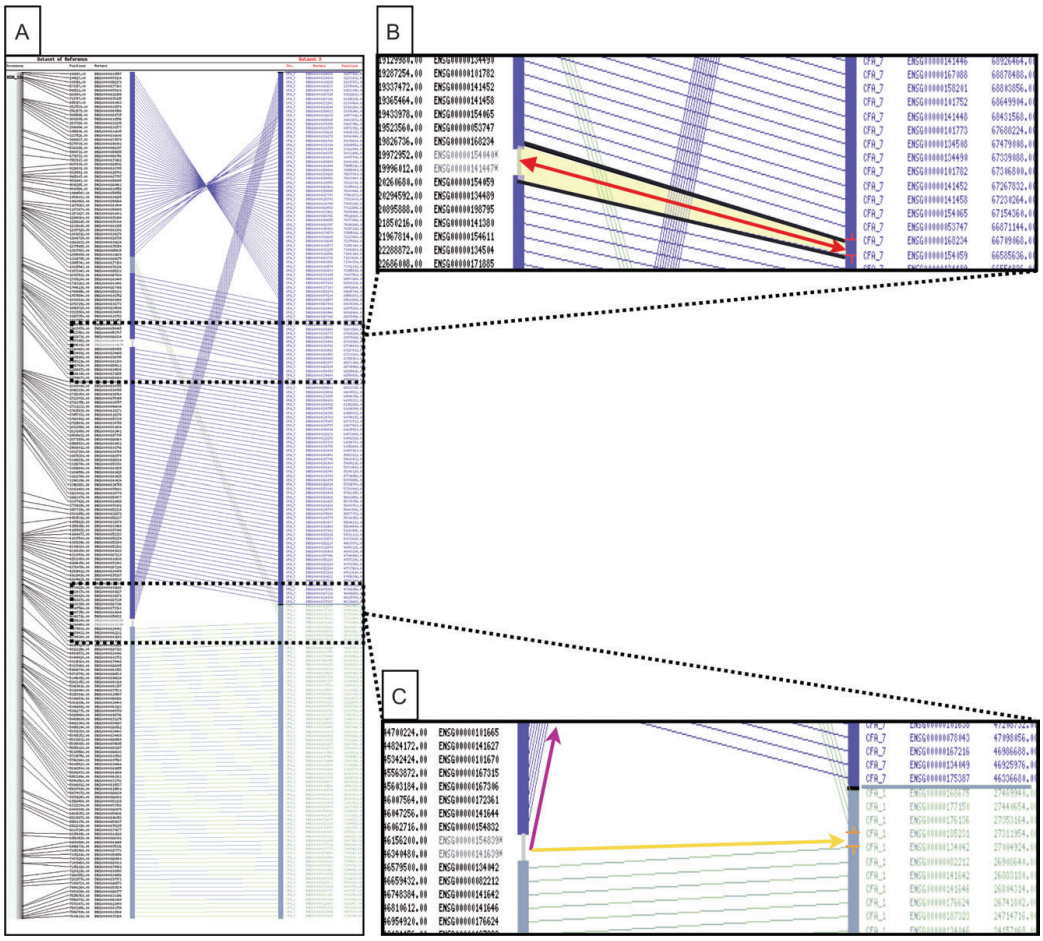


Figure 1. Human–dog synteny map: example of human chromosome 18. (**A**) Chromosome 18 from human, the reference species, is shown on the left-hand side of the figure. The bar on the right corresponds to the canine genome. For both species, genes are identified by their ID and genomic coordinates. The colored lines connecting orthologous genes identify 2 CS and 5 CSO and show the conservation of gene-order within the CSOs. chrUn genes are indicated on the synteny map and are shown in gray, with an asterisk, in the reference species. (**B**) This panel defines a canine ortholog interval, delimitated by 2 flanking one-to-one orthologous genes, predicted by inference to contain 2 chrUn genes. (**C**) If the canine ortholog interval was predicted to lie in the junction between 2 CS, arrows indicate the 2 possible dog chromosomes to which this interval could be mapped.

conservation within chromosomal segments inherited without rearrangements from the last common ancestor of dogs and humans. The proportion of genes for which order does not appear to be conserved (6%) in the human–dog synteny map may correspond to microrearrangements that have shaped genome evolution or to dubious orthologous relationships associated with the expansion or contraction of gene families (Goodstadt and Ponting 2006), making it difficult to identify true orthologs, as sequence assembly errors cannot be eliminated.

Inference of Orthologous Chromosomal Intervals from the Synteny Map

We used one-to-one orthologous gene-based comparative anchors to ensure the construction of a dense and reliable human–dog synteny map (Figure 1A). However, the accuracy of comparative maps may be decreased by errors in the definition of orthology, spurious gene annotations, or dubious genomic coordinates leading to the misinterpretation of gene-order conservation. We assessed the robustness of

Table 2. Characteristics of canine chromosomal intervals inferred from the human–dog and multispecies synteny maps: the example of human chromosome 19. (**A**) Canine chromosomal intervals characteristics from the human–dog synteny analysis. (**B**) [1] Multispecies analyses conducted with human (HSA), chimp (PTR), mouse (MMU), and rat (RNO) confirmed and narrowed down the canine interval size. The asterisk indicates which species led to the size refinement. [2] Size of the refined interval. [3] Blat alignment sequence results on CanFam2 for both the chrUn gene sequence (middle value) and flanking orthologs (flanking values).

A	Ensembl canine ID	Ensembl human ID	Human localization chr: start–stop	Canine chromosomal interval inferred (CanFaml)	Canine interval size (kb)
	ENSCAFG00000015323	ENSG00000167676	HSA_19: 4455471–4464704	CFA_20: 58006144–58063918	57.7
	ENSCAFG00000015415	ENSG00000188368	HSA_19: 47498124–47506813	CFA_1: 114237805–114252633	14.8
	ENSCAFG00000014192	ENSG00000204519	HSA_19: 62885189–62891468	CFA_1: 102524506–103202628	678.1
	ENSCAFG00000014217	ENSG00000179909	HSA_19: 62900547–62912370	CFA_1: 102524506–103202628	678.1
	ENSCAFG00000015129	ENSG00000083838	HSA_19: 63679564–63684403	CFA_1: 102142751–102263320	120.5

B	Ensembl canine ID	Ensembl human ID	Multispecies analysis [1]	Canine chromosomal interval inferred (CanFaml) [2]	Blat results (CanFam2) [3]
	ENSCAFG00000015323	ENSG00000167676	HSA-MMU*	CFA_20: 58010160–58063918	CFA_20: 58125693 < 58185309 < 58143701
	ENSCAFG00000015415	ENSG00000188368	HSA-PTR	CFA_1: 114237805–114252633	CFA_1: 114957992 < 114969059 < 114973083
	ENSCAFG00000014192	ENSG00000204519	HSA-PTR-MMU	CFA_1: 102524506–103202628	CFA_1: 102888975 < 103511307 < 103551598
	ENSCAFG00000014217	ENSG00000179909	HSA-PTR	CFA_1: 102524506–103202628	CFA_1: 102888975 < 103535021 < 103551598
	ENSCAFG00000015129	ENSG00000083838	HSA-PTR-MMU*-RNO	CFA_1: 102142751–102155210	CFA_1: 102503909 < 102510764 < 102627178

the conserved gene-order approach (Zheng et al. 2005), by selecting, at random, 1000 genes widely dispersed over all the human chromosomes and using these genes to evaluate the likelihood of identifying the chromosomal position of the dog gene based on the gene coordinates in the reference species (human). We masked canine genomic localization and mapped the 1000 reference species genes on the human–dog synteny map. We applied the conserved gene-order rule to identify the orthologous segment on the canine map, based on the closest flanking one-to-one orthologs (Figure 1B). We found that 93.2% of the 1000 genes tested were correctly localized on canine chromosomes, within an interval of 347 kb on average, as defined by the closest flanking one-to-one orthologs. For 3.1% of the genes, it was possible to infer position to within 170 kb, based on the interval defined by the closest flanking orthologs. This may be due to short microarrangements interrupting colinearity. It may also reflect inaccuracy in the coordinates of the gene in the human or dog genome sequence. For 2.4% of the genes, the predicted interval mapped to the junction of 2 CS of the synteny map, thereby precluding assignment to a single canine chromosome (Figure 1C). For the remaining 1.3%, the position on the canine interval was not correctly inferred because they correspond to singleton genes.

Mapping the chrUn Genes on the Synteny Map

When producing the human–dog synteny map, we mapped the human genes orthologous to the canine 132 chrUn genes in the reference species. We then identified the segment defined by the closest flanking genes with a one-to-one canine ortholog (Figure 1B). We applied the conserved gene-adjacency approach to identify the orthologous segment in the canine sequence most likely to contain the chrUn gene.

Considering only segments corresponding to CSOs containing at least 3 genes, we were able to identify a corresponding canine segment for 110 chrUn genes, the remaining 22 mapping to break point regions, preventing their assignment to a single interval on a single canine chromosome. The 110 corresponding canine intervals were distributed over 18 chromosomes, including 4 large segments on chromosomes 6, 18, 28, and 31, with up to 22 chrUn genes predicted on CFA28 (Table 1). Not surprisingly, these 4 segments correspond to genes also mapping in clusters at the ends of 4 human chromosomes (HSA 7, 10, 17, and 21; Table 1). The size of the 110 dog chromosomal segments was then evaluated as the distance between the 2 closest flanking one-to-one orthologs (or one ortholog and the chromosome end) (Table 2 and Supplementary Table 3). We were able to define 44 intervals based on flanking orthologous genes on either side of the

interval, and 66 intervals at the end of the chromosome were defined based on a flanking gene on one side and the telomere on the other. The mean length of the defined intervals was 415 kb, markedly longer than the interval size determined in the pilot analysis on 1000 randomly selected genes (347 kb), presumably due to overestimation of the size of intervals defined by an ortholog on one side and the telomere on the other. The large number of predicted intervals mapping to chromosome ends may be accounted for by difficulties in the assembly of these regions, due to problems retaining chromosome ends during construction of the clone library. Sequence coverage is lower for these regions than for other regions, and they are therefore less likely to be assembled.

We searched for gaps in the canine assembly and calculated the cumulative size of gaps, represented as stretches of *N* in the sequence (Karolchik et al. 2004; http://genome.ucsc.edu/cgi-bin/hgTables). Gap sequences were found to account for 9% of the intervals predicted by inference to contain a chrUn gene. This value is significantly higher than that for random intervals ($n = 1000$), in which gaps account for 1.5% of the sequence. We also found that 32% of the inferred intervals in dog contained large sequence gaps (arbitrarily set to 1000 bp in the assembly), corresponding to regions at the junction between supercontigs and related only by the physical map (UCSC server). The corresponding proportion was only 0.4% for random studies of 1000 intervals. Thus, intervals predicted, by inference, to contain chrUn genes contained a significantly higher proportion of gap sequences than randomly selected intervals. These findings are consistent with gaps in the inferred interval corresponding to chrUn gene sequences grouped in the chrUn pool.

Multispecies Comparative Analysis

We constructed pairwise synteny maps between the dog as tested genome and chimpanzee, mouse and rat species as reference genomes. These maps contained 12 204, 14 171 and 13 309 comparative anchors annotated by Ensembl between dog and chimpanzee, mouse and rat, respectively. We mapped 95% (126/132) of the chrUn genes on the reference species on at least one of the pairwise synteny maps and 56.8% (75) of these genes could be mapped on all 4 pairwise synteny maps, making use of the one-to-one orthologous relationships described between dog and the 4 species considered (human, chimpanzee, rat, mouse). From all pairwise synteny maps, genomic intervals in the dog sequence were predicted for 115 chrUn genes, 82% (92) of which were identified in at least 2 species. Furthermore, the multispecies analysis made it possible to shorten and refine the interval, using 2 flanking orthologous genes, for 8 and 5 intervals, respectively (Table 2). For 5 of the 22 canine chrUn genes initially assigned to break points, making it impossible to identify a single CFA with the human–dog synteny map alone, multispecies analysis led to the identification of a single CFA.

The use of the multispecies comparative approach made it possible to examine several sets of one-to-one orthologs differing in both number and nature. These differences may correspond to real biological differences. For example, a one-to-zero relationship for an olfactory receptor between dog and human may correspond to a one-to-one relationship between mouse and dog, both of which have a highly developed sense of smell (Quignon et al. 2005). Differences in orthologous data sets may also arise from inaccuracies in ortholog identification (Goodstadt and Ponting 2006).

Gene Sequence Analysis

Although an updated sequence assembly, CanFam2, has been released (Lindblad-Toh et al. 2005), the corresponding gene annotation set is not yet available from Ensembl database version 41. We have, however, performed sequence alignment analyses on the CanFam2 assembly (see Materials and Methods), using the Blat algorithm (Kent 2002) for the 115 dog chrUn genes and flanking orthologs localized in this study. We detected 78 of these gene sequences on canine chromosomes, with 94.8% (74/78) of the identified chromosomal locations consistent with the canine chromosomal interval determined in this study. The 4 gene sequences not confirmed by sequence alignment analysis may correspond to microrearrangements between dog and human undetectable on the synteny map or the incorrect definition of orthologous relationships. Twenty-five gene sequences were aligned to the chrUn pool, 7 were involved in major contig reassembly in the CanFam2 release, and 5 chrUn genes could not be mapped to any annotated chromosome in CanFam2 with significant sequence alignment. For these sets of genes, comparative genomics approaches are likely to be the most efficient for inferring the most probable location.

RH Mapping Validation

Among the 110 genes with new locations described in this work, a subset of 17 genes have been previously mapped in the canine 10 000 gene RH map built with the 9000-rad panel. Sixteen gene (16/17) placements were found in agreement with the computational approach, and one was slightly discrepent. These results indicate both the accuracy of the computational inference and the power of RH mapping to place chrUn genes.

Conclusion

In this study, we addressed the question of how comparative genomics can be used to map genes currently in the chrUn pool of the dog genome assembly. This approach is based on the construction of dense, accurate synteny maps with one-to-one gene-based comparative anchors and the use of gene-order conservation. It then makes use of the capacity to observe the conservation of gene order to infer orthologous intervals as candidates for the localization of chrUn genes. The evolutionary distance between the target and reference organisms is a critical factor with this method. It is essential to select phylogenetically close species (Margulies et al. 2005), such as pairs of mammals, and to consider multiple species, to maximize analysis efficiency. We constructed several synteny maps (human–dog, chimpanzee–dog, rat–dog, and mouse–dog) to identify, by inference, the canine segments

most likely to contain the 115 chrUn genes. We found that the frequency of gap sequences was high in the inferred intervals, supporting our findings, and we validated 95% of the inferred intervals by sequence analysis on the updated release of the dog assembly.

Multispecies comparative analysis should be improved by the use of additional species, including species more phylogenetically distant than chimpanzee and human or mouse and rat. The availability of high coverage (>6×) genome assemblies, such as those for cow, macaque, and opossum, will provide additional data that could be integrated into studies of this type. In contrast, the use of low-coverage (2×) sequence projects is of limited interest for this aspect of the topic, due to the lack of continuity in chromosomal assignation.

Our results demonstrate that the inference of canine chromosomal intervals for chrUn genes from a multispecies comparative genomics approach is efficient and can be rapidly achieved. Similar strategies could be used to localize chrUn genes of any species in mammals with complete sequence assemblies.

Supplementary Material

Supplementary Table 3 is available online at http://genoweb.univ-rennes1.fr/tom_dog/J_Hered_supplementary.html.

Acknowledgments

We thank the GenOuest Bioinformatics Platform for hosting the mySQL database and the AutoGRAPH server. We also thank the French Centre National de la Recherche Scientifique for supporting this work and the Conseil Regional de Bretagne for awarding a fellowship to T.D.

References

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, et al. 2005. The Vertebrate Genome Annotation, Vega database. Nucleic Acids Res. 33:D459–D465.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al. 2006. Ensembl 2006. Nucleic Acids Res. 36:D556–D661.

Chen Z-Q, Lautenberger JA, Lyons LA, McKenzie L, O'Brien SJ. 1999. A human genome map of comparative anchor tagged sequences. J Hered. 90:477–484.

Derrien T, Andre C, Galibert F, Hitte C. 2007. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics. 23(4):498–499.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, et al. 2004. Genome evolution in yeasts. Nature. 430:35–44.

Fischer G, Neuveglise C, Durrens P, Gaillardin C, Dujon B. 2001. Evolution of gene order in the genomes of two related yeast species. Genome Res. 11(12):2009–19.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493–521.

Goodstadt L, Ponting CP. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol. 2(9):e133.

Hardison RC. 2003. Comparative genomics. PLoS Biol. 2:E58.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695–716.

Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C, et al. 2005. Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. Nat Rev Genet. 8:643–648.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431:946–957.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent W.J. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 14(1):160–9.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, et al. 2003. The dog genome: survey sequencing and comparative analysis. Science. 301(5641):1898–1903.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Liang C, Dandekar T. 2006. InGeno—an integrated genome and ortholog viewer for improved genome to genome comparisons. BMC Bioinformatics. 7:461.

Lindblad-Toh K, Wade CM, Mikkelsen T, Karlsson E, Jaffe DB, Zody MC, Clamp M, Kamal M, Kulbokas EJ, Chang JL. et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 438:803–819.

Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC, Clamp M, NISC Comparative Sequencing Program. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci USA. 102:4795–4800.

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. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 309:613–617.

O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG. 1993. Anchored reference loci for comparative genome mapping in mammals. Nat Genet. 2:103–112.

Pan X, Stein L, Brendel V. 2005. SynBrowse: a synteny browser for comparative sequence analysis. Bioinformatics. 21:3461–3468.

Pevzner P, Tesler G. 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res. 13(1):37–45.

Quignon P, Giraud M, Rimbault M, Lavigne P, Tacher S, Morin E, Retout E, Valin AS, Lindblad-Toh K, Nicolas J, et al. 2005. Genome Biol. 6(10):R83.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J, Mural R. 2005. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics. 6:703–710.

Corresponding Editor: Urs Giger

2.2. Affiner l'annotation du génome du chien par une approche combinant synténie et alignements ciblés.

(Article 4 : Derrien et al., en préparation)

L'annotation en gènes codant pour des protéines dans le génome du chien (~19 000) s'est avérée être sensiblement inférieure aux autres génomes mammifères (~22 000 et 24 000, par exemple, pour l'homme et la souris, respectivement).
Plusieurs hypothèses peuvent être formulées pour expliquer ces différences. Pratiquement, ces observations sont basées sur le système d'annotation automatique standard d'Ensembl et, par conséquent, peuvent résulter de difficultés inhérentes au programme d'annotation. D'autre part, certains gènes ont pu être dupliqués dans les deux lignées rongeurs et primates aboutissant à des relations 1:0 avec le génome du chien. Enfin, il est possible que des gènes canins aient tellement divergé que leurs séquences nucléotidiques et protéiques ne permettent plus un alignement correct avec leurs orthologues murin et humain.

2.2.1. Objectifs - Résumé

Dans cette étude, nous nous sommes focalisés sur un sous-ensemble de gènes présents chez quatre espèces de référence (homme:chimpanzé:rat:souris) et a priori absents chez le chien (génome testé). Nous avons ainsi identifié 412 gènes en relation 1:1:1:1:0 (homme:chimpanzé:rat:souris:chien) pour lesquels nous avons combiné une analyse de conservation de l'ordre des gènes avec des alignements de séquences ciblées sur le génome testé. Nous confirmons l'existence de 285 nouveaux gènes parmi lesquels nous clarifions 100 nouvelles relations d'orthologie 1:1 entre le chien et les génomes référents. Pour 92 gènes qui demeurent sans prédiction, différents arguments impliquant des scénarios de pertes de gènes canins sont exposés à la fin de cette étude.

Par ailleurs, nous utiliserons le terme de gènes orphelins du chien pour désigner les 412 gènes présents chez les quatre génomes référents et absents chez le chien. En effet, le terme de "gène orphelin" possède un double sens qui, soit correspond à une région codante sans fonction, soit fait référence à une région codante sans correspondance dans d'autres génomes (Domazet-Loso *et al.*, 2003).

2.2.2. Résultats

• Analyse des 412 orthologues de référence

Préalablement à la recherche des 412 gènes dans le génome canin, nous avons entrepris une caractérisation structurale et fonctionnelle de ces gènes pour les quatre génomes de référence. Nous montrons que ces gènes sont significativement plus petits et possèdent un plus faible nombre d'exons que des gènes pris aléatoirement dans le génome (taille des transcrits : 4,6 vs 8,4 kb ; nombre d'exons : 5,8 vs. 9,8). De plus, l'annotation fonctionnelle de ces gènes par le programme GOTree Machine (Gene Ontology : Processus Biologique) révèle que les 412 gènes de référence appartiennent à des catégories fonctionnelles bien distinctes (**Table 4**) et caractéristiques de familles de gènes impliquées dans la détection et la réponse à des stimuli extérieurs, la régulation de la transcription ou encore la fertilisation.

Catégories GO	Probabilité critique (p-value)	catégories GO	Probabilité critique (p-value)
potassium ion transport	8.9E-3	fertilization	4.5E-4
response to wounding	8.9E-3	fertilization (sensu Metazoa)	4.0E-4
detection of chemical stimulus	8.8E-3	regulation of cellular physiological process	6.2E-05
physiological response to wounding	7.7E-3	regulation of physiological process	1.4E-05
plasma membrane fusion	6.9E-3	regulation of metabolism	1.7E-06
microtubule nucleation	6.9E-3	regulation of cellular metabolism	1.2E-06
fusion of sperm to egg plasma membrane	5.2E-3	regulation of transcription	6.8E-07
response to external stimulus	4.3E-3	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.7E-07
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	6.0E-4	transcription	5.0E-07
regulation of biological process	5.2E-4	regulation of transcription, DNA-dependent	4.3E-07
regulation of cellular process	4.9E-4	transcription, DNA-dependent	2.03E-07

Table 4 : Table des catégories GO sur-représentées dans les 412 gènes orthologues de référence.
Le nom de la catégorie est présentée dans la colonne de gauche et la probabilité critique (p-value) associée dans la colonne de droite.

• Carte de synténie et définition des intervalles consensuels orthologues

Nous avons construit les cartes de synténie entre entre les paires de génomes homme/chien, chimpanzé/chien, souris/chien et rat/chien et identifié l'ensemble des segments conservés ordonnés (CSOs) comme décrit dans l'article 3. Les 412 gènes orphelins du chien (gènes en relation 1:0) ont ensuite été localisés sur les cartes de synténie et leur distribution au sein des CSOs a été analysée. La taille moyenne des CSOs contenant un gène 1:0 s'est avérée plus grande que la taille moyenne de la totalité des CSOs (18,4 vs 12,6 Mb) suggérant la présence de famille de gènes dans ces CSOs. Pour chaque gène référent, l'analyse de l'ordre des gènes dans les CSOs a entraîné l'identification d'un intervalle orthologue sur le génome du chien. La multiplicité des génomes référents nous a permis d'identifier des intervalles consensuels ou COIL (Consensus Orthologous IntervaL) quand au moins deux génomes référents déterminaient une région canine se recouvrant (**Fig.20**). À partir des 412 gènes de référence, 389 COILs ont pu être identifiés sur le génome du chien alors que 17 gènes étaient localisés dans une rupture de synténie communes aux quatre espèces de référence et ne permettaient pas, par conséquent, d'identifier un intervalle canin unique (6 ne fournissant pas un intervalle consensuel). Il est à noter que cette proportion de gènes localisés en rupture de synténie partagée (3,4% 17/412) est significativement supérieure à une proportion de gènes "test" pris aléatoirement sur le génome (1,8%, $p < 0,02$). Il est alors possible que ces cassures chromosomiques mises en évidence par les ruptures de synténie entre les génomes référents et le chien soient aussi impliquées dans la perte de gènes chez le chien.

• Alignements des séquences de référence sur le génome du chien

Nous avons procédé à un alignement des séquences des 412 gènes de référence sur le génome canin. Puis, pour chaque gène, les localisations résultantes de ces alignements ont été comparées à la localisation des COILs précédemment définis. Au total 348/389 (89%) correspondances ont été trouvées entre les deux approches (COILs/alignements de séquences sur le génome du chien). De façon intéressante, 73/348 (~21%) des alignements chevauchants les COILs ne correspondaient pas aux meilleurs alignements issus du logiciel d'alignement utilisé (Exonerate - Slater *et al.*, 2005). Les meilleurs alignements pourraient correspondre à des pseudogènes rétro-transposé aléatoirement dans le génome et non au niveau de l'environnement synténique attendu (Goodstadt *et al.*, 2006).

• Prédiction de gènes canins dans les COILs

Le logiciel GeneWise (Birney *et al.*, 2004) a été utilisé pour la prédiction d'une structure de gènes canins dans les COILs. L'avantage de l'approche synténique nous a permis de délimiter la zone de

prédiction de gènes à un espace de recherche très réduit correspondant uniquement à la taille du COIL (~347 kb en moyenne). Au total, 285 gènes canins ont été identifiés dont l'identité de séquence protéique était supérieure à 40% avec les orthologues de référence. La comparaison avec les gènes canins préalablement annotés par Ensembl a révélé que 100/285 chevauchaient un gène canin annoté par Ensembl. Autrement dit, pour ces 100 gènes, une nouvelle relation d'orthologie non-définie par Ensembl est mise en évidence avec les gènes de référence. Pour les 185 gènes restant, soit la prédiction d'Ensembl pouvait être considéré comme douteuse (**Fig.21**), soit aucune correspondance avec des gènes canins annotés par Ensembl n'a été trouvée aboutissant à l'identification de nouveaux gènes canins et donc, de nouvelles relations d'orthologie avec les génomes référents. Enfin, nous montrons que 50% de ces 185 nouvelles prédictions s'alignent avec un ESTs et plus de 80% montrent un motif peptidique après alignements avec les séquences protéiques de la base de données InterPro.

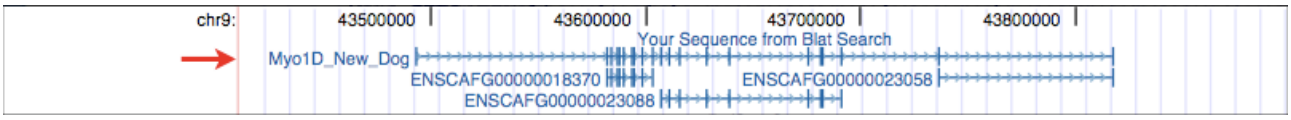


Fig.21 : Comparaison des prédictions de gènes avec les prédictions de gènes Ensembl (Captures d'écran à partir du Genome Browser UCSC (Kuhn *et al.*, 2007)).
Les gènes sont alignés sur la séquence génomique du chien (CanFam2.0) et sont représentés par des traits horizontaux bleu ciel. Les traits verticaux correspondent aux exons et les flèches à l'orientation du gène.
Le gène prédit par notre méthode est symbolisé par le premier trait horizontal bleu clair et identifié par la flèche rouge sur la figure.
Dans cet intervalle canin, Ensembl identifie trois gènes qui chevauchent notre prédiction et sont symbolisés par les identifiants ENSCAF*. Dans ce cas, le gène prédit est considéré comme étant un nouveau gène et définit donc une nouvelle relation d'orthologie avec les quatre gènes de référence.

- Caractérisation des COILs sans prédiction de gène canin

À partir de 389 COILs mis en évidence dans l'analyse synténique référent/chien, 285 nouveaux gènes canins et/ou nouvelles relations d'orthologie avec les génomes référents ont été identifiés. Par ailleurs, 104 gènes référents (389-285) demeurent en relation d'orthologie 1:1:1:1:0 avec le génome du chien. En effet, pour 41 gènes (39,4%) aucun alignement des séquences de transcrits de référence n'est en accord avec le COIL défini sur le génome du chien. Et 63 gènes canins définis par notre méthode montrent une identité protéique avec les orthologues de référence inférieure à 40% (dont les trois quarts ont une identité protéique inférieure à 25%). Par la suite, nous avons caractérisé, ces 104 intervalles canins afin d'identifier des caractéristiques de structure pouvant être corrélées à l'absence de prédiction de gène chez le chien.

Comme observé dans l'article 3 (Derrien *et al*., 2007), les 104 COILs montrent un taux de "gap" trois fois plus élevé que pour des intervalles pris aléatoirement dans le génome canin (4,4% vs 1,4%). L'analyse détaillée du contenu en "gap" de séquence a démontré que 12 intervalles contenaient plus de 10% de "gap" pouvant largement expliquer l'impossibilité de prédiction de gènes. De plus, 49% (51/104) des COILs se situent dans les régions télomériques des chromosomes canins illustrant ainsi les difficultés d'assemblage et d'annotations de séquences dans ces régions.
Par la suite, nous nous sommes focalisés sur un sous ensemble COILs pour lesquels la proportion de "gap" n'excédait pas 10% de la taille totale de la région. Une analyse des termes GO (Gene Ontology) des 92 gènes référents sélectionnés montre un enrichissement en gènes impliqués dans une fonction de "réponse à un stimulus extérieur" ($p < 0.001$), fonctions généralement assurées par des familles de gènes qui se prêtent à des événements d'extension ou de contraction. L'examen de la taille en paires de bases des 92 COILs a montré une taille significativement plus petite comparée aux intervalles "contrôles" (281 kb vs 375 kb). Deux hypothèses peuvent expliquer cette observation, l'absence de séquence du gène recherché où la présence de famille de gènes regroupées en cluster sur cette partie du génome qui permet de réduire les intervalles consensus.
Par ailleurs, nous avons caractérisé le contenu en GC des 92 intervalles et montrons que le taux de GC est supérieur à celui calculé pour des COILs pris aléatoirement dans le génome (47% vs 45%). Les régions riches en GC sont souvent associées à un taux de substitution plus élevée et corrèlent avec le modèle de la "conversion génique biaisée" (Biased Gene Conversion) lors de la recombinaison (Meunier *et al*., 2004 ; Galtier *et al*., 2007).

Dans cette étude, nous avons intégré l'analyse comparée de l'ordre des gènes entre plusieurs espèces et la prédiction de structures de gènes par analyse de séquences pour en déduire la position, la présence et l'annotation de gènes non préalablement annoté par le système standard d'Ensembl. À partir de 412 gènes présents en relation 1:1:1:1:0 entre quatre espèces de référence et le chien, nous apportons des arguments complémentaires pour mettre en évidence l'existence de 285 gènes chez le chien (dont 185 nouveaux gènes et 100 nouvelles relations d'orthologie). L'analyse des gènes non prédits est toujours en cours au laboratoire et différents mécanismes évolutifs sont suggérés mettant en relation la nature des gènes, la présence de famille de gènes et la composition en séquences pour expliquer cette perte de gènes chez le chien.

Article n°4

Refining the Dog Gene Repertoire using Multispecies Consensus Orthologous Intervals

Derrien T, Andre C, Galibert F, Hitte C.

(En préparation)

Multispecies Consensus Orthologous Intervals Search for Gene Predictions and Gene Losses in the Canine Genome

Derrien T. et al.

Summary

Although a number of species have been sequenced at high genome coverage, exhaustive identification of conserved and species-specific functional sequences remains a challenging task. Among mammals, the canine genome has an intriguing low number of annotated protein-coding genes (~17,500) compared to other mammalian species (> 20,000). We have analyzed the canine genome with objectives to specifically search genes presumably absent in dog but annotated in human, chimpanzee, mouse and rat genomes. We exploited the property of gene adjacency conservation between several species to infer canine Consensus Orthologous IntervaLs (COIL) on which were performed in-depth gene structure predictions. The species-specific rearrangements and the additive collective differences of one-to-one orthologue sets between species allowed to delimit short COILs reducing the search space for gene prediction from a genome-wide scale to a short targeted genomic interval of 350 kb, on average. Here, we show that accurate definition of COILs coupled with in-depth gene prediction enabled the identification of 285 new canine genes with structural and functional evidences. In contrast, for 92 genes, our results support scenario of gene losses in the canine species. For these 92 genes, we infered the length of time genes were functional before loss during the past 310 million years. This analysis enabled to differentiate canine gene loss events as losses of long-lived genes that remained functional long after their birth from lineage-specific gene gains in rodents and primates. Although comparative genomics analysis has been widely used in genome annotation, we applied it with a different scenario for optimizing gene prediction and as a strong predictor for detecting and characterizing losses in the dog genome.

INTRODUCTION

Comparative genomics plays a key role in understanding organisms evolution, refining functional annotation and identifying orthology relationships. Whole-genome sequence assemblies are available for many mammals with high coverage sequence and more recently for species from new clades such as turtle and lizard. These achievements allow to compare a newly sequenced genome with other species to study evolutionary rearrangements and to improve its genome annotation. Furthermore ongoing sequencing projects and sequencing technical development allowing with very high throughput will soon make available new additional resources from many model organisms commonly used as surrogates in human biology studies (http://www.genome.gov /10002154).

Computational approaches in the comparative genomics field, attempt to provide exhaustive and genome-scale predictions of functional sequence by exploiting sequence similarities (Twinscan, Korf *et al.* 2001; GENEWISE, Birney *et al et al* 2004; SGP2 Parra *et al.* 2003), phylogenetic analysis (Goodstadt *et al.*, 2006, Dufayard *et al.* 2005), evolutionary statistical model (Lunter *et al.*, 2006, Pollard *et al.*, 2007) or gene order

conservation information (Pevzner *et al.*, 2002 ; Derrien *et al.*, 2007). These comparative approaches are often complementary and have been developed to estimate and improve the identification of functional sequences for both newly sequenced species and reference species. Multispecies comparison is now used to refine protein-coding genes annotation for all the species of a given lineage using, as templates, the highest-quality gene sequence available from closely related species. This method has already been used successfully in predicting the protein-coding transcripts for each of eleven Drosophilia species, using D. melanogaster transcripts as templates (Heger *et al.*, 2007). Orthologues and paralogues are also distinguished using a combination of a well-suited phylogenetic approach (PhyOP pipeline, Goodstadt *et al.*, 2006 ; RAP Dufayard *et al.*, 2005) and the homologous synteny segment identification (AutoGRAPH, Derrien *et al.*, 2007). Multispecies genome scale comparison makes also possible to analyze the rates of duplication events for species-specific paralogues genes called in-paralogues (Sonnhammer *et al.*, 2002; O'Brien *et al.*, 2005) and examine which protein-coding genes are more prone to be duplicated. In-paralogues are investigated in order to better understand their fate, such as sub- or neo-functionalization and pseudogenezation processes and (Lynch *et al.*, 2000) and their connection to species-specific biological processes.

With the availability of numerous species sequences, it is becoming more and more relevant to gather all the datasets within fine-scale comparative maps constructed using orthologues. Such maps allow visualization and characterisation of conserved segments and gene order colinearity between the species. Conserved gene order between species maintained by natural selection suggests a functional significance with for example co-expressed genes, in particular housekeeping genes clustered along chromosomes more often than expected by chance. Alternatively, it has also been shown (Semon *et al.*, 2006) that constraints on gene orders have a limited impact on mammalian genome organization, and are the results of neutral evolutionary genome rearrangements. In addition, conserved gene order between species is exploited to identify potentially retrotransposed genes as these correspond mostly to pseudogenes inserted in non-syntenic chromosomal regions. As part of the characterization of gene content of a genome, it is tempting to take advantage of the conserved gene order between species as a strong predictor for gene structure identification and gene loss events investigation (Van Baren *et al.*, 2006).

In this study, we exploit the property of gene adjacency conservation throughout the separate evolution of related species coupled with targeted sequence alignments. We focused on a set of 412 genes that were not identified and presumably absent from the dog genome but annotated in four closely related mammals ; human, chimpanzee, mouse and rat. Using targeted gene prediction analysis on accurate orthologous intervals, we identified 285 new canine genes for which sequence identity with known proteins and expressed sequence tags (EST) validation strongly suggest an underestimate of the initial annotation process. The absence of gene structure prediction and functional evidences converged to support scenarios of gene losses in the canine species. For these losses, we inferred the length of time genes were functional before loss over the past 310 million years with the objectives to differentiate canine gene losses of long-established genes from recent gene gain acquired in rodent and primate lineages.

RESULTS

Structural and functional characterization of genes presumably absent in the dog genome

From all annotated genes of five mammalian species, human, chimpanzee, mouse, rat and dog, we extracted 412 genes and gathered them in a dataset we called the '412 set'. This 412 set consist of genes considered to be absent in the dog genome while annotated as protein-coding in the other four genomes and thus can be represented with a 1:1:1:1:0 orthology relationship pattern. We first examined the structural features of the 412 set and compared them to a random set of 400 genes independent from the 412 set. We analyzed gene and transcript sizes, the number of exons and the Guanine plus Cytosine (GC) content. Gene mean size of the 412 set was significantly smaller than the random set (Significance of differences was tested using *t* test; $P = 6.8\ 10^{-11}$; n= 400). The mean transcript size showed a similar result with a size ~50% smaller than determined in a random set ($P = 2.6\ 10^{-9}$) (Table1). The mean size of transcripts for human and mouse genes was significantly higher than those of chimpanzee and rat ($P = 5.0\ 10^{-3}$ and $P = 8.2\ 10^{-8}$, respectively). This may be explained by a lower quality annotation of the rat and chimpanzee transcripts due to the existence of alternative splicing events less investigated in these last two species. Mean number of exons was smaller (5.8 vs 9.8 ; $P = 3.7\ 10^{-13}$) than randomly expected. Single-exon genes were also found over-represented in comparison to that of the random set expectation (+15%). To ensure that single-exon genes in the four reference species were functional and not processed pseudogenes, we determined that they did not have accumulated degenerative mutations and EST were found for all of them. Finally, we observed a higher GC content in the 412 set (for human) compared to the random expectation set. GC content showed a strong negative correlation with the size of the gene (Spearman $\rho = -0.591$; $p = 3.8\ 10^{-10}$) and the size of the transcript (Spearman $\rho = -0.25$; $P = 3.16\ 10^{-7}$).

To test the underlying assumption that the structure features found in the 412 set may be implied in particular biological pathways, we examined their functional annotation in the context of gene ontology. Functional annotation by GO terms (Gene Ontology Tree Machine tool ; Zhang B. et al. 2004) showed that the 412 set is enriched in genes implicated in physiological pathways of immunity and organism responses to pathogens (12 genes), olfaction (16) and regulation of transcription (63) (**Table 2)** 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 or gain of genes (Goodstadt *et al.*, 2006).

Constructing synteny maps with 1:1 orthologs

Of the annotated gene set from Ensembl (v42), we extracted pairwise sets of 14,997 ; 14,798 ; 14,667 and 14,065 one-to-one protein-coding orthologues between human-dog, (H-D) chimpanzee-dog (C-D), mouse-dog (M-D) and rat-dog (R-D), respectively. Using those 1:1 orthologues as comparative anchors, we built

four fine-scale pairwise synteny maps (fig 1 and table 3) using the program AutoGRAPH, we recently developped to automate the construction of multispecies comparative maps. AutoGRAPH was designed to allow micro-rearrangements in order to relaxe the rigidity of conserved segments ordered (CSO) definition ie chromosomal segments that contains markers in the same order. We set to three the number of markers that was allowed to account for an interruption of colinearity. We identified 218, 229, 326 and 325 CSO, between H-D, C-D, M-D and R-D respectively, with a mean distance between consecutive genes of ~180 kb. Mean size of primate-dog CSOs was 12 Mb (range 0.02 Mb – 154 Mb) comprising on average 68 genes while the mean size of rodent-dog CSOs was 7 Mb (range 0.02 Mb – 55 Mb) comprising 40 genes on average. The higher number of CSOs identified between the rodent and dog genomes compared to primates and dog reflects the high level of chromosomal rearrangements in the rodent lineage (Murphy et al.). In all synteny maps, CSOs cover almost 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 (Pevzner and Tesler, 2003).

Mapping orphan genes on synteny maps

The 412 set comprises genes that are in a 1:0 orthology relationship relative to the dog which we refer to as ‘orphan genes’ in this study. In each pairwise synteny maps, we mapped the orphan genes on the reference genome. Of the 412 orphan genes, 92.3% mapped within CSOs while 7.7% mapped in breakpoints (table 4). Although orphan genes spread on all chromosomes for all reference species, their distribution greatly varied per chromosome (1 to 44 per human -HSA-chromosome in the example of the human-dog synteny map). The mean size of CSOs containing orphan genes is higher than the mean size of all CSOs (18.44 Mb vs 12.58 Mb ; P = 6.7 10^{-3}) in the example of the human-dog map) (table 4). This difference is possibly explained by gene clusters as for instance the 44 members of the KRAB zinc finger gene family that reside on a CSO between HSA19 and CFA1.

As the method we employed for detecting CSO is capable of relaxing the gene colinearity common to both genomes, we assessed the magnitude of gene order conservation within CSOs. We determined the rate of strict colinearity, corresponding to the percentage of genes in the same order in the tested (dog) and the reference species. Mean colinearity rate was 95.1% (range 94.8 to 95.7% for the four pairwise synteny maps) underlying that conserved 1:1 gene order within CSOs is a rule that is rarely violated.

Reducing genome to short Consensus Ortholog IntervaLs (COILs)

A significant contribution of synteny maps is that orthologues within CSO provide focal points of homology between species and provide landmarks for cross-genome navigation. As a consequence, the localization of un-annotated gene can be inferred from gene order conservation and may considerably help the prediction of gene structure. We used multiple closely related species to identify in the dog genome, short Consensus Orthologous IntervaL (COIL) delimited by the closest flanking 1:1 orthologs on either side the tested gene. Using multiple pairwise maps enabled to shorten COILs size and therefore reduced genomic

interval to search for gene prediction.

To assess the performance of COILs identification and the probability that a COIL contains the queried gene, we extracted 12,179 orthologs with a 1:1:1:1:1 orthology pattern between the five species studied and we masked the canine gene localization to simulate dog orphaned genes. Using the four pairwise synteny maps, we identified 11,934 (98.0%) canine COILs from at least two reference species (89,2% were identified from the four species). A total of 222 (1.8%) COIL localized in breakpoints, preventing their assignment to a single interval on a single canine chromosome and 23 (0.2%) did not exhibit a consensus orthologue interval in at least two reference species. Next, we determined that COILs could be used to infer close to perfection the localization of the simulated orphaned canine gene. Indeed, 99% of COILs (n=11,853) inferred on the canine genome included the expected dog gene. The remaining 1% that did not match the expected gene localization may be explained by short micro-arrangements interrupting colinearity or reflecting inaccuracy in the assembly or in the annotation of the gene in the reference or dog genome sequence.

We applied the conserved gene order rule to infer on the canine chromosome the shortest canine COILs potentially containing un-annotated canine orthologs of the 412 set. We identified 389 COILs (94.4%) with a mean size of 347 kb that localized in large conserved segments while 17 COILs (4.1%) localized in common breakpoint regions (ie overlapping between at least two species) (Murphy et al, Alekseyev and Pevzner PLos Comp. 2007). For six orphan genes (1.4%), no COIL could be determined because of the absence of consensus interval.

Genome-scale sequence analysis

Using the transcript sequences of each reference gene (412 from each species), we aligned mRNA against the canine genome sequence (CanFam2) with the Exonerate program (Slater *et al*., 2005). Matches were obtained for 389 genes (94.4%) of which we retained the best five matching sequence to relaxe the limitations of conventional best-match methods (Van Baren *et al*., 2006). Indeed, considering only the best ranked match may lead to gene predictions that are impaired by the presence of processed pseudogenes, intronless copies of genes or relics of reverse transcription. Using this analysis, we compared the sequence alignments localization with the expected canine COILs position. Localization were highly concordant for 348 sequence alignments (89.4% of the comparisons) for which at least two reference species matched the inferred COILs. Concordance was defined when matching sequences were totally embedded in COILs, whereas partially overlap were not retained. Interestingly, 6.4% (n=22) of the alignments concordant with COIL correspond to matches that ranked from second to five highest (but not first) for any reference sequence queries. These were carefully investigated in further sequence analysis.

COIL-targeted gene structure prediction

Within each canine COIL, we used GeneWise program (Birney *et al*., 2004) to splice and align protein sequence of each reference species to predict dog gene structure. We retained gene predictions produced by

at least two reference species protein templates. As a result, GeneWise produced 285 gene structure predictions with amino acid identity > 40%, 63 with protein identity < 40% and 41 for which no gene structure could be obtained. A subset of 53 genes (18,6%) were predicted using only rodent protein sequence as templates 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 prediction and 75% exhibit at least a canonic splice site (GT/AG). Twelve gene predictions showed ORF-disrupting mutations consisting of at least one stop codon and one frameshift in coding sequence raising the possibility of pseudogenes. ORF-disrupted predictions are made possible by GeneWise which has a tolerance for frameshits allowing for assembly or sequencing errors.

Next, we identified sequence alignment between gene predictions and canine ESTs (dbEST -may 2007-) in 47.7% of cases with on average seven ESTs per gene prediction (range 1-84; sequence similarity > 96% on at least 150 bp). In addition, we searched for cross-species sequence alignment between reference species mRNA and canine ESTs and identified 33% of significant matches with conservative criteria (sequence similarity > 90% ; overlap > 150 bp). We next aligned all canine ESTs that matched gene predictions to the canine genome sequence, and found that 87% resided in the expected COIL. It is worth noting that we validated 19/22 gene predictions with canine ESTs for which cross-species sequence alignment initially ranked from second to five highest. These data confirm the power of conserved synteny to control for impaired alignment with the best-match that could lead to annotate potential pseudogenes. A protein motif was found from the InterPro database for 80.5% of the gene prediction providing additional evidences of gene identification. For the category of the 63 gene predictions, with protein similarity <40%, we obtained cross-species alignment between reference species mRNA and canine ESTs for 9.5% only.

Although we analyzed a dataset of genes initially not identified in the dog Ensembl database, we tested for possible overlaps between gene predictions found in this study and canine genes referenced in Ensembl. We aligned the 285 predicted protein sequences with all canine protein sequences collated from Ensembl (v42) and found that 65% (n = 185) of the predicted genes did not overlap any Ensembl annotated protein-coding genes (**Fig. 3**). This analysis supports the identification of new canine genes, which we classify as definite gene identification along with the delineation of new orthology relationships with the four reference species.

For the remaining 35% (n=100), we determined a match with an annotated gene (protein identity > 75%) indicating dubious orthology relationships previously reported by Ensembl. As we used protein-based alignment approach, we rule out the possibility of nested genes partially or totally embedded in introns and functionally related genes that could be closely located (Peng Yu genomics 2005). For those 100 gene predictions, we refine new orthology relationships between the four reference species and dog to a 1:1:1:1:1 pattern. Interestingly, these 100 genes showed a higher tendency to belong to gene family (79%) than a random set (n=100, 69%) that may explain their dubious orthology previously reported.

Are there evidences of gene losses in the canine genome ?

For 63 genes we found canine gene predictions with weak protein identity with any of the four reference

genomes (mean of 23%, and 75% showed protein identity < 20%) that contrast with the mean percentage protein identity within the four reference genomes (mean of 63.5% between human and mouse). These predictions were considered as candidate for gene losses if ORF disrupting mutations such as stop-codon or frameshift were detected or when no significant match (sequence similarity < 80% and overlap < 150 bp) between ESTs and the gene prediction could be detected. Based on these criteria, 50 prediction status were identified as potential canine losses. In addition, for 41 genes, no sequence matches were obtained by Exonerate, and no gene prediction could be achieved with Genewise. These were classified as genes absent in the dog genome.

To eliminate trivial artefact such as gap sequence, we search for gap content in the 41 COILs that did not permit gene prediction. We calculated the cumulative size of gaps, represented as stretches of 'N' in the sequence (Karolchik *et al.*, 2004; http://genome.ucsc.edu/cgi-bin/hgTables). For six COILs (14.6% of the 41 COILs) gap sequences were 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%). Gap content analysis for the 63 COILs (protein identity < 40%) identified six COILs with significant gap sequence (>10%) all corresponding to weak gene predictions (average protein identity = 22%). Significant gap sequence likely prevented from accurate gene prediction and may largely explain the lack of canine gene identification. However, gap content did not explain the absence of gene predictions for 92/104 that remain candidate for genes loss in dog. For subsequent analysis, we focused on the subset of 92 COILs. To further assess gene loss scenario, we examined GC and repeat content, chromosomal localization, and mean size of the 92 COILs.

Examination of repeat element content (SINE, LINE, LTR) (Kirkness *et al.*, 2005) did not show an enrichment compared to the repeat content of a random set (n = 1000) of canine intervals (fraction of repeat content = 35.3% vs 35.8% for the random set ; $P > 0.4$). GC content distribution among the 92 COILs showed no significant increase in GC base composition compared to random expectation (47 vs 45.6% ; $P > 0.09$). We determined that approximatly 50% of the 92 COILs localized on canine chromosome-ends encompassing telomeric and sub-telomeric regions. Mean size of the 92 COILs (281 kb) was strikingly lower than COILs for which we could predict a gene with protein identity > 40% (347 kb) raising the possibility of the absence of gene sequence. Such features, weak cross-species sequence similarity, ORF-disrupted gene predictions and telomeric localization provide pattern of gene loss scenario as described in 'birth and death' evolutionary models (Nei *et al.*, 1997).

Inferring the length of time genes were functional before loss

To estimate the length of time genes were functional before loss in dog, we determined the interval between gene birth and death in dog branch. The timing of the gene birth was defined by finding a lower bound timing birth event during the past 310 Mya. We identified the 1:1 functional orthologues (Ensembl v46) for the 92 genes lost in the canine genome, in nine species that represent the mammalian tree. In addition to human, chimp, mouse and rat genome sequences, we used scaffold assemblies of elephant, tenrec and armadillo from the Afrotheria and Xenarthra superorder and two non-placental genome sequences of monodelphis and platypus. We included the chicken genome sequence in the analysis to infer gene birth that occurred further back in the evolutionary past down to 310 Mya. Using this method, we were able to infer

the timing of the gene birth for 83/92 genes from which we determined gene birth events prior to the chicken branch (before 310 Mya) for 26.5% of the genes, prior to opossum/platypus (before 140 Mya) for 22%, prior to elephant-tenrec or armadillo (before 100 Mya) and prior to rodent (before 60 Mya) for 14.5%.

Loss of gene long after their birth vs recent gene gain

A direct contribution of this analysis is that we could differentiate gene birth events that occurred after the canine branch from gene births that occurred long before the dog branch. Using a threshold of 140 Mya for gene birth (before opossum/platypus), we classified the loss events as gene losses we refer to as 'long-lived' genes, and using a threshold of 60 Mya, (before rodent but after dog) gene absences were classified as recent birth-induced genes in rodent and primates lineages. Interestingly, 14.5% (n=12) of the gene absences in dog are classified as birth-induced genes in rodent and primate lineages, which is 10-fold higher than for a random set analysis (n=1000, 1.5%).

If gene birth occurred after dog branch and before the last common ancestor of rodent and primates (~60 Mya), it suggests either lineage-specific retrotransposition events that reinserted at any location in the genome a functional copy of the parental gene or lineage-specific segmental or tandem duplication of genes resulting in in-paralogues. If in-paralogues occurred in euarchontoglires (rodent and primates), they can be detected by the presence of a parental gene and its duplicate. Among the 12 birth-induced canine gene absences we found five genes with in-paralogues (Ensembl v46) in both human and mouse species and for four of them (HLA-SX-alpha, ZNF524, PNMA5, PNMA3), in-paralogues located within 1Mb suggesting tandem duplications. In seven cases, there are no annotated paralogues, suggesting other mecanisms that had formed new genes in rodent and primate lineages.

Among the canine gene losses, we identified 50.5% (n=42) as long-lived losses with 26.5% (n=22) being functional for more than 310 Mya. Of these 22 genes, nine are members of large family such as Serpins whose members are organized in clusters making difficult the unambiguous identification of orthologues. Therefore, focusing on the remaining 13 genes, we identified three with no biological function characterized in human and 10 with functional annotations. Canine gene losses include *PROZ*, a vitamine K-dependent protein Z precursor involved in blood coagulation pathway, *FANCF*, a anemia group protein whose defect causes chromosome breakage and chromosomal instability, *EDPDR-1* a ependym-related protein for which three pseudogenes have been identified in human (Apoustolopoulos *et al*., 2001).

DISCUSSION

This study presents a multispecies comparative genomics approach that improve gene prediction and alternatively gene loss detection through the identification of targeted short genomic intervals. A basic application of gene order based approach is the capacity to detect conserved genomic context that allow to reduce the space search for gene prediction to a short genomic interval such as COIL (mean size ~350 kb). Such targeted interval allow to balance sequence alignment sensitivity against accuracy. Multispecies pairwise orthologous interval approach illustrates the complementary contributions of different genomes to

reduce the size of COILs. Furthermore, by using conserved genomic context, detection of all the real exons is restricted to a genome sub-region that greatly reduces the cost of detecting false-positives. Alternatively, similarity for more divergent sequence may be relaxed while search is improved in short and pre-defined space search such as COILs as the background noise is significantly reduced compared to a genome scale search (Zhang *et al.,* 1998). In addition to obvious evolutionary rearrangements that disrupt gene colinearity, evolutionary forces that shape genomes organization, such as, recombination and gene conversion may limit the utility of gene order for gene prediction. However by increasing the number of species to be compared we greatly reduce the probability of such events. Although a minority of genes may not be in conserved order, if they exist they should be covered by alignments with sufficient similarity.

We found 185 genes that could be predicted with conserved structure and functional evidences. Examples include the prediction of several conserved functions, such as transcrition factors (SOX14 ,SOX18), splicing factor (SF3B5), apoptosis regulation patways (TRIAP1). Our analysis may miss gene-prediction (false-negative) because of high divergence at the protein or nucleotides level. However, in this study we investigated genes with comparable evolutionary rate in the four species used as references. While scenarios of genes with different evolutionary rate and fate in different species exist, it can be legitimately postulated that using genes with one-to-one orthology relationship between several reference species (1:1:1:1 pattern) limit the possibility that a gene evolves neutrally along one lineage. As a consequence, while no definitive prediction can be made on the basis of one-to-one orthology and gene order conservation, multispecies approach for genes present in four species such as human, chimpanzee, mouse and rat is expected to be more powerful for structure prediction in closely related species.

Alternatively, this analysis may contains false positive gene predictions arising through several factors that can lead to errors including poor sequence alignement, gene model used in gene prediction analysis. In addition, the absence of experimental validation for all predictions, as no EST matches could be determined for each of them, can be considered as a vulnerability of the evaluation procedure. However, no EST validation can be due to problems inherent to EST quality such as chimerism or poor sequence, to rarely expressed genes for which it is unlikely to detect EST and to the paucity of available canine ESTs as compared to human DB_EST. The use of cross-species sequence comparison helped to make the analysis less prone to produce false postives. In this study, we identified 100 predictions for which we provide a major clarification of the orthology relationships that, in majority belong to gene families. This can be exemplified by gene members that belong to kallikrein (n=4), to zinc finger (n=53) or to olfactory receptors (n=8) large gene-family. These predictions emphasise that gene family orthology still remains to be uncovered, and conserved gene order between genomes is a strong predictor to infer orthology relationships.

We showed that the identification of COILs from multiple species targets efficiency intervals for in-depth gene predictions. A corollary is that the absence of gene prediction within targeted intervals can be a strong predictor of gene losses for the queried genes. Gene losses arise through retrotransposition events, segmental or tandem duplications (one copy becomes inactivated by mutations due to functional redundancy) or degenerative mutations causing non-functionality. It is expected to find genomic signatures left behind retrotransposition and duplication scenario, such as traces of alignment and conserved exon-intron structure with the parental gene, respectively. Among the 92 gene losses candidates, we inferred a strikingly high

number (14.5% vs 1.5% for a random set n=1000) of gene birth events that occurred prior to euarchontoglires (rodent and primate lineages) but after the canine branch. In addition, we determined that these genes comprise fewer exons (mean of 3.3). These results suggest that lineage-specific genes in-paralogues were gained in rodent and primates through tandem or segmental duplication mechanisms rather than truly lost in dog. These results are in agreement with recent studies (Goodstadt *et al.,* 2006, PhyOP) that identified 3-fold more gene duplications in the human genome than in dog genome and that in-paralogues are generally shorter and have fewer exons than 1:1 orthologues.

In contrast to gene gain, we identified 42 long-lived losses as genes are functional for more than 140 Mya from platypus to human but lost in the dog. It is intuitive to consider losses of long-lived gene in a given species as an adaptive event that may confer selective advantages to an organism (Zhu *et al*., 2007). Adaptive losses (ocurring ~85 Mya ago for dog) are expected to leave genomic signatures such as poor sequence conservation due to sequence degeneration and consequently altered exon-intron structure causing non-functionality. Such signatures allow to differentiate gene-losses from genes under positive selection or under relaxation of evolutionary constraints that remain functional. On this basis, we defined the 42 long-lived losses for which ORF-disrupting sequence mutation and absence of EST validation were identified. It is worth wondering which selective advantage could have confered gene losses in dog. Among the gene loss we identifed, *PROZ*, a vitamine K-dependent protein Z precursor gene loss is correlated to blood coagulation pathway. Mammalian blood coagulation is initiated and regulated by a complex network of interactions in which it has been suggested that genes have evolved in new function (Davidson C. et al 2003, Jiang Yong 2003). Gene-loss of *FANC* a gene involved in anemia physiological pathway is related to cellular response to external stimulus. Although it is difficult to assess the precise impact of gene loss, adaptation can arise at a severe cost for the fitness of a species and therefore it is reasonable to predict that small sets of such costly adaptation will be few among species.

Our analysis does obsviously not provide an exhaustive list of gene losses of the canine species for the past 140 Mya. Gene predictions inaccuracies may be generated by sequence artefacts although we filtered this potential problem by analysing the gap content. Validation of candidate losses by the absence of canine EST match is not a definite evidence as canine ESTs are far from representative of the complete dog transcriptome. Therefore, we used cross-species EST alignments (gene predictions were aligned against human ESTs) to further validate our findings, and we determined a very low number of matches (9% that can come from alignment with close homologues) as compared to 33% for the set of 185 new gene predictions.

A rough correlation exists between gene repertoire conservation and evolutionary distances between species However, on the basis of co-occurrence of orthologue genes, only for ~70% of the gene repertoires a one-to-one correspondence among mammals is shown. Orthology relationships remain to be deciphered for gene family, genes with divergent sequence and species-specific sets of genes that have emerged through evolution. We show in this study that multispecies gene-order-based approach is efficient to infer targeted genomics interval. Combination of multispecies comparative genomics, improved gene prediction, accurate consideration of phylogenetic relationship and timing of gene births events help to better predict gene

structure and gene losses in newly sequenced genomes and, in turn, benefit to reference genomes.

METHODS

Gene data sets and synteny maps

We use Biomart version 0.5 (Ensembl v.42) to collect orthologous protein-coding genes from the five genomes of interest: Human (NCBI 36) - Chimp (Chimp 2.1) - Dog (CanFam 2.0) - Mouse (NCBI m36) - Rat (RGSC 3.4). Ensembl Gene Id, orthology relationship, locations in base pair on each species-chromosomes were downloaded and inserted in a MySQL database. We extracted 412 protein-coding genes unannotated on the dog genome (1:1:1:1:0 Human:Chimp:Mouse:Rat:Dog).

We used AutoGRAPH (Derrien et al., 2007) to construc pairwise synteny maps between reference genomes and tested genome. In brief, AutoGRAPH has been designed to construct synteny 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 can then be identified with respect to the ranking order. We only considered CS and CSO containing a minimum of three genes. AutoGRAPH inferred the colinearity rate within CS and CSO corresponding to the longest increasing gene order sequence between the two species divided by the size of the CS/CSO. We discarded CS/CSO that had a colinearity rate less than 0.5 meaning.

Reference species cDNA alignments

Reference mRNA were collated from Ensembl (v.42). and aligned on the dog genome using Exonerate v1.2 (Slater et al., 2005) with a minimum perfect match of 18 bases (in order to increase speed) to begin an alignment (--dnawordlen 18) and retained the 5 best alignments (--bestn 5).

Gene prediction with GeneWise

The GeneWise program (Birney et al., 2004) was used with default parameters to align each reference protein on the dog COIL sequence as described by Curwen et al. Prediction with the highest GeneWise score and with a minimum of 40% identity with the reference protein was retained.

New Orthology vs. New genes

Dog protein inferred from the gene prediction analyses were aligned to all canine transcripts as defined by Ensembl.

EST analysis

Dog predicted protein were aligned on canine DB_EST (est.fa 05/19/07 from UCSC download browser) using exonerate (--model protein2genome).

GO annotation

The Gene Ontology Tree Machine (GOTM) (Zhang et al., 2004) was used to retrieve GO term associated with ensembl gene identifiants. An 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 overepresented ($p < 10^{-3}$) were considered.

Statistical tests

We used the R package (R Development Core Team 2006) to test the statistical significance in comparing distinct distributions at each step of the method (Mann-Withney, chi square and Student's test).

REFERENCES

- Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004 May;14(5):988-95.PMID: 15123596

- Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. 2005 Dec;21(12):673-82. Epub 2005 Oct 19. Review. PMID: 16242204

- 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 Jun 1;21(11): 2596-603. PMID: 15713731

- Galtier N, Duret L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 2007 Jun;23(6):273-7. Epub 2007 Apr 5. Review. PMID: 17418442

- Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol. 2006 Sep 29;2(9):e133. PMID: 17009864

- Heger A, Ponting CP. Evolutionary rate analyses of orthologues and paralogues from twelve Drosophila genomes. 2007 (submit)

- Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000 Nov 10;290(5494): 1151-5. PMID: 11073452.

- Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers SN, Milan D, Ostrander EA, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'brien SJ, Pevzner PA, Lewin HA. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 2005 Jul 22;309(5734):613-7. PMID: 16040707

- Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997 Jul 22;94(15):7799-806. PMID: 9223266

- O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D476-80. PMID: 15608241

- Semon M, Duret L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 2006 Sep;23(9):1715-23. Epub 2006 Jun 6. PMID: 16757654

- Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005 Feb 15;6:31. PMID: 15713233

- Van Baren MJ, Brent MR. Iterative gene prediction and pseudogene removal improves genome annotation. Genome

Res. 2006 May;16(5):678-85.PMID: 16651666

- Wang W, Kirkness EF. Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res. 2005 Dec;15(12):1798-808. PMID: 16339378

- Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle T, et al. . Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage. PLoS Computational Biology. 2007. Vol. 2, No. 5, e47

TABLES AND FIGURES:

Table 1:

(1) Reference gene and transcript (2) sizes are significantly smaller in orphan set than in a random expectation set.

Reference GC content in genes (3) and in transcripts (4) are higher in orphans than in random set.

(5) The number of exon per gene are significantly smaller in orphan than in the random set.

	Tested set (n=412)				Random expectation set (n=400)			
Reference species	**hsa**	**ptr**	**mmu**	**rno**	**hsa**	**ptr**	**mmu**	**rno**
(1) Size gene (bp)	31015	30389	26385	25758	65396	65173	49217	47044
(2) Size cDNA (bp)	4699	1973	2099	1613	8421	3098	3117	2480
(3) GC gene (%)	49.78	49.89	48.60	49.34	46.77	46.80	46.68	46.96
(4) GC transcripts (%)	52.55	53.00	52.11	53.62	51.35	51.64	51.52	52.53
No of transcript/ gene	1.78	-	1.30	-	2.05	-	1.43	-
(5) No of exon(s)/ gene	5.79	-	5.77	-	9.84	-	9.74	-

Table 2:

GO categories overrepresented in the 412 starting set.

p-value	GO category (BP)	p-value	GO category (BP)
0.008970025556679	potassium ion transport	0.00045451746485333	fertilization
0.008902320120899	response to wounding	0.00040299007694943	fertilization (sensu Metazoa)
0.0088375773202095	detection of chemical stimulus	6.2941602638017E-05	regulation of cellular physiological process

p-value	GO category (BP)	p-value	GO category (BP)
0.0070783033944581	physiological response to wounding	1.4569323904298E-05	regulation of physiological process
0.0069479680074864	plasma membrane fusion	1.7202210352273E-06	regulation of metabolism
0.0069479680074864	microtubule nucleation	1.2026321510082E-06	regulation of cellular metabolism
0.0052673840107958	fusion of sperm to egg plasma membrane	6.8760533901414E-07	regulation of transcription
0.004307225374419	response to external stimulus	5.7435783302335E-07	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
0.00060351043892835	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.0349916873545E-07	transcription
0.00052973081439737	regulation of biological process	4.3662427217568E-07	regulation of transcription, DNA-dependent
0.00049971977537453	regulation of cellular process	2.0446677466923E-07	transcription, DNA-dependent

Table 3:
Characterization of Conserved Segments Ordered (CSO) identified by AutoGRAPH and based on the four pairwise synteny maps between reference species and the dog genome.

Reference:Tested Species	No of 1:1 orthologues identified	No Conserved segments (CS/CSO)	CSO Min sizes (number of ortholog : size (Mb)	CSO Max sizes (number of ortholog : size (Mb)	CSO Mean sizes (number of ortholog : size (Mb)	Colinearity
Human:Dog	14 997	218	3 : 0,04	515 : 154,63	68 : 12,58	95,6%
Chimp:Dog	14798	229	3 : 0,02	459 : 152,8	62 : 11,93	94,8%
Mouse:Dog	14667	326	3 : 0,02	452 : 51,84	44 : 7,19	95,7%
Rat:Dog	14065	325	3 : 0,03	326 : 55,26	42 : 7,47	94,6%

Table 4:
Characterization of Conserved Segments Ordered (CSO) that include/contain reference gene in 1:0 orthology relationships with the dog genome.

Reference Species	Gene in CS : Gene in Bkpt	Uniq (CS/ CSO)	CSO Min sizes (number of ortholog : size (Mb)	CSO Max sizes (number of ortholog : size (Mb)	CSO Mean sizes (number of ortholog : size (Mb)	Colinearity
Human	391:21	120	5 : 0,28	515 : 154,63	103:18,44 - (/120) 184:30,87- (/391)	96,1%
Chimp	379:21 (+12chr_random unanalysed)	122	5 : 0,31	459 : 152,8	96:18,00 - (/122) 172:29,93 - (/379)	95,4%
Mouse	376:36	150	4 : 0,18	452 or 51,84	72:11,46 - (/150) 108:15,92- (/376)	95,8%
Rat	375:37	159	3 or 0,13	326 or 55,26	65:11,35 - (/159) 95:14,74- (/375)	94,3%

Table 5:
Characterization of canine intervals at each step of the workflow.

	Mean Size (bp)	Fraction of GAP content (%)	Fraction of repeat content (%)	Fraction of GC content (%)	Fraction of gene in telomeric region (%)
412	None dog interval predicted				
23	None dog interval predicted				
389	347,401	2.23	37.33	46.23	44.4
41	231,900	5.96	35.24	48.86	53.6
348	361,008	1.79	37.57	45.92	43.4
63	287,821	3.35	35.36	48.09	46.0
285	377,187	1.44	38.06	45.43	42.8
Random set 1000	375,929	1.3233	35.8865	45.63	31.0

FIGURES

Figure 1:

Comparative method that allows the definition of canine interval based on colinearity within CSO from pairwise synteny maps. Example of the BCLAP gene. (1) Collecting orthologs localisation for each reference species. (2) Constructing pairwise comparative maps between reference species and the dog genome. (3) Mapping 1:0 orthology relationships on the comparative maps. (4) Defining a consensual canine interval on the dog genome.

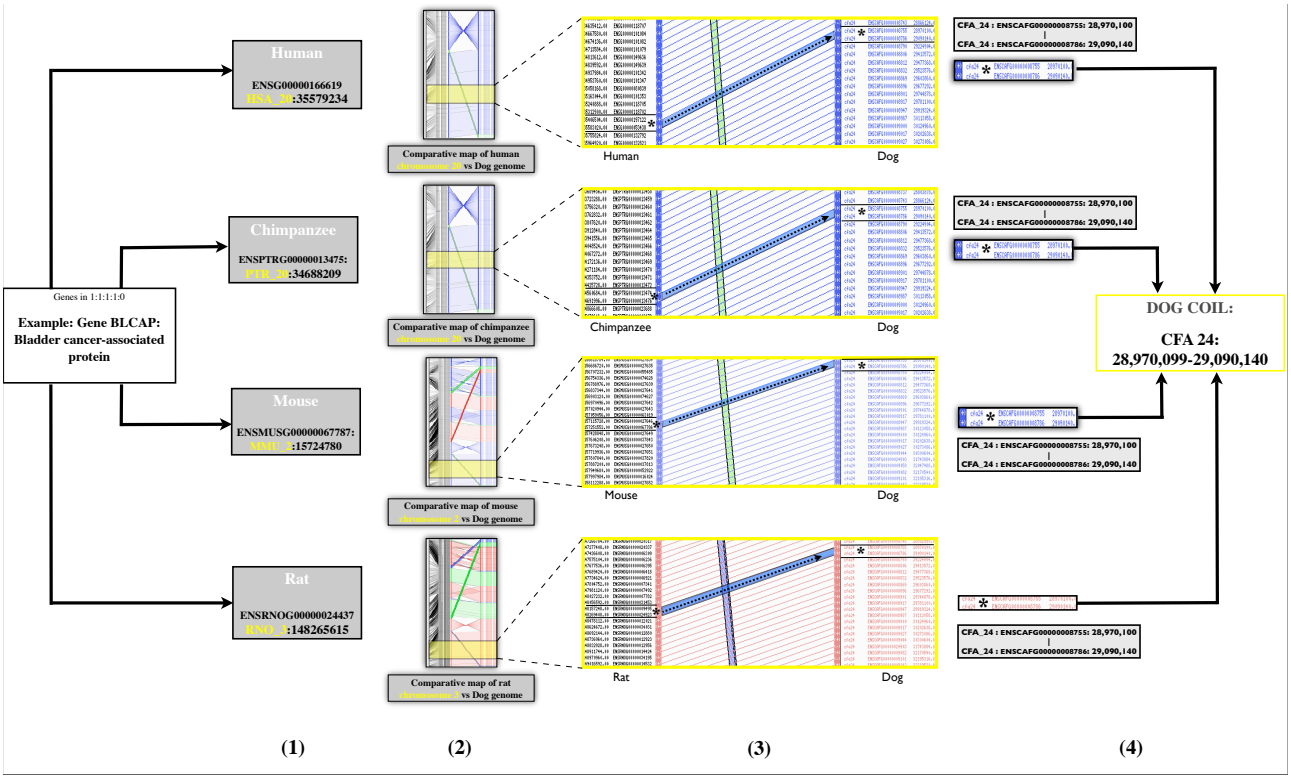


Figure 2: Comparative map of the chromosome X between the dog in the middle, human on the right and mouse on the left. Comparative anchors between dog and human are colinear all along the chromosome X and lead to the identification of ust one CSO. Conversely, comparative map between dog and mouse leads to the identification of eight large-scale rearrangements.

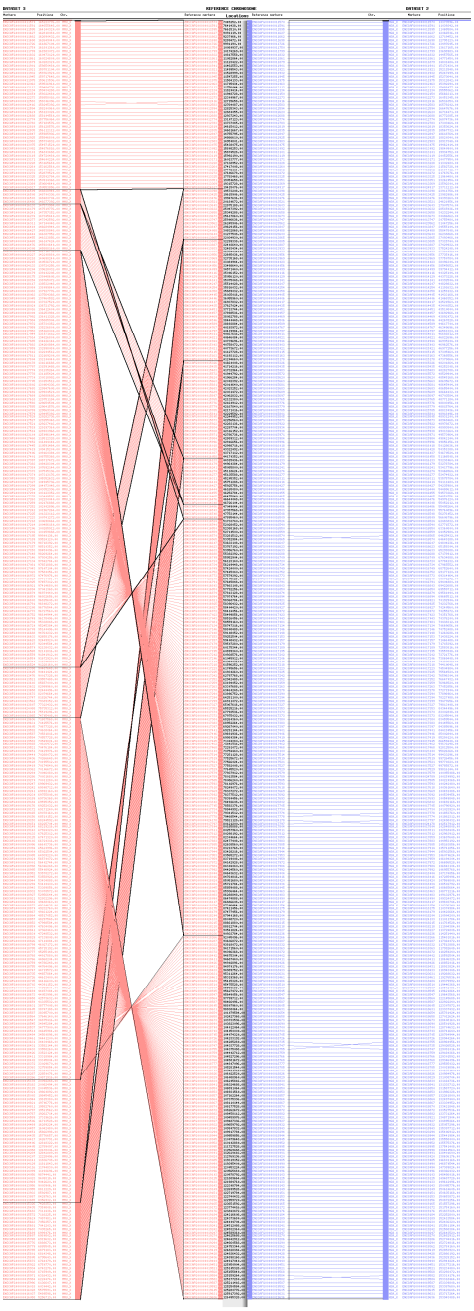
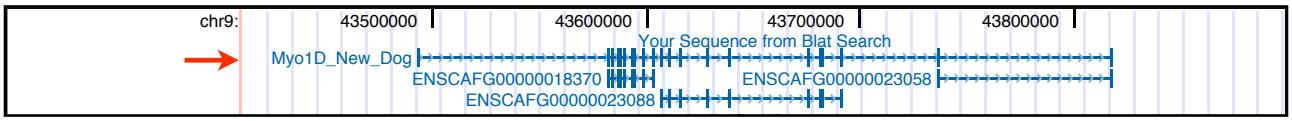


Figure 3:

Screenshot from UCSC genome browser of the dog sequence assembly. Ensembl gene Myo1D has been identified in the four reference species as one-to-one orthologues. Our new dog gene prediction (red arrow) overlaped 3 Ensembl canine genes ENSCAFG00000018370, ENSCAFG00000023088 and ENSCAFG00000023058. Our gene prediction exhibited 98% amino acid identity with reference ortholog. In this case, we considered that our workflow have defined a new canine gene.



3. Apport de la génomique comparée à l'identification de locus de susceptibilité à des maladies génétiques

(Article 5 : Hédan et al., 2007)

Au cours de mon travail de thèse, j'ai eu l'occasion de collaborer à des projets qui ont nécessité l'apport de la génomique comparée et de l'analyse de la conservation de l'ordre des gènes entre les génomes du chien et d'autres organismes mammifères. Dans ce paragraphe, j'illustrerai l'intérêt de la génomique comparative pour caractériser des locus de susceptibilité à des maladies génétiques chez le chien homologues aux maladies rencontrées chez l'homme.

Le chien est un modèle génétique puissant pour identifier les gènes et les allèles responsables d'affections génétiques ou de variations phénotypiques. Toutes les races canines présentent des phénotypes, des aptitudes, des performances, bien distincts ainsi que des susceptibilités à de nombreuses maladies génétiques monogéniques ou complexes qui sont pour la plupart homologues de maladies humaines. La force du modèle canin est accentuée par (i) l'aspect spontané des maladies canines (contrairement au modèle souris, souvent induit en laboratoire), (ii) leur caractère souvent spécifique de races ou d'un petit groupe de races apparentées dû à une mutation fondatrice sélectionnée et amplifiée dans la ou les races apparentées (iii) la forte incidence de ces maladies dans les races prédisposées (Galibert *et al.*, 2006; Wayne *et al*. 2007).

Notre équipe est très impliquée dans la recherche des bases génétiques des maladies spécifiques du chien afin de mettre en évidence de nouveaux gènes, de nouvelles fonctions chez le chien, transposables a l'homme. Parmi les différents projets, une étude du phénotype merle, une couleur de robe grisonnante dans certaines races canines, a été engagée depuis trois ans au laboratoire. Le phénotype merle est caractérisé par une dilution de la couleur des robes chez le chien et est associé à des anomalies de développement oculaire et auditif dans les cas les plus graves (homozygotie pour le locus). Ce phénotype est très ressemblant à celui observé chez l'homme dans les cas de syndrome de Waardenburg. Chez l'homme, des mutations dans quatre gènes de la voie de pigmentation des mammifères ont été mises en évidence dans les quatre types de syndromes. Pour autant, tous les cas ne sont pas expliqués par ces mutations et de nouveaux gènes ou de nouvelles mutations restent à découvrir. Par conséquent, nous avons combiné une approche par gènes candidats (gènes impliqués dans la voie de pigmentation des mammifères) avec des analyses de liaison génétique. Des microsatellites choisis dans, ou proches, des gènes candidats ont été

génotypés dans une famille d'une centaine de chiens présentant des chiens "merle" et un locus a été mis en évidence sur le chromosome 10 du chien (CFA10). Un des objectifs était alors de caractériser les segments conservés du CFA 10 avec les génomes de l'homme et de la souris, et plus particulièrement, le segment conservé renfermant le locus d'intérêt, afin d'avoir la meilleure idée possible du contenu en gènes de ce locus et des ruptures de synténie éventuelles avec l'homme. La construction de la carte de synténie a été réalisée par le programme AutoGRAPH et les résultats sont présentés dans la figure 22 (**Fig.22**).

Le locus candidat identifié par les analyses de liaison génétique sur la famille de bergers australien correspondait à une région d'environ 3 Mb en partie centromérique du CFA10 (de 3Mb à 6Mb). Cette région est incluse dans le premier CSO localisé sur le génome du chien de 3 à 19 Mb et orthologue des chromosomes humain HSA12 de 54 à 74 Mb et murin MMU10 de 111 à 129 Mb. Ce CSO est délimité dans sa partie inférieure par une rupture de synténie commune aux deux génomes testés et contient plus de 120 gènes en relation 1:1 en forte colinéarité entre les trois génomes. Cette forte colinéarité nous a permis de repérer les régions humaine et murine orthologues du locus canin de susceptibilité. Ce locus de 3 Mb contient au total une centaine de gènes dans les trois génomes mais seulement une douzaine impliqués dans la voie de pigmentation des mammifères. Ainsi, l'analyse comparative des génomes a permis de faciliter la mise en évidence de ce "locus merle" dans les trois génomes d'intérêt. Dans le même temps, une équipe américaine a découvert que les chiens "merle" possédaient une insertion d'un transposon de type SINEC_Cf dans le gène *SILV (*Clark *et al*., 2006). Ce gène localisé à ~3,3 Mb sur le chromosome 10 du chien appartient au "locus merle" identifié dans l'analyse de liaison génétique et est impliqué dans la voie de pigmentation des mammifères. Même si le phénotype mutant du gène *SILV* n'est pas encore connu chez l'homme, ce gène représente désormais un bon candidat pour les désordres pigmentaires humains dont les syndromes de Waardenburg.

Article n°5

Coat colour in dogs: identification of the merle locus in the Australian shepherd breed.

Hedan B, Corre S, Hitte C, Dreano S, Vilboux T, Derrien T, Denis B, Galibert F, Galibert MD, Andre C.

BMC Vet Res. 2006

BMC Veterinary Research

BioMed Central

Research article

Open Access

Coat colour in dogs: identification of the *Merle* locus in the Australian shepherd breed

Benoit Hédan[1], Sébastien Corre[1], Christophe Hitte[1], Stéphane Dréano[1], Thierry Vilboux[1], Thomas Derrien[1], Bernard Denis[2], Francis Galibert[1], Marie-Dominique Galibert[1] and Catherine André*[1]

Address: [1]UMR 6061 CNRS, Génétique et Développement, Faculté de Médecine, Université de Rennes1, 35043 RENNES Cédex, France. and [2]5 avenue Foch 54200 Toul, France.

Email: Benoit Hédan - benoit.hedan@univ-rennes1.fr; Sébastien Corre - sebastien.corre@univ-rennes1.fr; Christophe Hitte - christophe.hitte@univ-rennes1.fr; Stéphane Dréano - stephane.dreano@univ-rennes1.fr; Thierry Vilboux - thierry.vilboux@univ-rennes1.fr; Thomas Derrien - thomas.derrien@univ-rennes1.fr; Bernard Denis - denis.brj@wanadoo.fr; Francis Galibert - francis.galibert@univ-rennes1.fr; Marie-Dominique Galibert - marie-dominique.galibert-anne@univ-rennes1.fr; Catherine André* - catherine.andre@univ-rennes1.fr

* Corresponding author

Published: 27 February 2006

*BMC Veterinary Research*2006, **2**:9 doi:10.1186/1746-6148-2-9

Received: 11 November 2005
Accepted: 27 February 2006

This article is available from: http://www.biomedcentral.com/1746-6148/2/9

Abstract

Background: Coat colours in canines have many natural phenotypic variants. Some of the genes and alleles involved also cause genetic developmental defects, which are also observed in humans and mice. We studied the genetic bases of the *merle* phenotype in dogs to shed light on the pigmentation mechanisms and to identify genes involved in these complex pathways. The *merle* phenotype includes a lack of eumelanic pigmentation and developmental defects, hearing impairments and microphthalmia. It is similar to that observed in *microphthalmia* mouse mutants.

Results: Taking advantage of the dog as a powerful genetic model and using recently available genomic resources, we investigated the segregation of the *merle* phenotype in a five-generation pedigree, comprising 96 sampled Australian shepherd dogs. Genetic linkage analysis allowed us to identify a locus for the *merle* phenotype, spanning 5.5 megabases, at the centromeric tip of canine chromosome 10 (CFA10). This locus was supported by a Lod score of 15.65 at a recombination fraction $\theta = 0$. Linkage analysis in three other breeds revealed that the same region is linked to the *merle* phenotype. This region, which is orthologous to human chromosome 12 (HSA12 q13-q14), belongs to a conserved ordered segment in the human and mouse genome and comprises several genes potentially involved in pigmentation and development.

Conclusion: This study has identified the locus for the *merle* coat colour in dogs to be at the centromeric end of CFA10. Genetic studies on other breeds segregating the *merle* phenotype should allow the locus to be defined more accurately with the aim of identifying the gene. This work shows the power of the canine system to search for the genetic bases of mammalian pigmentation and developmental pathways.

Background

Coat colours in mammals depend on skin and hair pigment synthesis. Melanocytes manufacture two types of melanin: the black/brown photo-protective eumelanin pigment, and the red-yellow cytotoxic phaeomelanin pigment. Several paracrine factors secreted primarily by surrounding keratinocytes are involved in the melanogenic pathway by stimulating the switch between phaeomelanin and eumelanin [1]. In this pathway, *microphthalmia transcription factor* (*MITF*) plays a central role by regulating the expression of the *TYR* (*Tyrosinase)*, *TRP-1* (*Tyrosine Related Protein*) and *DCT* (*Dopachrome Tautomerase*) genes that encode enzymes involved in pigment manufacture [2,3].

Coat colour is highly polymorphic in dogs. In 1957, Little described, after observing the possible phenotypes, more than 20 loci affecting coat colours [4,5]. Until recently, only a few genes were recognised as involved in pigmentation. However, more and more genes, alleles and new interactions are being discovered: variants of *melanocortine 1 receptor* gene (*MC1R*), (locus previously called *extension* E) [6-8], variants of Agouti, the antagonist ligand of MC1R [9,10], variants of *tyrosinase-related protein 1* (*TYRP1*) [11] and variants of *melanophillin* [12]. Three mutations responsible for the brown coat colour versus black coat colour were described in *TYRP1* in several dog breeds including the Australian Shepherd dog [11]. Genomic tools are now fully available in canine genetics: dense radiation hybrid maps with 1500 polymorphic microsatellite markers and anchored BAC markers [13,14], a radiation hybrid map comprising 10,000 canine gene-based markers [15], and a whole sequence assembly of the canine genome, build 2.1 [16]. Altogether, the dog appears to be a good model for understanding better the genetics of pigmentation in mammals and for isolating new genes, new variants and interactions between alleles of different loci.

We are interested in the *merle* phenotype because of its involvement in coat colour and developmental impairments. The *merle* phenotype is a dominant trait, with heterozygous dogs presenting a coat colour in which eumelanic regions are incompletely and irregularly diluted, leaving intensely pigmented patches. *Merle* is found throughout the body except on the pheomelanic regions of the black and tan coat colour (Figure 1A, 1B). These dogs often have heterochromia iridis or blue eyes and often have a lack of retinal pigment visible on the fundus. Homozygous *merle* dogs display a more severe phenotype. The dogs are usually very pale, sometimes completely white and present developmental defects with an incomplete penetrance, microphthalmia and hearing loss (Figure 1C, 1D). In *merle* European lineages, microphthalmia and/or hearing loss are not frequently observed as breeders avoid mating *merle* dogs to avoid these developmental defects. However, several veterinary studies on the "*merle* syndrome", reported retinal defects [17], microphthalmia and coloboma [18]. The non-survival or degeneration of melanocytes in the cochlea have been suggested to explain hearing loss [19].

When analysing the genetic basis of the *merle* phenotype, Little suggested that a unique locus (called M) was responsible for the *merle* phenotype in different breeds [4]. It was proposed that the *merle* coat colour may be due to a transposable element, after the observation of two germinal reversions out of 66 *merle* offspring of a homozygous *merle* female [20]. Recently, the *Kit Ligand*, *KITLG*, was excluded as a candidate gene for the *merle* phenotype in dogs [21] and the candidate gene approach has not yet give any conclusive results.

We searched for candidate genes for the *merle* phenotype in dogs by considering well-described pigment disorders in mice. Mutations in the gene of the Mitf pathway cause specific coat colour phenotypes, some of which are similar to the *merle* phenotype in dogs. These include dilution of the coat colour in patches and complete or mild microphthalmia (Figure 1E, 1F). Also, the complete abolition of functional Mitf results in loss of the melanocyte lineage, causing a white coat colour due to a lack of pigment cell manufacturer, and additional eye (microphthalmia) and inner ear disorders. Heteroallelic combinations of *MITF* variants produce animals with normal sized dark eyes and yellowish-brown to grey spotted checker board-like coat colours [22,23]. Mitf is also involved in human Waardenburg syndromes, including pigment cell migration disorders [24] and developmental defects such as deafness. Pax3 (Paired box gene 3) and Sox10 (SRY – Sex determining region Y-box 10), which regulates *MITF* gene expression, are also associated with this syndrome [25,26]. This genetic evidence suggest that *MITF*, *PAX3* and *SOX10* genes may be candidate genes for the *merle* phenotype.

We collected a pedigree of Australian shepherd dogs and used a genetic linkage approach with microsatellite markers flanking the *MITF*, *PAX3* and *SOX10* candidate genes to search for the genetic bases of the *merle* coat colour in canines. Although these three genes were excluded for the *merle* phenotype in dogs, we successfully identified the *merle* locus on canine chromosome 10, close to the centromere, 20 Mb away from Sox10. This locus was restricted to a 5.5 Mb interval and was further confirmed by analysing families of other breeds segregating the *merle* phenotype.

BMC Veterinary Research 2006, **2**:9 http://www.biomedcentral.com/1746-6148/2/9

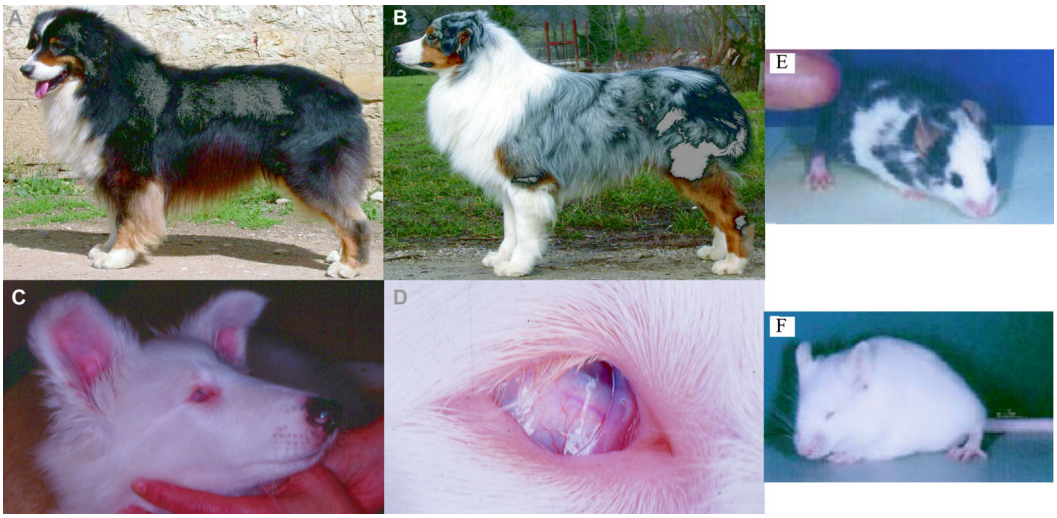


Figure 1
Pictures of none *merle* and *merle* dogs and mice microphthalmia mutants. A: Black and tan Australian Shepherd dog. B: Heterozygous *merle* Australian shepherd dog (pictures from Elevage du Paradis Sauvage de Ménestruel, Poncin, France) [40]. C: Six-month homozygous *merle* collie: the coat colour is totally white and the dog is blind and deaf. D: The left eye of the same dog, shows a microphthalmia with microcornea and a cataract (pictures from Dr Gilles Chaudieu, DVM, Dip. ECVO, Chamalières, France). E: $Mitf^{mi\text{-}sp}/Mitf^{mi\text{-}vga9}$ mouse. F: $Mitf^{mi\text{-}vga9}/Mitf^{mi\text{-}vga9}$ mouse with severe microphthalmia (pictures from Steingrímsson [23]).

Results

Pedigrees

A pedigree comprising 96 Australian shepherd dogs (43 brown and 53 black dogs) was collected. This pedigree, called the "complete pedigree", included 42 *merle* dogs. A sub-pedigree of 38 dogs, including 17 *merle* dogs, derived from the complete pedigree was used for genotyping (Figure 2). Isolated families from different breeds segregating the *merle* coat colour were also collected, including three dachshund families (14 dogs); a Beauce shepherd family (five dogs) and a Border collie family (13 dogs).

Genetic linkage analysis of the brown coat colour in the pedigrees

We evaluated the linkage power of the pedigrees by investigating the genetic linkage between the brown phenotype and the *TYRP1* gene in the complete pedigree. As the *TYRP1* gene was previously associated with the brown coat colour in dogs [11], we genotyped markers linked to *TYRP1*. These markers, FH2319 and REN105I03, are 1.18 and 5.17 Mb from *TYRP1*, respectively (see Additional file 1). The Lod scores between the brown phenotype and FH2319 and REN105I03 markers obtained by a two-point analysis on the complete pedigree, were 7.2 and 3.5 respectively, with a recombination fraction of θ = 0. For the sub-pedigree, the Lod scores were 3.6 and 2.4 respectively at θ = 0. The number of informative meiosis for the brown phenotype was 58 in the complete pedigree and 16 in the sub-pedigree. As the number of informative meiosis for the *merle* phenotype significantly increased to 81 and 33 in the complete pedigree and sub-pedigree, respectively, we expected these present pedigrees to be valuable for linkage analysis of the *merle* phenotype.

Genetic linkage analysis of the merle *trait*

As expected, the transmission mode in the collected pedigrees was consistent with an autosomal dominant segregation of the *merle* phenotype.

Using the sub-pedigree, we carried out genetic linkage analysis on polymorphic markers either flanking or within the intronic part of *TYRP1* gene and the candidate genes *MITF*, *PAX3*, and *SOX 10*. Two-point analysis showed no significant linkage between the *merle* phenotype and markers flanking the *MITF* and *PAX3* genes (Table 1). Therefore, we could exclude *MITF* and *PAX3* being involved in this phenotype. However, we found significant linkages (Lod scores ranging from 3.09 to 3.65) with markers flanking *SOX10*, with recombination fractions, θ, ranging from 0.08 to 0.14 (Table 2). This sug-

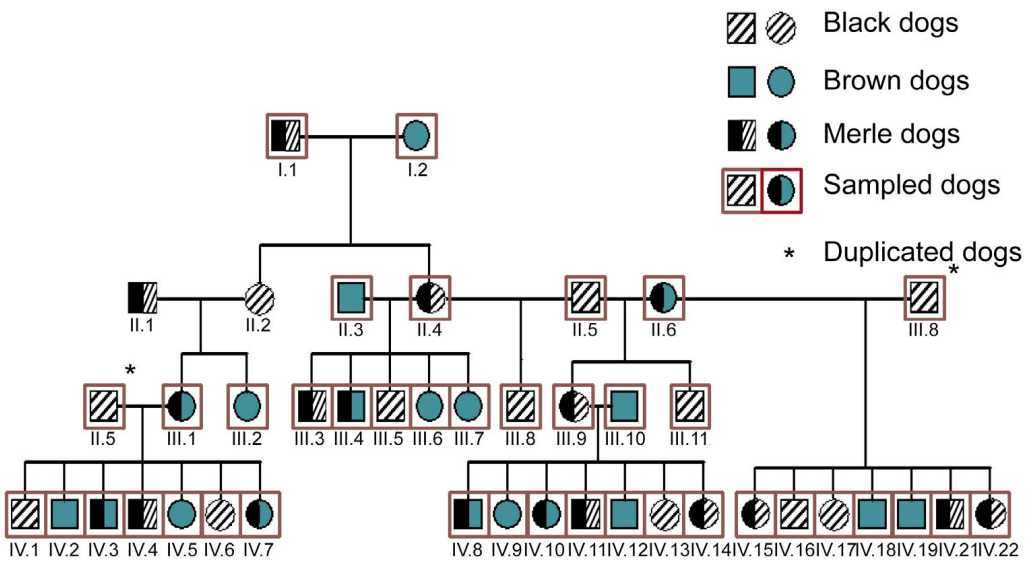


Figure 2
Sub-pedigree of 38 sampled Australian shepherd dogs (19 brown and 19 black dogs).

gested that the *merle* locus was about 10 cM from *SOX10*. As part of a "chromosome walking strategy", we selected 30 new polymorphic markers spanning a 27 Mb region from the *SOX10* region to the centromeric tip of CFA10, as the C10.769 marker telomeric to *SOX10* had a decreasing Lod score (Table 2).

Linkage analysis allowed us to identify seven markers close to the centromere, CFA10.1 to CFA10.8, which cosegregate with the *merle* phenotype with significant Lod scores (> 3) (Table 2, see Additional file 2).

We extended the genetic linkage by analysing the nine most centromeric markers in the "complete pedigree" and in three nuclear families of dachshund, Beauce shepherd and Border collie segregating the *merle* phenotype. We obtained increased Lod scores for markers CFA10.1 to CFA10.8, with maximum Lod scores for CFA10.7 and CFA10.8 (Lod scores at $\theta = 0$ of 15.65 and 14.90 in the complete pedigree and 19.87 and 19.57 in the complete pedigree plus the three other families, respectively). The CFA10.9 marker (telomeric to CFA10.8) is unlinked to the phenotype (Lod scores at $\theta = 0$ of -14.38 in the complete pedigree and -11.97 in the complete pedigree plus the three other families, respectively, Table 2). Haplotype analyses of this region allowed us to detect a recombination event between the *merle* phenotype and the CFA10.9 microsatellite, thus limiting further the *merle* locus (Figure 3). These data, as well as the previous results for *SOX10* flanking markers (Table 2), allowed us to exclude the *SOX10* gene as being involved in the *merle* phenotype. Our results show that the *merle* locus is located in a 5.5 Mb region between the end of the centromere, arbitrarily located at 3 Mb, (represented by CFA10.1 located at 3.1 Mb) and the CFA10.9 marker (located at 8.5 Mb) defining the telomeric limit of the critical interval (see Additional file 2).

Discussion

In the present study, we used a genetic linkage approach on a pedigree of Australian shepherd dogs segregating the *merle* phenotype. We identified with high statistical support a 5.5 Mb locus at the centromeric tip of CFA10 in which the gene responsible for this phenotype should be found. Dog samples were collected from breeders. As phenotyping was easily and immediately detectable after birth by breeders themselves and then officially declared to the breed club, we have been able to obtain an informative family with as many as five generations of Australian shepherd dogs. Moreover, parentage testing ensured good

BMC Veterinary Research 2006, **2**:9 http://www.biomedcentral.com/1746-6148/2/9

Table 1: Linkage data between the *merle* phenotype and markers flanking the *TYRP1* gene and *MITF* and *PAX3* candidate genes. Two point Lod scores and recombination fractions (θ) between the *merle* phenotype and the markers flanking *TYRP1*, *MITF* and *PAX3* genes calculated on the sub-pedigree. [a] CFA: *Canis familiaris* chromosome. [b] Starred markers were selected from the CanFam 1.0 canine sequence draft. Significant exclusions (<-2) are indicated in bold.

			θ at Lod score max		Lod scores at θ						
CFA[a]	*Gene*/Markers[b]	Gene/Marker Position in bp	θ	Lod score max	0.001	0.01	0.05	0.1	0.15	0.2	0.3
CFA11	***TYRP1***	34,664779–34,683148									
	FH2319	35,863893–35,864169	0.5	0.0	**-23.083**	**-13.146**	**-6.442**	**-3.802**	**-2.429**	-1.582	-0.661
	Ren105I03	39,857761–39,857999	0.5	0.0	**-15.888**	**-8.939**	**-4.278**	**-2.471**	-1.549	-0.995	-0.417
	Ren96j16	22,981531–22,981669	0.5	0.0	**-13.194**	**-8.198**	**-4.721**	**-3.24**	**-2.384**	-1.786	-0.963
	Ren130E03	22,981511–22,981683	0.5	0.0	**-13.194**	**-8.198**	**-4.721**	**-3.24**	**-2.384**	-1.786	-0.963
	Ren159M20	23,456266–23,456464	0.5	0.0	**-19.789**	**-11.82**	**-6.37**	**-4.141**	**-2.915**	**-2.1**	-1.061
CFA20	***MITF***	24,701418–24,735483									
	Ren100j13	25,668407–25,668570	0.5	0.0	**-24.585**	**-14.621**	**-7.791**	**-4.988**	**-3.442**	**-2.417**	-1.131
	Ren178E07	26,911251–26,911383	0.5	0.0	**-12.593**	**-7.605**	**-4.164**	**-2.729**	-1.923	-1.377	-0.671
	Ren105M20	19,606338–19,606486	0.4	0.087	**-8.992**	**-5.016**	**-2.327**	-1.264	-0.709	-0.367	-0.011
	Ren67C18	22,361131–22,361265	0.5	0.0	**-11.992**	**-7.012**	**-3.606**	**-2.218**	-1.462	-0.969	-0.379
	Pax3.2*	30,185679–30,186031	0.27	0.147	**-9.89**	**-4.957**	-1.763	-0.644	-0.159	0.063	0.137
CFA37	***PAX3***	31,332182–31,429995									
	Pax3.1*	31,368786–31,369020	0.5	0.0	**-20.085**	**-11.159**	**-5.204**	**-2.924**	-1.779	-1.101	-0.418
	Ren75L05	31,649572–31,649607	0.5	0.0	**-9.294**	**-5.321**	**-2.65**	-1.605	-1.065	-0.732	-0.36

reliability of the pedigrees. Such collected pedigrees from existing families means that housing of dogs is not required, limiting housing costs and ethical issues.

We focused on three candidate genes belonging to the coat colour pathway: *MITF, PAX3* and *SOX 10*. The *merle* phenotype shares similarities with *Mitf* mouse mutants in coat colour and ocular and hearing defects, and also with human Waardenburg patients. However, our genetic linkage study ruled out these three genes as being involved in the *merle* phenotype. Genetic analysis of the *SOX10* region in the sub-pedigree, using a "chromosome walking" strategy on CFA10, allowed us to identify the *merle* locus. It spans a 5.5 Mb region 20 Mb away from *SOX10*. Genetic linkage analyses on the complete pedigree and on small families from other breeds confirmed that the *merle* locus was located between the CFA10 centromere (3 Mb) and the CFA10.9 marker (8.5 Mb), with the highest Lod score of 19.87.

The corresponding orthologous human region is HSA12q13-q14 (position from 54.36 Mb to 60.94 Mb) and mouse region is MMU10D3 (position from 122.8 Mb to 128.7 Mb). These orthologous dog, human and mouse regions correspond to a unique conserved ordered segment, which has the same orientation in dogs and humans but is inverted between dogs and mice. In the dog region, 99 genes are predicted and 48 are known (Broad1), in the human region, 134 genes are predicted and 98 are known (NCBI 35), and in the mouse region, 112 genes are predicted and 95 are known (NCBI M34) [27]. These gene numbers may vary due to slight changes in the annotated genes as the versions are updated. In humans, mice and dogs, the conserved segments are totally ordered, making annotation of the dog segment easy, thus helping determination of candidate *merle* genes. This locus has many candidate genes, with at least a dozen being potential metabolic candidates as they, or their paralogs, belong to the pigmentation pathway. These include proteins involved in neural crest development (such as ERBB3), melanosome motility and transfer to surrounding keratinocytes (such as Silv/Pmel-17 and rab, kinesin, dynactin, myosin proteins).

Although the *MITF* gene itself has been excluded, the *merle* mutation should affect a gene interacting directly with the *MITF* gene in the pigmentation pathway. Alternatively, a more complex mechanism could explain the incomplete penetrance of eye defects observed in homozygous *merle* dogs. Although hearing loss may be due to an extreme white phenotype, including the absence of melanocytes in the cochlea, as in other white canine breeds [28], less is known about the origins of microphthalmia and other ocular defects. These may be due to another mutation in the same locus.

The *merle* phenotype occurs in several breeds and is commonly encountered in mongrel dogs. Breeds segregating

BMC Veterinary Research 2006, **2**:9 http://www.biomedcentral.com/1746-6148/2/9

Table 2: Linkage data between the *merle* phenotype and CFA10 markers. List of the CFA10 genotyped markers in the sub-pedigree (left), the complete pedigree (middle) and all the studied pedigrees (right), with their two point Lod scores and recombination fractions (θ). All markers were selected from CanFam 1.0 sequence draft except FH2293 and C10769 taken from Guyon et al.[13]. The markers are ordered in the table from the centromere to the telomere. [a] markers flanking the *SOX10* gene. Values in bold correspond to the highest Lod scores obtained in the three conditions.

			Two point Lod scores values								
			θ at Lod max		Lod scores at θ						
	Gene/Marker names	Marker Position (in bp)	θ	Lod max	0.001	0.01	0.05	0.1	0.15	0.2	0.3
Lod scores in the sub-pedigree	CFA10.1	3,109459–3,109703	0.00	6.32	6.31	6.22	5.79	5.23	4.64	4.02	2.68
	CFA10.2	3,150322–3,150695	0.00	6.92	6.91	6.82	6.37	5.78	5.16	4.50	3.05
	CFA10.3	3,169429–3,169640	0.00	3.01	3.01	2.95	2.72	2.42	2.10	1.76	1.06
	CFA10.4	3,452451–3,452757	0.04	5.94	4.82	5.71	5.90	5.55	5.04	4.44	3.04
	CFA10.5 [c]	3,533969–3,534312	0.00	6.92	6.91	6.81	6.35	5.73	5.09	4.40	2.90
	CFA10.6	4,050670–4,050915	0.00	3.01	3.01	2.96	2.74	2.46	2.17	1.86	1.20
	CFA10.7	4,901919–4,902276	0.00	**7.23**	7.21	7.11	6.65	6.04	5.39	4.71	3.20
	CFA10.8	6,938580–6,938928	0.00	**7.23**	7.21	7.11	6.65	6.04	5.39	4.71	3.20
	CFA10.9	8,539383–8,539710	0.50	0.00	-20.08	-11.15	-5.16	-2.85	-1.68	-0.98	-0.31
	CFA10.10	9,459108–9,459324	0.07	2.62	1.21	2.15	2.60	2.57	2.40	2.16	1.52
	CFA10.11	10,599594–10,599989	0.04	5.08	3.91	4.82	5.07	4.79	4.35	3.84	2.64
	CFA10.12	12,963916–12,964094	0.15	2.55	-4.49	-0.58	1.79	2.43	2.55	2.44	1.83
	CFA10.13	13,290922–13,291305	0.18	0.75	-2.99	-1.03	0.21	0.60	0.73	0.74	0.57
	CFA10.14	13,687824–13,688059	0.11	3.44	-1.49	1.42	3.09	3.43	3.36	3.10	2.25
	CFA10.15	15,287513–15,287805	0.14	2.22	-2.99	-0.06	1.69	2.15	2.21	2.08	1.51
	CFA10.16	16,325750–16,325992	0.09	3.67	0.61	2.53	3.55	3.66	3.49	3.19	2.33
	CFA10.17	16,883188–16,883575	0.12	2.95	-2.09	0.83	2.53	2.92	2.91	2.72	2.02
	CFA10.18	17,574178–17,574420	0.08	3.87	0.91	2.82	3.79	3.84	3.61	3.25	2.31
	CFA10.19	18,220318–18,220600	0.10	4.20	-0.58	2.31	3.93	4.20	4.07	3.74	2.75
	CFA10.20	19,075239–19,075433	0.12	3.19	-1.79	1.12	2.81	3.17	3.13	2.90	2.10
	CFA10.21	19,999011–19,999231	0.26	0.51	-11.69	-5.76	-1.88	-0.50	0.12	0.41	0.47
	CFA10.22	21,607022–21,607316	0.10	4.45	-0.28	2.61	4.20	4.45	4.30	3.95	2.90
	CFA10.23	21,956041–21,956337	0.10	4.45	-0.28	2.61	4.20	4.45	4.30	3.95	2.90
	CFA10.24	23,113457–23,113656	0.25	0.92	-13.18	-6.27	-1.77	-0.17	0.52	0.83	0.85
	CFA10.25	24,031569–24,031859	0.10	3.51	-1.19	1.70	3.28	3.51	3.34	2.97	1.93
	CFA10.26	25,504535–25,504764	0.14	2.73	-4.19	-0.29	2.05	2.65	2.73	2.57	1.89
	CFA10.27	25,930447–25,930692	0.12	1.84	-1.49	0.45	1.58	1.83	1.81	1.67	1.16
	CFA10.28	26,491052–26,491288	0.13	1.66	-1.79	0.16	1.32	1.62	1.65	1.56	1.16
	CFA10.29	27,391154–27,391817	0.18	2.36	-6.88	-1.98	1.09	2.04	2.33	2.33	1.87
	CFA10.30	28,230039–28,230419	0.30	0.69	-18.28	-9.37	-3.47	-1.27	-0.23	0.33	0.69
	CFA10.31[a]	29,232270–29,232511	0.13	3.51	-3.28	0.61	2.90	3.46	3.49	3.27	2.45
	CFA10.32[a]	29,355913–29,355989	0.14	3.09	-3.88	0.02	2.37	2.99	3.08	2.94	2.24
	CFA10.33[a]	29,483066–29,483109	0.13	2.40	-2.69	0.23	1.95	2.37	2.38	2.22	1.58
	CFA10.34[a]	29,826073–29,826290	0.14	2.52	-2.69	0.24	2.00	2.45	2.51	2.38	1.81
	CFA10.35[a]	29,834441–29,834690	0.16	2.74	-6.28	-1.39	1.63	2.51	2.73	2.67	2.08
	SOX10	29,856920–29,867728									
	CFA10.36[a]	29,861299–29,861547	0.08	3.60	0.61	2.52	3.51	3.58	3.37	3.02	2.10
	CFA10.37[a]	29,868038–29,868267	0.14	3.38	-3.58	0.32	2.67	3.29	3.37	3.21	2.47
	CFA10.38[a]	30,176995–30,177378	0.11	3.65	-1.19	1.71	3.35	3.65	3.55	3.26	2.38
	FH2293[a]	31,696028–31,696274	0.13	3.56	-3.28	0.61	2.93	3.50	3.54	3.34	2.53
	C10.769[a]	36,663349–36,663563	0.18	1.50	-5.99	-2.06	0.42	1.20	1.46	1.49	1.18
Lod scores in the complete pedigree	CFA10.1	3,109459–3,109703	0.00	11.74	11.72	11.54	10.70	9.61	8.46	7.27	4.71
	CFA10.2	3,150322–3,150695	0.02	11.97	11.13	11.91	11.63	10.64	9.48	8.20	5.38
	CFA10.3	3,169429–3,169640	0.00	6.62	6.61	6.50	5.98	5.30	4.60	3.86	2.33
	CFA10.4	3,452451–3,452757	0.03	13.82	11.73	13.46	13.65	12.67	11.36	9.88	6.55
	CFA10.5 [c]	3,533969–3,534312	0.02	8.77	7.82	8.65	8.58	7.88	7.01	6.03	3.86
	CFA10.6	4,050670–4,050915	0.00	6.92	6.91	6.79	6.24	5.51	4.76	3.99	2.41
	CFA10.7	4,901919–4,902276	0.00	**15.65**	15.63	15.39	14.30	12.88	11.40	9.84	6.54
	CFA10.8	6,938580–6,938928	0.02	**14.90**	14.13	14.87	14.39	13.16	11.72	10.17	6.76

Table 2: Linkage data between the *merle* phenotype and CFA10 markers. List of the CFA10 genotyped markers in the sub-pedigree (left), the complete pedigree (middle) and all the studied pedigrees (right), with their two point Lod scores and recombination fractions (θ). All markers were selected from CanFam 1.0 sequence draft except FH2293 and C10769 taken from Guyon et al.[13]. The markers are ordered in the table from the centromere to the telomere. [a] markers flanking the *SOX10* gene. Values in bold correspond to the highest Lod scores obtained in the three conditions. *(Continued)*

	CFA10.9	8,539383–8,539710	0.19	2.48	-14.38	-5.54	0.02	1.78	2.38	2.47	1.89
Lod scores in the complete pedigree plus families of dachshund, Beauce shepherd and Border collie	CFA10.1	3,109459–3,109703	0.04	10.67	8.43	10.21	10.61	9.91	8.91	7.74	5.09
	CFA10.2	3,150322–3,150695	0.04	10.97	8.73	10.51	10.91	10.21	9.21	8.04	5.39
	CFA10.3	3,169429–3,169640	0.00	6.02	6.01	5.91	5.44	4.84	4.22	3.56	2.20
	CFA10.4	3,452451–3,452757	0.03	14.38	12.33	14.05	14.19	13.15	11.79	10.25	6.83
	CFA10.5 [c]	3,533969–3,534312	0.02	8.77	7.82	8.65	8.58	7.89	7.02	6.05	3.90
	CFA10.6	4,050670–4,050915	0.03	6.99	6.01	6.86	6.85	6.23	5.46	4.61	2.82
	CFA10.7	4,901919–4,902276	0.00	**19.87**	19.83	19.52	18.11	16.28	14.37	12.37	8.15
	CFA10.8	6,938580–6,938928	0.00	**19.57**	19.53	19.22	17.82	15.99	14.08	12.09	7.91
	CFA10.9	8,539383–8,539710	0.16	4.23	-11.97	-3.16	2.25	3.83	4.22	4.10	3.02

merle are from the collie lineage (Group 1-FCI – Federation Cynologique Internationale classification): Shetland sheepdog, Border collie, collie, Australian shepherd dog, etc., and from other unrelated breeds belonging to different FCI groups and different clusters, as defined by Parker *et al*. [29], such as dachshund (Group 4-FCI):, Beauce shepherd (Group 1-FCI), great Dane (Group 2-FCI), Welsh corgi cardigan and Pyrenean shepherd. The *merle* phenotype is most probably very old, with the *merle* coat colour being reported in old books [30,31], from which drawings of *merle* dogs have been selected and reproduced [32].

It is not yet known whether the genetic cause of the *merle* phenotype is the same in all breeds and mongrels segregating this phenotype. A unique locus has been suggested as responsible for the *merle* coat colour [4]. In the present study, the increased Lod scores observed for genotyped markers from the *merle* locus in dachshund, Beauce shepherd and Border collie families is consistent, at least in these breeds, with there being a unique locus for the *merle* coat colour. If all *merle* dogs share a common ancestor chromosome, all breeds segregating *merle* could be used to refine the locus. The sharing of the *merle* locus by several breeds and also by mongrels may be due either to a common ancestor chromosome region being transmitted throughout canine evolution and/or to backcrosses that introduced a *merle* haplotype in several breeds at different times.

Conclusion

Using genetic linkage analysis, we excluded the involvement of the *MITF, PAX3* and *SOX 10* candidates genes in the *merle* phenotype. However, we identified the *merle* locus at the centromeric end of CFA10 in pedigrees of Australian shepherd dogs, dachshund, Beauce shepherd dogs and Border collies segregating the *merle* phenotype. This locus spans 5.5 Mb and is linked to the *merle* coat colour with a maximum Lod score of 19.87 and a recombination fraction of 0. We are currently analysing this locus in several breeds segregating *merle*, with a high density of single nucleotide polymorphic markers (SNP). This should help in identifying the *merle* gene. As well as benefiting breeding practices and canine veterinary medicine, identifying the *merle* gene will also help in understanding the genetic bases of mammalian pigmentation and developmental pathways.

Methods

Genomic DNA extraction

No dogs were housed for research purposes, and all dogs were privately owned pets.

Blood samples and the accompanying pedigree and coat colour data (with pictures when possible) were collected by DVM veterinarians. All data were entered into a database. Genomic DNA was extracted from 5 ml of blood collected on EDTA, using the nucleon BACC 3 kit (Amersham Biosciences, Piscataway, NJ, USA). For low concentration samples, the extracted DNA was "whole genome amplified" using the genomiphi kit (Amersham Biosciences).

Canine pedigree

Pedigrees were constructed using the Cyrillic software (Cyrillic2.1) [33], which allows haplotypes from the genotyping data to be drawn and the data to be exported in different formats for use in genetic linkage analysis. We carried out genotyping of 20 polymorphic microsatellites from four different chromosomes (CFA10, 11, 20 and 37) to check and validate the parentage compatibility.

Markers selection and Genotyping experiments

Microsatellite markers were selected from RH map data [13,34] or from the CanFam 1.0 draft of the canine genome sequence [35]. Markers were selected from their

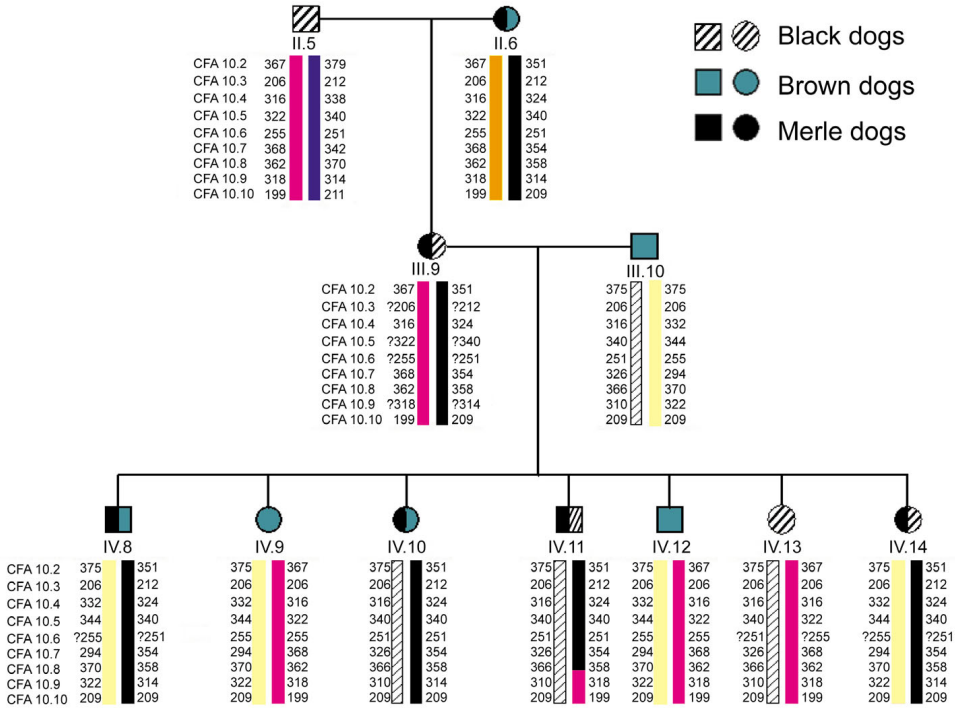


Figure 3
Haplotype analysis of the centromeric end of CFA10. Analysis revealed a crossing-over between CFA10.8 and CFA10.9 markers (dog IV.11), limiting the *merle* locus in between the centromeric end and the CFA10.9 marker. The *merle* chromosome is shown in black. Alleles with a question mark indicate that the parental origin of the allele could not be determined.

position and their polymorphism level (see Additional file 1). We used Primer3 software to design PCR primers [36].

Microsatellite markers were labelled using a two-step-PCR fluorescent labelling procedure [37]. The first step was carried out on 50 ng of dog genomic DNA using a classical PCR protocol and a touchdown program of 61°C to 51°C. The second step consisted of a one strand labelling PCR as previously described [37]. The PCR products were purified using Sephadex G50 fine column filtration in a 96 format (Amersham Biosciences). An aliquot of 3 μl fluorescent purified PCR product was mixed with 0.3 μl (0.2 nM) of fluorescent geneScan-500 ROX size standard (Applied Biosystems) and 8.7 μl of formamide, and then loaded onto a 3130 XL genetic analyser (Applied Biosystems). Results were analysed using GeneMapper software v3.7 (Applied Biosystems) and the genotyping data were used by Cyrillic software for the genetic linkage analysis.

Genetic linkage analysis

Haplotypes were constructed using the Cyrillic software. Two-point linkage analysis was carried out between each marker and the *merle* phenotype using M-LINK software through the GLUE web interface [38] and MultiMap software [39]. We used the 'prepare' option of CRI-MAP to check for Mendelian segregation. The linkage between each pair of markers was carried out with the TWOPOINT option of CRI-MAP. Lod scores were calculated assuming an autosomal dominant transmission with full penetrance and affected individuals were scored as heterozygous at the phenotype locus.

Authors' contributions

BH collected samples, constructed the pedigrees and performed genotyping experiments; BH also interpreted all dataand actively participated in writing the manuscript. SC helped with the genotyping experiments and interpretation of the dataand participated in writing the manu-

script. CH carried out the statistical analyses and data interpretation, and critically revised the manuscript. SD helped with the genotyping experiments. TV extracted DNA from blood samples and commented critically on the work and manuscript. TD carried out the synteny analyses. BD contributed with knowledge on canines coat colours and critically revised the manuscript. FG provided intellectual input and critically revised the manuscript. MDG helped conceive and design the work and helped in the writing of the manuscript. CA conceived and designed the work and drafted the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Characteristics of the markers used in the genetic linkage studies. *[a] CFA :* Canis familiaris *chromosome.[b] Starred markers were selected from the CanFam 1.0 canine sequence draft, [c] marker corresponding to marker FH2537 from Guyon et al. [13], [d] number of alleles as determined from the sub-pedigree. [E] Markers flanking* SOX10 *gene.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1746-6148-2-9-S1.doc]

Additional File 2

Scheme of two-point linkage analysis of the* merle *phenotype in the Australian Shepherd dog pedigrees on CFA10. *Two-point linkage analysis of the* merle *phenotype in the Australian shepherd dog sub-pedigree (in black) and complete pedigree (in brown) (Lod scores at theta = 0) is shown on the right. An ordered list of genotyped markers (right) and genes (left) and their position in Mb are indicated in the middle. An ideogram of the canine chromosome 10 is shown on the left with the corresponding human chromosomal conserved segments. NB: genomic sequence systematically starts at an arbitrary coordinate of 3 Mb to include the non-sequenced centromeric region.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1746-6148-2-9-S2.jpeg]

Acknowledgements

We gratefully acknowledge the CNRS, the Conseil Regional de Bretagne for their financial support, as well as Australian shepherd breeders, (especially Mme Boutarfa and Mme Coosemans, Elevage du Paradis Sauvage de Ménestruel, Poncin, France; Mme Bernard, Elevage de l'Orée des Charmois, Bois Le Roi, France), Mme Rougé, Elevage des Corsaires de Feu, Britany, France and Dr. Gilles Chaudieu, DVM and Dr. Philippe Pilorge, DVM for sampling and pictures.

References

1. Imokawa G: **Autocrine and paracrine regulation of melanocytes in human skin and in pigmentary disorders.** *Pigment Cell Res* 2004, **17:**96-110.
2. Busca R, Ballotti R: **Cyclic AMP a key messenger in the regulation of skin pigmentation.** *Pigment Cell Res* 2000, **13:**60-69.
3. Widlund HR, Fisher DE: **Microphthalamia-associated transcription factor: a critical regulator of pigment cell development and survival.** *Oncogene* 2003, **22:**3035-3041.
4. Little CC: **The inheritance of coat color in dogs.** Edited by: Press CU. Ithaca, NY, ; 1957.
5. Sponenberg DP, Rothschild MF: **Genetics of coat colour and hair texture.** In *The Genetics of the Dog* Ruvinsky A , Sampson J edition. New York, NY, CABI Publishing; 2001:61 -685.
6. Everts RE, Rothuizen J, van Oost BA: **Identification of a premature stop codon in the melanocyte-stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour.** *Anim Genet* 2000, **31:**194-199.
7. Newton JM, Wilkie AL, He L, Jordan SA, Metallinos DL, Holmes NG, Jackson IJ, Barsh GS: **Melanocortin 1 receptor variation in the domestic dog.** *Mamm Genome* 2000, **11:**24-30.
8. Schmutz SM, Berryere TG, Ellinwood NM, Kerns JA, Barsh GS: **MC1R studies in dogs with melanistic mask or brindle patterns.** *J Hered* 2003, **94:**69-73.
9. Berryere TG, Kerns JA, Barsh GS, Schmutz SM: **Association of an Agouti allele with fawn or sable coat color in domestic dogs.** *Mamm Genome* 2005, **16:**262-272.
10. Kerns JA, Newton J, Berryere TG, Rubin EM, Cheng JF, Schmutz SM, Barsh GS: **Characterization of the dog Agouti gene and a nonagoutimutation in German Shepherd Dogs.** *Mamm Genome* 2004, **15:**798-808.
11. Schmutz SM, Berryere TG, Goldfinch AD: **TYRP1 and MC1R genotypes and their effects on coat color in dogs.** *Mamm Genome* 2002, **13:**380-387.
12. Philipp U, Hamann H, Mecklenburg L, Nishino S, Mignot E, Gunzel-Apel AR, Schmutz SM, Leeb T: **Polymorphisms within the canine MLPH gene are associated with dilute coat color in dogs.** *BMC Genet* 2005, **6:**34.
13. Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, Quignon P, Lowe JK, Renier C, Gelfenbeyn B, Vignaux F, DeFrance HB, Gloux S, Mahairas GG, Andre C, Galibert F, Ostrander EA: **A 1-Mb resolution radiation hybrid map of the canine genome.** *Proc Natl Acad Sci U S A* 2003, **100:**5296-5301.
14. Breen M, Hitte C, Lorentzen TD, Thomas R, Cadieu E, Sabacan L, Scott A, Evanno G, Parker HG, Kirkness EF, Hudson R, Guyon R, Mahairas GG, Gelfenbeyn B, Fraser CM, Andre C, Galibert F, Ostrander EA: **An integrated 4249 marker FISH/RH map of the canine genome.** *BMC Genomics* 2004, **5:**65.
15. Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C, Evanno G, Pullar B, Cadieu E, Oza V, Lourgant K, Jaffe DB, Tacher S, Dreano S, Berkova N, Andre C, Deloukas P, Fraser C, Lindblad-Toh K, Ostrander EA, Galibert F: **Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping.** *Nat Rev Genet* 2005, **6:**643-648.
16. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulkobas EJ 3rd, Zody MC, *et al.*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438:**803-19.
17. Akcan A, Wegner W: **Changes in the visual pathways and visual centers in Merle syndrome in the dog.** *Z Versuchstierkd* 1983, **25:**91-99.
18. Gelatt KN, Powell NG, Huston K: **Inheritance of microphthalmia with coloboma in the Australian shepherd dog.** *Am J Vet Res* 1981, **42:**1686-1690.
19. Strain GM: **Congenital deafness and its recognition.** *Vet Clin North Am Small Anim Pract* 1999, **29:**895-907, vi.
20. Sponenberg DP: **Germinal reversion of the merle allele in Australian shepherd dogs.** *J Hered* 1984, **75:**78.
21. Schmutz SM, Berryere TG, Sharp CA: **KITLG maps to canine chromosome 15 and is excluded as a candidate gene for merle in dogs.** *Anim Genet* 2003, **34:**75-76.
22. Hansdottir AG, Palsdottir K, Favor J, Neuhauser-Klaus A, Fuchs H, de Angelis MH, Steingrimsson E: **The novel mouse microphthalmia mutations Mitfmi-enu5 and Mitfmi-bcc2 produce dominant negative Mitf proteins.** *Genomics* 2004, **83:**932-935.
23. Steingrimsson E, Arnheiter H, Hallsson JH, Lamoreux ML, Copeland NG, Jenkins NA: **Interallelic complementation at the mouse Mitf locus.** *Genetics* 2003, **163:**267-276.
24. Tassabehji M, Newton VE, Read AP: **Waardenburg syndrome type 2 caused by mutations in the human microphthalmia (MITF) gene.** *Nat Genet* 1994, **8:**251-255.
25. Potterf SB, Furumura M, Dunn KJ, Arnheiter H, Pavan WJ: **Transcription factor hierarchy in Waardenburg syndrome: regu-**

lation of MITF expression by SOX10 and PAX3. *Hum Genet* 2000, **107:**1-6.
26. Watanabe A, Takeda K, Ploplis B, Tachibana M: **Epistatic relationship between Waardenburg syndrome genes MITF and PAX3.** *Nat Genet* 1998, **18:**283-286.
27. Ensembl: **.** [http://www.ensembl.org/index.html].
28. Strain GM: **Deafness prevalence and pigmentation and gender associations in dog breeds at risk.** *Vet J* 2004, **167:**23-32.
29. Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L: **Genetic structure of the purebred domestic dog.** *Science* 2004, **304:**1160-1164.
30. Brehm AE: **La vie des animaux illustrée.** Edited by: Bailleres JB. Paris, ; 1868.
31. Buffon D: **Origine des espèces.** *Volume 3*. Edited by: Bureau-de-la-société-des-Publications-illustrées . Paris, ; 1839.
32. Société-Française-de-Cynotechnie: **Nos chiens d'antan.** Edited by: Maradi . France, ; 1994.
33. Cyrillic: **.** [http://www.cyrillicsoftware.com].
34. Canine-Hybrid-Radiation-project: **.** *UMR 6061* [http://www-recomgen.univ-rennes1.fr/doggy.html].
35. UCSC: **.** [http://genome.ucsc.edu].
36. Primer3: **.** [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi].
37. Jouquand S, Cheron A, Galibert F: **Microsatellite analysis using a two-step procedure for fluorescence labeling of PCR products.** *Biotechniques* 1999, **26:**902-905.
38. Glue: **.** [http://menu.hgmp.mrc.ac.uk/menu-bin/GLUE/glue.pl].
39. Matise TC, Perlin M, Chakravarti A: **Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map.** *Nat Genet* 1994, **6:**384-390.
40. Elevage-Menestruel: **.** [http://menestruel.free.fr].

III. Discussion - Conclusions

1. L'analyse comparée des génomes : applications en cartographie

Nous avons montré que la combinaison de plusieurs ressources pour un même génome permet d'exploiter de façon optimale l'information génétique pour l'organisme d'intérêt. La construction de cartes d'un génome optimise la phase d'assemblage du séquençage profond et est essentielle pour les données issues d'un projet de séquençage léger. Par exemple, le "Mammalian Genome Project" entrepris dès 2005 au Broad Institute aux États-Unis (http://www.broad.mit.edu/mammals/), avait pour mission initiale de séquencer, avec une couverture de 2X, 15 génomes mammifères d'intérêt phylogénétique et/ou considérés comme "modèles" en biologie. Neuf nouveaux projets de séquençage léger (~2X) (dont le lama *Lama glama*, le dauphin *Tursiops truncatus* ou bien les damans *Hyracoidea*...) ont été ajoutés à cette liste initiale. Finalement, 24 génomes mammifères sont ou vont être séquencés à des couvertures de 2X ou 7X. Parmi ces projets, certains organismes possèdent déjà des ressources de cartographie RH (le chat *Felis catus*, la vache *Bos taurus*) qu'il sera très utile de pouvoir combiner afin de maximiser réciproquement la qualité des cartes et du séquençage. Par ailleurs, d'autres génomes, en cours de séquençage, ne disposent pas de ressources supplémentaires de type cartographie. Il serait alors envisageable de produire des cartes RH de ces génomes afin de comparer et soutenir ces projets de séquençage mais sous la condition, d'une part, d'améliorer la construction des panels RH et, d'autre part, d'optimiser l'étape de génotypage. Ces considérations font l'objet de projets de recherche dans différentes équipes (McKay *et al*., 2007) dont la notre où l'objectif est d'automatiser le génotypage de marqueurs par la technologie des puces Illumina© dans le cadre de la cartographie RH de 1 500 marqueurs du génome du poisson bar *Dicentrarchus labrax*.

Cependant, tous les organismes d'intérêt ne bénéficient pas de projets de séquençage de leurs génomes. Par conséquent, seule la construction de cartes permettra de rendre compte de la structure globale des génomes et, plus localement, des chromosomes et des locus qui les composent. L'amélioration des méthodes de cartographie existantes et le développement de nouvelles méthodologies de construction de cartes constituent un pré-requis essentiel à l'étude d'un génome non séquencé.

À partir des données expérimentales, la construction de cartes RH repose sur un traitement statistique dont la fiabilité peut être évaluée, à la fois par la mesure statistique et probabiliste que constitue le calcul de la meilleure carte possible, et aussi, par l'apport de la cartographie comparée avec d'autres ressources quand celles-ci sont disponibles. Dans ce dernier cas, le respect de l'ordre des marqueurs entre plusieurs ressources permet alors de renforcer mutuellement ces approches et illustre la reproductibilité et la qualité des différentes méthodes employées. Le développement du projet AutoGRAPH a grandement facilité la comparaison des cartes dans le cadre de la cartographie comparée du génome du chien.

Il est, toutefois, relativement rare qu'un même organisme dispose de plusieurs cartes de son génome et ceci est d'autant plus vrai pour les organismes qui ne sont pas considérés comme des modèles de laboratoire. Aussi, une démarche complémentaire en cartographie comparée consiste à analyser l'ordre des marqueurs par une approche synténique avec des génomes de référence, ceci afin de rendre compte de la robustesse de la construction de la carte. L'ordre des gènes dans les génomes n'est pas aléatoire et cet ordre est d'autant plus forte que les organismes comparés sont proches au niveau phylogénétique. Dans le cadre du projet de cartographie de 10 000 marqueurs de gènes sur le génome canin, une évaluation de l'ordre des marqueurs a été fournie par la construction des cartes de synténie entre les 10 000 gènes canins et le génome de l'homme (Hitte *et al.*, 2005). Le programme AutoGRAPH a permis d'automatiser la construction des cartes de synténie et d'évaluer rapidement la colinéarité des marqueurs entre la carte RH du chien et les segments conservés ordonnés du génome humain. Lorsque les relations d'orthologie entre les marqueurs cartographiés par hybrides d'irradiation et un génome référent sont disponibles, il apparaît alors judicieux de pouvoir intégrer ce critère de synténie directement dans le processus de construction de la carte RH. Le groupe du Dr Thomas Faraut (Laboratoire de Génétique Cellulaire INRA - Toulouse) a ainsi développé une formalisation algorithmique permettant de prendre en compte la conservation de synténie avec un génome de référence pour améliorer l'ordonnancement de marqueurs peu supporté par l'analyse statistique lors de la construction de la carte seule (Faraut *et al.*, 2007). Un projet de collaboration s'est ainsi engagé entre les laboratoires de Toulouse et de Rennes dans le but d'appliquer cette nouvelle méthodologie au projet de carte RH de 10 000 gènes canins. En effet, chaque marqueur de gène canin avait été identifié par des alignements réciproques des séquences canines avec le génome humain (utilisé comme génome de référence). Finalement, nous avons estimé la validité de la méthode en la comparant, d'une part, à l'approche de cartographie sans apport de la synténie, et d'autre part, au séquençage profond du génome canin qui représentait la référence de l'expérience. Les résultats de l'analyse du chromosome 17 du chien (CFA17) sont

présentés dans la figure (**Fig.23**). Graphiquement, la colinéarité des marqueurs fournie par la carte RH intégrant l'apport de synténie est nettement améliorée quand on la compare avec l'ordre des marqueurs sans apport de synténie. Par ailleurs, nous avons évalué quantitativement cette amélioration en calculant le LIS (Longest Increasing Subsequence) correspondant à la plus longue séquence commune entre deux ressources. L'apport de la synténie dans la construction des cartes RH du chien a permis d'améliorer l'ordre des marqueurs de ~30% estimé par les valeurs de LIS en moyenne sur l'ensemble des cartes RH construites.

Cette nouvelle méthodologie offre de nombreuses perspectives quant à l'amélioration de la construction de cartes en utilisant la conservation de l'ordre des marqueurs avec un génome référent. La puissance de l'approche réside dans la quantité et la qualité des informations apportées par les ancres de comparaison avec le génome référent. Il serait judicieux de pouvoir intégrer, non plus seulement un seul génome, mais plusieurs génomes référents lors de la construction statistique de la meilleure carte possible. La contribution relative des génomes référents et leur impact dans l'amélioration de la carte pourraient être évalués en fonction (i) de la position phylogénétique de l'organisme cartographié par rapport aux génomes référents, (ii) du nombre de segments conservés ordonnés (CSO) et des ruptures de synténie associées entre chacun des génomes référents et le génome à cartographier et, enfin, (iii) de la colinéarité des ancres de comparaisons au sein des CSOs. Une stratégie de construction de carte RH supportée par l'identification des CSOs est à l'étude actuellement. L'analyse des cartes RH du génome du chien, par exemple, ne serait plus seulement considérée pour les 39 chromosomes canins pris individuellement mais plus efficacement, pour les ~250 CSOs identifiés entre le chien et le génome humain pris comme référence. Le choix du (ou des) organisme(s) de référence est en cours afin de mesurer l'impact du nombre de CSOs sur la qualité des cartes RH construites avec l'apport de synténie.

2. L'analyse comparée des génomes dans le contexte de l'annotation en gènes

2.1. Les options méthodologiques

2.1.1. Le choix des génomes de référence

Dans l'article 4 (Derrien *et al*., en préparation), nous avons utilisé la conservation de l'ordre des gènes entre quatre génomes de référence (homme - chimpanzé - rat - souris) et le chien pour définir des intervalles potentiels de localisation (COIL = Consensus Orthologous IntervalLs) des gènes orphelins chez le chien. Les quatre génomes de référence correspondent à deux ordres distincts que sont les primates et les rongeurs. L'avantage de combiner deux organismes par lignée a été de renforcer mutuellement la fiabilité des prédictions de gènes et des relations d'orthologies 1:1:1:1 annotées par le système standard d'Ensembl (Curwen *et al*., 2004). Cependant, cela a entraîné une certaine redondance des distances phylogénétiques des deux lignées prises individuellement par rapport au génome du chien. Il serait alors envisageable de multiplier les génomes de référence positionnés à des distances évolutives graduelles dans l'arbre phylogénétique des mammifères. Comme nous l'avons vu dans l'article 3, cela permettrait de diminuer la proportion de gènes de référence localisés dans des ruptures de synténie partagées qui empêchent la définition d'un intervalle unique sur le génome du chien. De plus, la réduction de l'espace de recherche par l'approche synténique préalable favorise des temps de calcul relativement raisonnables et autorise la prédiction d'un gène canin à partir de plusieurs gènes/génomes de référence. Utiliser des génomes distants du chien, tels que l'éléphant, favorisera l'identification de gènes très conservés au cours de l'évolution (empreinte phylogénétique ou phylogenetic footprinting) alors que la disponibilité récente d'autres génomes carnivores (chat, furet... http://genome.gov/page.cfm?pageID=10002154) permettra de mettre en évidence des gènes spécifiques à la lignée carnivore (ombrage phylogénétique ou phylogenetic shadowing). Très récemment, l'analyse des gènes et des taux d'évolution entre orthologues et paralogues pour 12 génomes de drosophiles a été réalisée (Heger *et al*., soumis). Cette étude illustre la possibilité d'analyser, non plus des paires de génomes uniquement, mais plus largement plusieurs génomes d'une même clade. Développer cette approche sur un nombre (sûrement plus restreint) de génomes carnivores (en cours ou déjà séquencés)

constitue une orientation importante pour l'analyse du génome du chien et de l'ensemble de la lignée des carnivores.

2.1.2. L'annotation des familles de gènes

L'analyse des 412 gènes en relation de synténie 1:0 entre les génomes de référence et le génome du chien a révélé la présence de familles de gènes impliquées dans les mécanismes d'adaptation de l'organisme à son environnement. La plupart de ces familles de gènes naissent par duplication en tandem entraînant un respect de l'ordre des gènes et, plus particulièrement, des paralogues issus de la duplication. Cependant, la fréquence de ces duplications n'est pas la même entre deux organismes et, par exemple, le génome humain contient trois fois plus de gènes dupliqués que le génome canin (Goodstadt *et al.*, 2006). Dans notre étude, nous nous sommes focalisés sur l'analyse des 412 gènes en relation 1:0 avec le chien sans avoir distingué individuellement toutes les familles impliquées dans le jeu de 412 gènes. Il serait intéressant d'utiliser AutoGRAPH dans le cadre de la caractérisation d'une famille de gènes entre plusieurs génomes. En effet, le long des chromosomes, le patron de distribution des relations 1:0 pour une famille de gènes est rapidement identifiable par notre programme. De plus, la possibilité de repérer ces relations 1:0 dans le contexte des segments conservés et des ruptures de synténie pourrait refléter les régions chromosomiques d'instabilité ou, au contraire, sous fortes contraintes sélectives associées à la perte ou au maintien des relations d'orthologie dans la famille considérée. Une étude récente de James Thomas sur la super-famille des cytochromes P450 a mis en évidence que les membres de la famille localisés dans des clusters n'étaient pas retrouvés en relation d'orthologie 1:1 avec d'autre génomes mammifères (Thomas *et al.*, 2007). À l'inverse, les gènes codant pour des cytochromes situés en dehors des clusters présentaient une forte conservation de séquence entre orthologues des espèces comparées. À cette distinction stabilité/instabilité des membres d'une famille de gènes, il serait intéressant d'étudier et d'intégrer leur contexte synténique (segment conservé/rupture de synténie) avec d'autres espèces, leur contexte chromosomique (régions télomèriques, isochores, régions recombinogènes...) et enfin, leur contexte génique (sélection positive ou négative) pour préciser les relations de causes à effets entre ces mécanismes évolutifs.

2.2. L'identification des réarrangements chromosomiques : limites et optimisations

L'identification des segments conservés entre deux génomes dépend de la nature et du nombre d'ancres de synténie utilisées. Dans les projets présentés, nous avons construit les cartes de synténie entre le chien et quatre génomes référents par l'utilisation de plus de 14 000 orthologues en relation 1:1 entre les paires d'organismes. Cependant, les gènes codant pour des protéines représentent à peine 2% du génome humain alors que plus de 40% de son génome s'aligne avec celui de la souris, par exemple. Il est donc envisageable d'utiliser, non seulement, les gènes, mais plus exhaustivement toutes les régions uniques et très conservées entre les génomes à comparer. Cette densification en points de comparaison permettra d'affiner la détection et, par conséquent, d'augmenter, probablement, le nombre de segments conservés identifiés. Ainsi, la distance évolutive estimée selon le nombre de réarrangements intra- et inter-chromosomiques entre plusieurs génomes comparés sera mieux représentée (Chaisson *et al*., 2006).

Bien qu'AutoGRAPH, dans son module de données pré-insérées inclut l'affichage graphique de l'orientation de la transcription des gènes (+/-), nous n'avons pas tenu compte de ce critère, parfois référencé sous le terme de micro-synténie, pour identifier les segments conservés. Or, plusieurs travaux (Tanimoto *et al*., 1999, Poyatos *et al*., 2006 ; Kolesov *et al*., 2007), montrent que l'ordre des gènes et l'orientation de la transcription sont corrélés en favorisant une régulation commune par les facteurs de transcription. Là encore, prendre en compte la conservation de l'ordre et de l'orientation entre différents gènes orthologues permettrait d'affiner la détection des segments conservés entre deux génomes.

Finalement, l'amélioration de l'identification des blocs conservés au cours de l'évolution nécessitera de perfectionner des programmes pré-existants, voire de créer de nouveaux outils permettant aussi d'intégrer et de comparer les nouveaux génomes séquencés. La disponibilité d'un nombre toujours plus grand de génomes séquencés implique de proposer de nouvelles interfaces graphiques de comparaison de génomes afin de faciliter la mise en évidence des ruptures de synténie partagées et des segments conservés. Dans ce but, le projet "HomoloBlock" a été initié très récemment au laboratoire. L'objectif du programme est de (i) construire des cartes de synténie entre un chromosome de référence et de multiples génomes de référence (ne plus se limiter à deux génomes testés), (ii) d'identifier les ruptures de synténie communes ou spécifiques à tous les organismes comparés et enfin, (iii) d'intégrer au niveau du chromosome de référence les caractéristiques de séquence (taux de GC) et les vecteurs évolutifs et fonctionnels tels que les

duplications segmentaires ou les familles de gènes (annotation Gene Ontology). La figure 24 présente l'ébauche du programme (**Fig.24**).

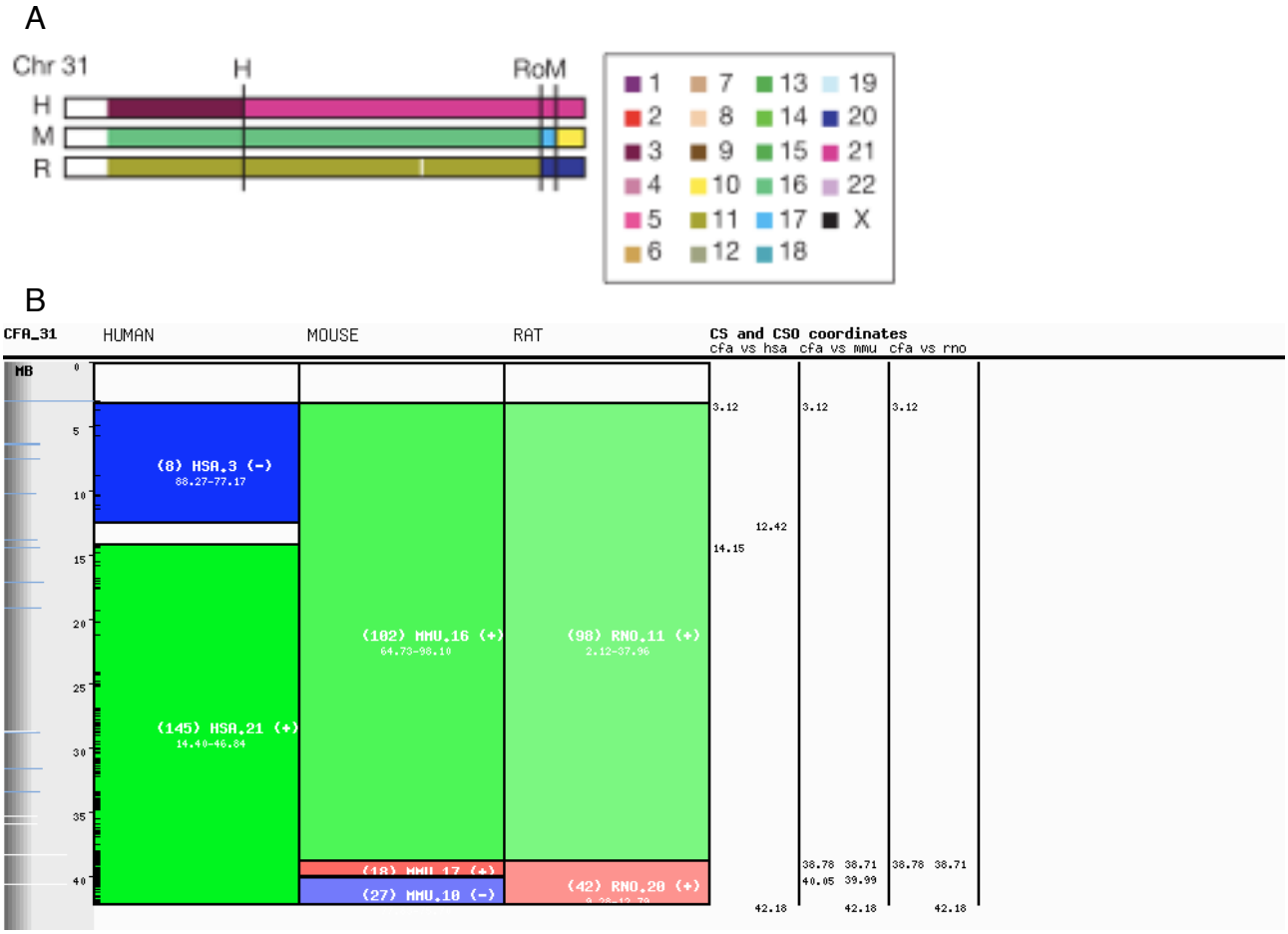


Figure 24 : Carte de synténie du chromosome 31 canin (CFA_31) avec les génomes humains, souris et rat.

A) Représentation graphique tirée de Lindblad-Toh *et al.*, 2005

B) Sortie graphique automatisée du programme Homoloblock. Les duplications intra et inter- chromosomiques sont illustrées par les barres blanches et bleues le long du chromosome 31 canin représenté en gris (à gauche). Chaque segment conservé ordonné est représenté en couleur et les limites des CSOs (correspondant aux ruptures de synténie) sont affichées sur la droite.

2.3. Intégrer l'histoire évolutive du chien à l'annotation de son génome

Dans les articles 3 et 4 (Derrien *et al.*, 2007 ; Derrien *et al.*, en préparation), nous avons amélioré l'annotation du génome du chien par une approche comparative en identifiant au total ~400 nouvelles prédictions de gènes canins et/ou de relations d'orthologies avec d'autre génomes mammifères (respectivement, 115 et 285 dans les deux études). Au-delà de l'identification de ces

gènes canins, il est intéressant de spéculer sur l'absence d'une centaine de gènes canins par ailleurs annotés en relation d'orthologie 1:1:1:1 entre quatre génomes mammifères. Nos travaux ont permis de montrer que plusieurs arguments, tels que l'absence d'alignement de séquences avec les orthologues référents, la présence de famille de gènes, la réduction de la taille des intervalles canins, l'absence d'alignements avec des ESTs, la localisation physique dans les régions télomériques et le contenu élevé en GC, offrent des hypothèses de scénarios de perte de gènes chez le chien. Mais, peut-on corréler le plus faible nombre de gènes canins avec son histoire évolutive particulière ? Le processus de domestication du chien a changé radicalement son mode de vie, et, par conséquent, les contraintes environnementales associées, telles qu'une disponibilité de nourriture plus abondante et régulière, une certaine forme de protection apportée par l'homme, un apport possible de soins ainsi que des échanges sociaux et cognitifs. De plus, le nombre relativement restreint d'individus fondateurs de la population canine domestiquée, les phénomènes migratoires humains (accompagnés de chiens), ont entraîné de fortes variations du pool et des fréquences alléliques et la fixation de mutations dans le génome du chien. Puis, le second goulet d'étranglement correspondant à la création des races canines modernes par l'homme a substitué l'effet de la sélection naturelle par une véritable sélection artificielle. Ainsi, les chiens sélectionnés pour des critères phénotypiques et/ou comportementaux ont pu survivre et se reproduire même si, dans le même temps, d'autre mutations moins avantageuses ont pu être sélectionnées. La relaxation des contraintes environnementales sur une population a pu modifier les pressions de sélection qui agissent sur le génome, en relaxant par exemple le maintien de l'intégrité de certaines régions fonctionnelles et permettre l'accumulation de mutations non-synonymes. Cela pourrait expliquer, en partie, que les alignements de séquences entre gènes orthologues ne permettent pas de retrouver des correspondances avec d'autres espèces.

Cependant, le taux de mutations dans le génome du chien est probablement trop faible et la vitesse d'accumulation des mutations trop lente pour rendre compte des pertes de gènes et des variations phénotypiques associées à la population canine depuis sa domestication il y a 14 000 ans ; sauf si l'on attribue un rôle déterminant aux contraintes exercées par l'homme dans la sélection du pool génétique des individus à l'origine de l'espèce canine actuelle. Ainsi, d'autres caractéristiques du génome canin sont à prendre en compte avec, en particulier, les variations de la longueur des répétitions dans les gènes et l'influence des rétrotransposons de type SINEC_Cf. Pour le premier cas, Fondon *et al*. ont reporté que des changements graduels de la longueur des éléments répétés de type microsatellite dans des gènes du développement (*Runx-2* ou *alx-4*, par exemple) entraînaient des variations des phénotypes associés (morphologies des membres ou de la face) (Fondon *et al*., 2004). Alors que peu de substitutions (essentiellement synonymes) sont retrouvées en comparant les

gènes du développement entre différentes races, les auteurs proposent que la diversité morphologique de la population canine serait corrélée à l'hyper-mutabilité des séquence répétées incluses dans les gènes du développement. Dans le deuxième cas, Wang *et al*. ont montré que près la moitié (48%) des gènes canins possédait au moins un SINE de type SINEC_Cf et que 15% des SINEC_Cf présents dans les gènes étaient polymorphes (présent/absent) (Wang *et al*., 2005). L'influence de ces SINEs selon leurs insertions aléatoires dans les gènes (exons, sites d'épissage...) ou les sites de régulations peut modifier considérablement l'expression des gènes associés et fournir des arguments à la perte de gènes canins et à la variabilité phénotypique entre les races. Ainsi, dans l'article 4, nous avons analysé tous les intervalles canins (COILs) pour plusieurs caractéristiques dont la proportion et la fréquence de séquences répétées (SINE et autres familles de séquences répétées). Il serait intéressant d'étudier les variations des longueurs des répétitions pour les gènes canins et de comparer ces variations par rapport aux gènes orthologues de référence. De plus, nous avons calculé le nombre et la proportion de SINEC_Cf dans les COILs pour chaque étape de notre processus d'identification de gènes canins. Nous n'avons pas mis en évidence de différence significative pour la proportion de SINEC_Cf entre des COILs pris aléatoirement dans le génome canin et les COILs testés dans notre étude. Il serait cependant envisageable d'analyser la répartition des SINEs dans les gènes canins et, non plus seulement, au sein des intervalles. En effet, outre le nombre, c'est plutôt la localisation des insertions de SINEs au sein même des séquences de gènes qui pourrait être responsable d'une pseudogénéisation et donc d'une absence de correspondance entre les gènes des génomes de référence et le génome canin.

Conclusion générale

Ces trois dernières années, la production de ressources de cartographie et de séquences du génome canin a fourni les entités biologiques nécessaires à la caractérisation de son génome et à l'établissement de ce modèle en génétique. Durant la première partie de ma thèse, j'ai développé le programme AutoGRAPH qui a permis, tout d'abord, d'intégrer les différentes données disponibles (séquences, carte RH, carte cytogénétique...), puis, de mettre en évidence que la complémentarité de ces ressources permettait d'exploiter l'information génétique disponible de façon optimale. Les fonctionnalités du programme ont été enrichies par la possibilité de construire des cartes comparées ou cartes de synténie entre deux ou trois génomes. La construction des cartes entre le génome du chien et d'autres génomes mammifères (homme - chimpanzé - souris - rat) a permis d'identifier et de caractériser l'ensemble des segments conservés (CS), des segments conservés ordonnés (CSO) et des ruptures de synténie ayant eu lieu au cours de l'évolution des génomes étudiés. De plus, la formalisation de la comparaison de génomes a révélé la forte conservation de l'ordre des gènes au sein des segments conservés ordonnés entre le chien et les génomes mammifères considérés.
Outre la caractérisation des blocs conservés au cours de l'évolution, la connaissance du génome canin nécessite d'identifier les séquences fonctionnelles qui le compose et, plus particulièrement, les gènes codant pour des protéines. Ainsi, durant la seconde partie de ma thèse, deux projets ont permis d'améliorer l'annotation du génome canin par une approche combinant à la fois le respect de l'ordre des gènes entre les génomes et des alignements de séquences ciblés sur le génome canin. Un premier projet a redéfini la localisation de plus d'une centaine de gènes préalablement assignés au chromosome non assemblé. Le second projet nous a permis de définir 285 nouveaux gènes canins et/ou nouvelles relations d'orthologie avec quatre génomes de référence (homme - chimpanzé - rat - souris) et différents mécanismes évolutifs sont suggérés mettant en relation la nature des gènes, la présence de famille de gènes et la composition en séquences pour expliquer la perte de gènes chez le chien.

La connaissance du génome du chien va sans nul doute être améliorée dans les années à venir. En effet, le chien présente le double avantage d'être un modèle pertinent pour l'identification de gènes responsables de maladies génétiques et aussi pour sa position en tant que groupe frère (Laurasiathérien) de l'homme et des rongeurs (Euarchontoglires) dans l'arbre phylogénétique des mammifères. Le génome du chien présente, par ailleurs, moins de duplications segmentaires que les génomes humain et murin, *a priori* moins de gènes et d'expansion de familles de séquences

répétées (excepté les SINEC_Cf). Il semble donc qu'il ait retenu des caractéristiques du génome ancestral des euthériens qui en font un bon "point d'ancrage" pour les études comparatives avec d'autres séquences des génomes mammifères telles que celles du "Mammalian Genome Project".

De plus, l'énorme diversité morphologique des ~400 races actuelles de chien offre des perspectives excitantes pour l'analyse de l'impact de l'histoire évolutive du chien (domestication puis sélection des races) sur les caractéristiques propres à son génome. Ainsi, de nombreuses questions restent sans réponse : Y a-t-il un biais dans l'usage des codons ? Dans la composition en base ? La sélection a-t-elle favorisé la présence de gènes plus courts chez le chien ? La comparaison du génome du chien avec d'autres génomes de la famille des canidés *(*loup, coyote, chacal*)* devrait sûrement nous permettre de corréler l'influence de ces événements sélectifs sur les caractéristiques propres au génome du chien.

Pour cela, les études de génomique comparative sont et, vont être, de plus en plus incontournables à la phase d'analyse fonctionnelle et évolutive de la séquence d'un organisme. Aussi, l'emploi du terme "décrypter" un génome, comme on a pu l'entendre dans les médias à propos du séquençage du génome humain, est sûrement légèrement abusif voire peut-être trop optimiste (à court terme). Il est certain que les techniques de séquençage des génomes vont continuer à se perfectionner pour des coûts encore plus faibles. À titre d'exemple, le séquençage d'un nouveau génome humain (celui du James Watson!) a été réalisé dans le temps record d'un mois et pour moins d'un million de dollars par la société 454 Life Science et le Baylor College of Medecine aux USA. Par comparaison, le coût du projet public du séquençage humain (Human Genome Consortium), complété il y a quatre ans, a été évalué à trois milliards de dollars pour 13 ans de travaux (REF http://www.nature.com/news/2007/070528/full/news070528-10.html). Cependant, comme nous l'avons vu, la séquence d'un génome ne représente pas une fin en soi mais plutôt le point de départ à de nouvelles découvertes. Six ans après la publication initiale du génome humain (Lander *et al*., 2001 ; Venter *et al*., 2001) la communauté scientifique s'oriente vers un consensus autour de 22 000 gènes humains codant pour des protéines, on ne perçoit que depuis peu l'influence des séquences fonctionnelles non-codantes. Les résultats du projet ENCODE (ENCyclopedia Of Dna Elements) (Encode Project Consortium, 2007), visant à annoter de façon exhaustive 1% (30 Mb) du génome humain, ont révélé que la plupart des bases de l'ADN sont transcrites, ce qui suggère des mécanismes beaucoup plus complexes pour la régulation d'un organisme.

L’amélioration des techniques de génomique comparative permettra de mieux identifier les séquences fonctionnelles (codantes et non-codantes) et les mécanismes évolutifs sous-jacents car "Rien n'a de sens en biologie, si ce n'est à la lumière de l'évolution” (Theodosius Dobzhansky).

IV. Bibliographie

1) Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, Volff JN, Schartl M. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics. 2002 May;161(1):259-67. PMID: 12019239

2) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. PMID: 2231712

3) Arumugam M, Wei C, Brown RH, Brent MR. Pairagon+N-SCAN_EST: a model-based gene annotation pipeline. Genome Biol. 2006;7 Suppl 1:S5.1-10. Epub 2006 Aug 7. PMID: 16925839

4) Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006 Nov 9;444(7116):171-8. PMID: 17086204

5) Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. Science. 2002 Aug 9;297(5583):1003-7. PMID: 12169732

6) Barrett JH. Genetic mapping based on radiation hybrid data. Genomics. 1992 May;13(1):95-103. PMID: 1577497

7) Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002 Jan;12(1):177-89. PMID: 11779843

8) Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004 May;14(5):988-95. PMID: 15123596

9) Bjornerfeldt S, Webster MT, Vila C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Res. 2006 Aug;16(8):990-4. PMID: 16809672

10) Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 2006;7(5):R43. Epub 2006 May 24. PMID: 16723033

11) Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science. 2003 Feb 28;299(5611):1391-4. PMID: 12610304.

12) Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. Nat Rev Genet. 2004 Jun;5(6):456-65. PMID: 15153998

13) Bourque G, Pevzner PA, Tesler G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res. 2004 Apr;14(4):507-16. PMID: 15059991

14) Breen M, Hitte C, Lorentzen TD, Thomas R, Cadieu E, Sabacan L, Scott A, Evanno G, Parker HG, Kirkness EF, Hudson R, Guyon R, Mahairas GG, Gelfenbeyn B, Fraser CM, Andre C, Galibert

F, Ostrander EA. An integrated 4249 marker FISH/RH map of the canine genome. BMC Genomics. 2004 Sep 13;5(1):65. PMID: 15363096

15) Breen M, Jouquand S, Renier C, Mellersh CS, Hitte C, Holmes NG, Cheron A, Suter N, Vignaux F, Bristow AE, Priat C, McCann E, Andre C, Boundy S, Gitsham P, Thomas R, Bridge WL, Spriggs HF, Ryder EJ, Curson A, Sampson J, Ostrander EA, Binns MM, Galibert F. 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. 2001 Oct;11(10):1784-95. PMID: 11591656

16) Breen M, Thomas R, Binns MM, Carter NP, Langford CF. Reciprocal chromosome painting reveals detailed regions of conserved synteny between the karyotypes of the domestic dog (Canis familiaris) and human. Genomics. 1999 Oct 15;61(2):145-55. PMID: 10534400

17) Brent MR, Guigo R. Recent advances in gene structure prediction. Curr Opin Struct Biol. 2004 Jun;14(3):264-72. Review. PMID: 15193305

18) Burge, C. B. and Karlin, S. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. . 1998. 8, 346-354. PMID: 9666331

19) Byrne KP, Wolfe KH. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D452-5. PMID: 16381909

20) Chaisson MJ, Raphael BJ, Pevzner PA. Microinversions in mammalian evolution. Proc Natl Acad Sci U S A. 2006 Dec 26;103(52):19824-9. Epub 2006 Dec 22. PMID: 17189424

21) Clark LA, Wahl JM, Rees CA, Murphy KE. Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc Natl Acad Sci U S A. 2006 Jan 31;103(5):1376-81. PMID: 16407134

22) Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. 2005 Dec;21(12):673-82. Epub 2005 Oct 19. Review. PMID: 16242204

23) Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. The Ensembl automatic gene annotation system. Genome Res. 2004 May;14(5):942-50. PMID: 15123590

24) Derrien T, Andre C, Galibert F, Hitte C. Analysis of the unassembled part of the dog genome sequence: chromosomal localization of 115 genes inferred from multispecies comparative genomics. J Hered. 2007;98(5):461-7. Epub 2007 Jun 15. PMID: 17573383

25) Derrien T, Andre C, Galibert F, Hitte C. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics. 2007 Feb 15;23(4):498-9. PMID: 17145741

26) Djebali S, Delaplace F, Crollius HR. Exogean: a framework for annotating protein-coding genes in eukaryotic genomic DNA. Genome Biol. 2006;7 Suppl 1:S7.1-10. Epub 2006 Aug 7. PMID: 16925841

27) Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 2003 Oct;13(10):2213-9. PMID: 14525923

28) 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 Jun 1;21(11):2596-603. PMID: 15713731

29) Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439-40. PMID: 16082012

30) ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14;447(7146):799-816. PMID: 17571346

31) Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug 19;5:113. PMID: 15318951

32) Eichler EE, Sankoff D. Structural dynamics of eukaryotic chromosome evolution. Science. 2003 Aug 8;301(5634):793-7. Review. PMID: 12907789

33) Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol. 2006 Sep;4(9):e286. PMID: 16933976

34) Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2006 Jul;2(7):e102. PMID: 16839186

35) Faraut T, de Givry S, Chabrier P, Derrien T, Galibert F, Hitte C, Schiex T. A comparative genome approach to marker ordering. Bioinformatics. 2007 Jan 15;23(2):e50-6. PMID: 17237105

36) Fischer G, James SA, Roberts IN, Oliver SG, Louis EJ. Chromosomal evolution in Saccharomyces. Nature. 2000 May 25;405(6785):451-4. PMID: 10839539

37) Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genet. 2006 Mar;2(3):e32. Epub 2006 Mar 10. PMID: 16532063

38) Fitch WM. Distinguishing homologous from analogous proteins. Syst. Zool. 1970.19:99-106 PMID: 8748022

39) Fitch WM. Uses for evolutionary trees. Philos. Trans. R. Soc. London Ser. 1995. B 349:93-102.

40) Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269 (5223): 496-512. PMID 7542800

41) Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T. EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res. 2003 Jul 1;31(13):3742-5. PMID: 12824408

42) Fondon JW 3rd, Garner HR. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A. 2004 Dec 28;101(52):18058-63. Epub 2004 Dec 13. PMID: 15596718

43) Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953. Apr 25;171(4356):740-1. PMID: 13054694

44) Galibert F, Andre C, Hitte C. Dog as a mammalian genetic model. Med Sci (Paris). 2004 Aug-Sep;20(8-9):761-6. French. PMID: 15361341

45) Galtier N, Duret L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 2007 Jun;23(6):273-7. Epub 2007 Apr 5. Review. PMID: 17418442

46) Galtier N. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 2003 Feb;19(2):65-8. Review. PMID: 12547511

47) Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004 Apr 1;428(6982):493-521. PMID: 15057822

48) Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol. 2006 Sep 29;2(9):e133. PMID: 17009864

49) Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006;7 Suppl 1:S2.1-31. Epub 2006 Aug 7. PMID: 16925836

50) Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003 Oct;52(5):696-704. PMID: 14530136

51) Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W557-9. PMID: 15980534

52) Guyon R, Kirkness EF, Lorentzen TD, Hitte C, Comstock KE, Quignon P, Derrien T, Andre C, Fraser CM, Galibert F, Ostrander EA. Building comparative maps using 1.5x sequence coverage: human chromosome 1p and the canine genome. Cold Spring Harb Symp Quant Biol. 2003;68:171-7. No abstract available. PMID: 15338615

53) Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, Quignon P, Lowe JK, Renier C, Gelfenbeyn B, Vignaux F, DeFrance HB, Gloux S, Mahairas GG, Andre C, Galibert F, Ostrander EA. A 1-Mb resolution radiation hybrid map of the canine genome. Proc Natl Acad Sci U S A. 2003 Apr 29;100(9):5296-301. PMID: 12700351

54) Guyon R, Pearce-Kelling SE, Zeiss CJ, Acland GM, Aguirre GD. 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. 2007 Jul 11;13:1094-105. PMID: 17653054

55) Hardison RC. Comparative genomics. PLoS Biol. 2003 Nov;1(2):E58. PMID: 14624258

56) Hedan B, Corre S, Hitte C, Dreano S, Vilboux T, Derrien T, Denis B, Galibert F, Galibert MD, Andre C. Coat colour in dogs: identification of the merle locus in the Australian shepherd breed. BMC Vet Res. 2006 Feb 27;2:9. PMID: 16504149

57) Heger A, Ponting CP. Evolutionary rate analyses of orthologues and paralogues from twelve Drosophila genomes. 2007 (submit)

58) Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C, Evanno G, Pullar B, Cadieu E, Oza V, Lourgant K, Jaffe DB, Tacher S, Dreano S, Berkova N, Andre C, Deloukas P, Fraser C, Lindblad-Toh K, Ostrander EA, Galibert F. Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. Nat Rev Genet. 2005 Aug;6(8):643-8. PMID: 16012527.

59) Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. Nucleic Acids Res. 2007 Jan;35(Database issue):D610-7. Epub 2006 Dec 5. PMID: 17148474

60) Hurst LD, Pal C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 2004 Apr;5(4):299-310. Review. No abstract available. PMID: 15131653

61) Huxley THH. The Origin of Species. Westminst. Rev. 1860. 17:541-70

62) International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec 9;432(7018):695-716. Erratum in: Nature. 2005 Feb 17;433(7027):777. PMID: 15592404

63) Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004 Oct 21;431(7011):946-57. PMID: 15496914

64) Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354-7. PMID: 16381885

65) Kellis M, Patterson N, Birren B, Berger B, Lander ES. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. J Comput Biol. 2004;11(2-3):319-55. PMID: 15285895

66) Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, Venter JC. The dog genome: survey sequencing and comparative analysis. Science. 2003 Sep 26;301(5641):1898-903. PMID: 14512627

67) Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA. How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci U S A. 2007 Aug 28;104(35):13948-53. Epub 2007 Aug 20. PMID: 17709750

68) Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38. PMID: 16285863.

69) Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ. The UCSC genome browser database: update 2007. Nucleic Acids Res. 2007 Jan;35(Database issue):D668-73. PMID: 17142222

70) Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. PMID: 11237011

71) Lee JM, Sonnhammer EL. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 2003 May;13(5):875-82. Epub 2003 Apr 14. Erratum in: Genome Res. 2004 Dec;14(12):2510. PMID: 12695325

72) Leonard JA, Wayne RK, Wheeler J, Valadez R, Guillen S, Vila C. Ancient DNA evidence for Old World origin of New World dogs. Science. 2002 Nov 22;298(5598):1613-6. PMID: 12446908

73) 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 Dec 8;438(7069):803-19. PMID: 16341006

74) Louis EJ. Evolutionary genetics: making the most of redundancy. Nature. 2007 Oct 11;449(7163):673-4. PMID: 17928851

75) Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000 Nov 10;290(5494):1151-5. PMID: 11073452.

76) Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000 Jan;154(1):459-73. PMID: 10629003

77) Margulies EH, Vinson JP; NISC Comparative Sequencing Program, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC, Clamp M. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A. 2005 Mar 29;102(13):4795-800. Epub 2005 Mar 18. PMID: 15778292

78) McKay SD, Schnabel RD, Murdoch BM, Aerts J, Gill CA, Gao C, Li C, Matukumalli LK, Stothard P, Wang Z, Van Tassell CP, Williams JL, Taylor JF, Moore SS. Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. Anim Genet. 2007 Apr;38(2):120-5. PMID: 17302794

79) Mellersh CS, Hitte C, Richman M, Vignaux F, Priat C, Jouquand S, Werner P, Andre C, DeRose S, Patterson DF, Ostrander EA, Galibert F. An integrated linkage-radiation hybrid map of the canine genome. Mamm Genome. 2000 Feb;11(2):120-30. PMID: 10656926

80) Menotti-Raymond M, David VA, Agarwala R, Schaffer AA, Stephens R, O'Brien SJ, Murphy WJ. Radiation hybrid mapping of 304 novel microsatellites in the domestic cat genome. Cytogenet Genome Res. 2003;102(1-4):272-6. PMID: 14970716

81) Meunier J, Duret L. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 2004 Jun;21(6):984-90. Epub 2004 Feb 12. PMID: 14963104

82) Mignot E, Bell RA, Rattazzi C, Lovett M, Grumet FC, Dement WC. An immunoglobulin switchlike sequence is linked with canine narcolepsy. Sleep. 1994 Dec;17(8 Suppl):S68-76. PMID: 7701203

83) Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SM, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, Cook A, Cuff J, Das R, Davidow L, Deakin JE, Fazzari MJ, Glass JL, Grabherr M, Greally JM, Gu W, Hore TA, Huttley GA, Kleber M, Jirtle RL, Koina E, Lee JT, Mahony S, Marra MA, Miller RD, Nicholls RD, Oda M, Papenfuss AT, Parra ZE, Pollock DD, Ray DA, Schein JE, Speed TP, Thompson K, VandeBerg JL, Wade CM, Walker JA, Waters PD, Webber C, Weidman JR, Xie X, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Graves JA, Ponting CP, Breen M, Samollow PB, Lander ES, Lindblad-Toh K. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 2007 May 10;447(7141):167-77. PMID: 17495919

84) Miller W, Makova KD, Nekrutenko A, Hardison RC. Comparative genomics. Annu Rev Genomics Hum Genet. 2004;5:15-56. PMID: 15485342

85) Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, et al.. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62. PMID: 12466850

86) Muller C, Denis M, Gentzbittel L, Faraut T. The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W429-34. PMID: 15215424

87) Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers SN, Milan D, Ostrander EA, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'brien SJ, Pevzner PA, Lewin HA. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 2005 Jul 22;309(5734):613-7. PMID: 16040707

88) Nadeau JH, Taylor BA. Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci U S A. 1984 Feb;81(3):814-8. PMID: 6583681

89) Neff MW, Broman KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander EA, Rine J. A second-generation genetic linkage map of the domestic dog, Canis familiaris. Genetics. 1999 Feb;151(2):803-20. PMID: 9927471

90) Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997 Jul 22;94(15):7799-806. PMID: 9223266

91) Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 2005;39:121-52. Review. PMID: 16285855

92) Novozhilov AS, Karev GP, Koonin EV. Biological applications of the theory of birth-and-death processes. Brief Bioinform. 2006 Mar;7(1):70-85. Review. PMID: 16761366

93) O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D476-80. PMID: 15608241

94) 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 Oct 15;286(5439):458-62, 479-81. PMID: 10521336

95) Owen R. On the Archetype and Homologies of the Vertebrate Skeleton. London: Murray. 1848

96) Page RDM. and Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. 1997. Mol. Phyl. Evol., 2, 231-40. PMID: 9126565

97) Pan X, Stein L, Brendel V. SynBrowse: a synteny browser for comparative sequence analysis. Bioinformatics. 2005 Sep 1;21(17):3461-8. PMID: 15994196

98) Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L. Genetic structure of the purebred domestic dog. Science. 2004 May 21;304(5674):1160-4. PMID: 15155949

99) Passarge E, Horsthemke B, Farber RA. Incorrect use of the term synteny. Nat Genet. 1999 Dec;23(4):387. PMID: 10581019

100) Passarge E, Horsthemke B, Farber RA. Incorrect use of the term synteny. Nat Genet. 1999 Dec;23(4):387. PMID: 10581019

101) Petsko GA. Homologuephobia. Genome Biol. 2001. 2: COMMENT 1002a

102) Pevzner P, Tesler G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res. 2003 Jan;13(1):37-45. PMID: 12529304

103) Pevzner P, Tesler G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A. 2003 Jun 24;100(13):7672-7. PMID: 12810957

104) Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006 Sep 14;443(7108):167-72. Epub 2006 Aug 16. PMID: 16915236

105) Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. The Ensembl analysis pipeline. Genome Res. 2004 May;14(5):934-41. PMID: 15123589

106) Poyatos JF, Hurst LD. Is optimal gene order impossible? Trends Genet. 2006 Aug;22(8):420-3. Epub 2006 Jun 27. PMID: 16806566

107) Priat C, Jiang ZH, Renier C, Andr C, Galibert F. Characterization of 463 type I markers suitable for dog genome mapping. Mamm Genome. 1999 Aug;10(8):803-13. PMID: 10430667

108) Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007 Jan;35(Database issue). PMID: 17130148

109) Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 2007 Jul 6;317(5834):86-94. PMID: 17615350

110) Quignon P, Giraud M, Rimbault M, Lavigne P, Tacher S, Morin E, Retout E, Valin AS, Lindblad-Toh K, Nicolas J, Galibert F. The dog and rat olfactory receptor repertoires. Genome Biol. 2005;6(10):R83. PMID: 16207354

111) Raudsepp T, Fronicke L, Scherthan H, Gustavsson I, Chowdhary BP. Zoo-FISH delineates conserved chromosomal segments in horse and man. Chromosome Res. 1996 Apr;4(3):218-25. PMID: 8793207

112) Reese MG, Guigo R. EGASP: Introduction. Genome Biol. 2006;7 Suppl 1:S1.1-3. Epub 2006 Aug 7. PMID: 1692583

113) Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001 Dec 14;314(5):1041-52. PMID: 11743721

114) Rhodes M, Straw R, Fernando S, Evans A, Lacey T, Dearlove A, Greystrong J, Walker J, Watson P, Weston P, Kelly M, Taylor D, Gibson K, Mundy C, Bourgade F, Poirier C, Simon D, Brunialti AL, Montagutelli X, Gu'enet JL, Haynes A, Brown SD. A high-resolution microsatellite map of the mouse genome. Genome Res. 1998 May;8(5):531-42. PMID: 9582196

115) Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T. Genetic evidence for an East Asian origin of domestic dogs. Science. 2002 Nov 22;298(5598):1610-3. PMID: 12446907

116) Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 2006 Mar 16;440(7082):341-5. PMID: 16541074

117) Scherthan H, Schweizer D, Loidl J. Delineation of individual chromosomes of Saccharomyces cerevisiae by two-colour in situ hybridization. Trends Genet. 1993 Feb;9(2):41. PMID: 8456499

118) Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-mouse alignments with BLASTZ. Genome Res. 2003 Jan;13(1):103-7. Erratum in: Genome Res. 2004 Apr;14(4):786. PMID: 12529312.

119) Semon M, Duret L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 2006 Sep;23(9):1715-23. Epub 2006 Jun 6. PMID: 16757654

120) Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50. Epub 2005 Jul 15. PMID: 16024819

121) Sinha AU, Meller J. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. BMC Bioinformatics. 2007 Mar 8;8:82. PMID: 17343765

122) Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005 Feb 15;6:31. PMID: 15713233

123) Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 2002 Dec;18(12):619-20. PMID: 12446146

124) Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y. DDBJ with new system and face. Nucleic Acids Res. 2007 Oct 25. PMID: 17962300

125) Susumu Ohno. Evolution by gene duplication. 1970. Springer-Verlag. ISBN 0-04-575015-7.

126) Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG, Quignon P, Johnson GS, Parker HG, Fretwell N, Mosher DS, Lawler DF, Satyaraj E, Nordborg M, Lark KG, Wayne RK, Ostrander EA. A single IGF1 allele is a major determinant of small size in dogs. Science. 2007 Apr 6;316(5821):112-5. PMID: 17412960

127) Tanimoto K, Liu Q, Bungert J, Engel JD. Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice. Nature. 1999 Mar 25;398(6725):344-8. PMID: 10192336

128) Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001 Jan 9(1):22-8. PMID: 11125040

129) Tesler G. GRIMM: genome rearrangements web server. Bioinformatics. 2002 Mar;18(3):492-3. PMID: 11934753

130) Van Baren MJ, Brent MR. Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res. 2006 May;16(5):678-85. PMID: 16651666

131) Varshavsky A. Spalog and sequelog: neutral terms for spatial and sequence similarity. Curr. Biol. 2004. 14:R181-3. PMID: 17846424

132) Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. PMID: 11181995

133) Vila C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK. Multiple and ancient origins of the domestic dog. Science. 1997 Jun 13;276(5319):1687-9. PMID: 9180076

134) Vila C, Seddon J, Ellegren H. Genes of domestic mammals augmented by backcrossing with wild ancestors. Trends Genet. 2005 Apr;21(4):214-8. PMID: 15797616

135) Wall DP, Fraser HB, Hirsh AE. Detecting putative orthologs. Bioinformatics. 2003 Sep 1;19(13):1710-1. PMID: 15593400

136) Wang M, Buhler J, Brent MR. The effects of evolutionary distance on TWINSCAN, an algorithm for pair-wise comparative gene prediction. Cold Spring Harb Symp Quant Biol. 2003;68:125-30. PMID: 15338610

137) Wang W, Kirkness EF. Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res. 2005 Dec;15(12):1798-808. PMID: 16339378

138) Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737-8. PMID: 13054692

139) Watson JD. ADN, le secret de la vie. (2003) (Ed, Odile JACOB).

140) Wayne RK, Ostrander EA. Lessons learned from the dog genome. Trends Genet. 2007 Oct 24. PMID: 17963975

141) Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. A second-generation linkage map of the human genome. Nature. 1992 Oct 29;359(6398):794-801. PMID: 1436057

142) Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res. 2002 Jan 1;30(1):13-6. PMID: 11752242

143) Wilkins MH, Stokes AR, Wilson HR. Molecular structure of deoxypentose nucleic acids. Nature. 1953 Apr 25;171(4356):738-40. PMID: 13054693

144) Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586-91. PMID: 17483113

145) Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997 Oct;13(5):555-6. PMID: 9367129

146) Zheng D, Gerstein MB. A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol. 2006;7 Suppl 1:S13.1-10. Epub 2006 Aug 7. PMID: 16925835

147) Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J, Mural R. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics. 2005 Mar;21(6):703-10. Epub 2004 Sep 30. PMID: 15458983