Application of genome-wide single-nucleotide polymorphism arrays to understanding dog disease and evolution

Javier Quilez Oliete

Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona

Doctoral thesis 2012

Directors: **Olga Francino Martí** i **Laura Altet Sanahujes**

Summary

The generation of vast SNP repertoires from genome sequencing projects together with rapid improvements in large-scale SNP genotyping allowed the development of high-density genome-wide SNP microarrays in many animal species. This thesis presents two examples of the application of SNP arrays to understanding disease and evolution in the dog, whose evolutionary history makes it a suitable animal model for trait mapping and a fascinating case of artificial selection.

In the first example, motivated by the fact that only a certain proportion of individuals infected with *Leishmania* are susceptible to develop clinical leishmaniasis disease we tried to dissect how and to what extent host genetics determines whether infected individuals progress to clinical disease. Firstly, we tried to map loci affecting the phenotype and for the strongest associations we tested whether their haplotype structure correlated with the affection status and examined their nearby genetic content. Secondly, we estimated the heritability of the trait and assessed the capability to predict the phenotype from genomic information.

In the second case, we searched for genomic footprints of selection in the Boxer breed. We presented a novel selective sweep of >8 Mb on chromosome 26. Hinted by the presence of another selective sweep on chromosome 1 previously associated with canine brachycephaly, characterized by severe shortening of the muzzle and a breed-defining trait of the Boxer, we investigated on the relationship between the selective sweep on chromosome 26 and this trait. We tried to prove the selective sweep is representative of the Boxer breed and that it is also present in other brachycephalic breeds but absent in non-brachycephalic breeds and wolf. Finally, we examined the genetic content of the selective sweep to find putative targets of selection and potential undesired health consequences for the breeds bearing the phenotype.

***Keywords**:* dog, genetic susceptibility to leishmaniasis, association mapping, heritability, phenotype prediction, selective sweep, brachycephaly

Resum

El descobriment d'un gran ventall de SNPs arrel dels projectes de seqüenciació de genomes, juntament amb les ràpides millores en el seu genotipatge a gran escala, van permetre el desenvolupament en moltes espècies animals de xips d'alta densitat de SNPs distribuïts pel genoma. Aquesta tesi presenta dos exemples de l'aplicació dels xips de SNPs per tal d'entendre malaltia i evolució en el gos, la història evolutiva del qual el converteix en un model animal apropiat per al mapatge de caràcters i en un fascinant cas de selecció artificial.

En el primer exemple, motivats pel fet que només una certa proporció dels individus afectats per *Leishmania* són susceptibles a desenvolupar clínicament la malaltia leishmaniasi, hem intentat diseccionar com, i fins a quin punt, la genètica de l'hoste determina que els individus infectats progressin cap a la malaltia clínica. Primer, hem intentat localitzar loci que afectin el fenotip, i per les associacions més fortes, analitzat si la seva estructura haplotípica correlaciona amb l'estat d'afectació, alhora que hem examinat el seu contingut genètic més proper. Segon, hem estimant la heretabilitat del caràcter i avaluat la capacitat de predir el fenotip a partir de la informació genòmica.

En el segon cas, hem buscat empremtes genòmiques causades per la selecció en la raça Bòxer. Presentem un nou *selective sweep* de >8 Mb en el cromosoma 26. Guiats per la presència d'un altre *selective sweep* en el cromosoma 1 prèviament associat amb la braquicefàlia canina, caracteritzada per un escurçament sever del musell i un tret característic del Bòxer, hem investigat sobre la relació del *selective sweep* en el cromosoma 26 i aquest tret. Hem intentat demostrar que el *selective sweep* és representatiu de la raça Bòxer i que està també present en altres races braquicèfales però absent en races no braquicèfales i en el llop. Finalment, hem examinat el contingut genètic del *selective sweep* per tal de trobar dianes per a la selecció així com conseqüències no desitjades per a la salut de les races que presenten el fenotip.

Paraules clau: gos, susceptibilitat genètica a leishmaniasi, mapatge per associació, predicció del fenotip, *selective sweep*, braquicefàlia

Contents

List of Boxes

List of Tables

List of Figures

LIST OF PUBLICATIONS

This thesis is based on the work of contained in the following papers:

Paper I **Quilez J**, Martínez V, Woolliams JA, Sanchez A, Pong-Wong R, Kennedy LJ, Quinnell RJ, Ollier WER, Roura X, Ferrer L *et al.*, (2012) Genetic control of canine leishmaniasis: genome-wide association study and genomic selection analysis. PloS one 7(4):e35349.

Paper II **Quilez J**, Short AD, Martínez V, Kennedy LJ, Ollier W, Sanchez A, Altet L, Francino O, (2011) A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. BMC Genomics 12:339.

RELATED WORK BY THE AUTHOR

(Not included in the thesis)

Martínez V, **Quilez J**, Sanchez A, Roura X, Francino O, Altet L, (2011) Canine leishmaniasis: the key points for qPCR result interpretation. Parasites & vectors 4:57.

Olsson M, Meadows JRS, Truvé K, Rosengren Pielberg G, Puppo F, Mauceli E, **Quilez J**, Tonomura N, Zanna G, Docampo MJ *et al.*, (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 7(3):e1001332.

ABBREVIATIONS

AUC	Area under the ROC curve
BMI	Body mass index
bp	Base pair
CanL	Canine leishmaniasis
CFA	*Canis familiaris* autosome
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
FDR	False discovery rate
g	Fraction of correct predictions
GBLUP	Genomic best linear unbiased prediction
gDNA	Genomic DNA
GEBV	Genomic estimate of breeding value
GWAS	Genome-wide association study
h^2	Heritability
HWE	Hardy-Weinberg equilibrium
Kb	Kilo bp
λ	Lambda or genomic inflation factor
LD	Linkage disequilibrium
MAF	Minor allele frequency
Mb	Mega bp
MCMC	Markov chain Monte Carlo
MDS	Multidimensional scaling
mtDNA	Mitochondrial DNA
OR	Odds ratio
π	Proportion of SNPs with zero effect on the phenotype
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
ROH	Region of homozygosity (ROHs for plural)
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphism (SNPs for plural)

r	Accuracy
VL	Visceral leishmaniasis
WGA	Whole-genome amplification
WGP	Whole-genome marker-enabled prediction

INTRODUCTION

From genome sequencing projects to high-density genome-wide SNP microarrays

Genome sequencing projects

The 2000s witnessed the completion of the genome sequencing projects of many animal species (**Table 1**). A direct implication of these sequencing efforts was the generation of unprecedented repertoires, in the order of millions, of single nucleotide polymorphisms (SNPs), which in many cases has led to extensive maps of genetic variation (**The International HapMap Consortium 2005**, **2007**; **Bovine HapMap Consortium** *et al.* **2009**). The domestic dog (*Canis lupus familiaris*) was not an exception and a first 1.5x coverage dog sequence (**Kirkness** *et al.* **2003**) was followed by the release of a high-quality 7.5x coverage sequence of the dog genome in 2005 (**Lindblad-Toh** *et al.* **2005**), providing a map of >2.8 million SNPs.

High-density genome-wide SNP microarray technology

The deposition of millions of SNPs into public databases was accompanied by rapid improvements in SNP genotyping technology (**Kennedy** *et al.* **2003**; **Gunderson** *et al.* **2005**), which enabled the development of commercially available high-density genome-wide SNP microarrays for several species (**Table 1**); hereafter referred as SNP arrays for simplicity. Essentially, these large-scale SNP arrays are based on three steps: DNA amplification, hybridization and single-base labeled extension (**Figure 1**).

Table 1. Genome sequencing projects. This table is not all-inclusive and focuses on *Homo sapiens* and on those species that are typically used as animal models in biomedicine. Although genome sequencing projects have provided the vast majority of the SNPs repertoires, current SNPs arrays have in some cases benefited from subsequent re-sequencing projects, especially with the arrival of next-generation sequencing technologies (e.g. **Vaysse** *et al.* **2011**).

Species genome	**Year of publication**	**References**	**# SNPs in array[1] (Company)**
Human (*Homo sapiens*)	2001	(**Lander** *et al.* **2001**; **Venter** *et al.* **2001**)	4,300,000 (Illumina)
Mouse (*Mus musculus*)	2002	(**Chinwalla** *et al.* **2002**)	623,000 (Affymetrix)
Rat (*Rattus norvegicus*)	2004	(**Gibbs** *et al.* **2004**)	__[2]
Chicken (*Gallus gallus*)	2004	(**Hillier** *et al.* **2004**)	54,000 (Illumina)
Domestic dog (*Canis lupus familiaris*)	2005	(**Lindblad-Toh** *et al.* **2005**)	174,000 (Illumina)
Pig (*Sus scrofa*)	2007	(**Humphray** *et al.* **2007**)	62,000 (Illumina)
Cattle (*Bos taurus*)	2009	(**Bovine Genome Sequencing and Analysis Consortium** *et al.* **2009**)	778,000 (Illumina)
Horse (*Equus caballus*)	2009	(**Wade** *et al.* **2009**)	54,000 (Illumina)
Sheep (*Ovis aries*)	__[3]	__	54,000 (Illumina)
Goat (*Capra aegagrus hircus*)	__[4]	__	53,000 (Illumina)

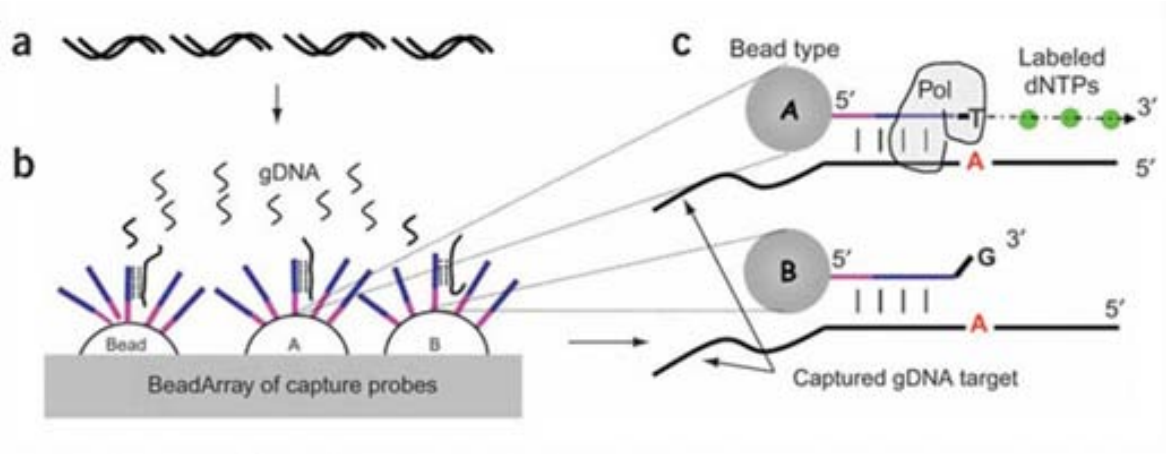
[1]Comercially available SNP array of highest density (as of March 2012).

[2]Expression arrays only.

[3]The sheep genome reference sequence is in progress (**International Sheep Genomics Consortium** *et al.* **2010**).

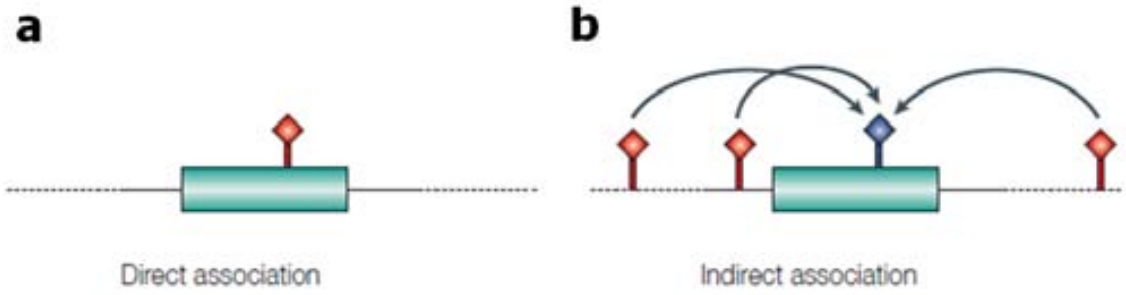
[4]The goat genome sequencing project is at an initial stage (**Personal communication**).

Figure 1. Whole-genome genotyping on DNA arrays. (a) whole-genome amplification (WGA) to generate large amounts of amplified genomic DNA (gDNA). **(b)** Hybridization of the WGA product to a specific and sensitive oligonucleotide probe array (50-mers). **(c)** An array-based allelic-specific primer extension reaction that incorporates multiple labeled dNTP nucleotides into the appropriate allelic probe, followed by a sensitive detection and signal amplification step to read the incorporated labels.



Gunderson et al. (2005) Nat Genet 37(5):549-554

Figure 2. GWAS exploit linkage disequilibrium to map trait-causing loci.



Modified from Hirschhorn & Daly (2005) Nat Rev Genet 6:95-108

Canine SNP arrays

To date, four whole-genome canine SNP arrays have been made commercially available. The first SNP array for the dog was developed in 2007 by Karlsson *et al.* (**Karlsson** *et al.* **2007**) and contained ~27,000 markers that were of high-quality and effective. SNP selection from the 2.8 million SNPs in the canine genetic map (**Lindblad-Toh** *et al.* **2005**) was on the basis of uniform spacing (average spacing = 87 Kb) and to be informative in the >10 breeds used. The SNP array was commercialized by Affymetrix [27K Canine Affymetrix v1, Affymetrix, Santa Clara, CA, http://www.affymetrix.com]. A second version (~50,000 SNPs, average spacing = 47 Kb) containing the majority of SNPs from the first version plus other markers from the same canine genetic map and using the same criteria was released later [50K Canine Affymetrix v2]. A third array with ~22,000 SNPs (average spacing = 125 Kb) was developed by Illumina [CanineSNP20 BeadChip panel, Illumina, San Diego, CA, http://www.illumina.com]. The three SNP arrays have been similarly used in canine genetic studies (**Table 2** and **Table 3**).

A canine SNP array of higher density with ~174,000 SNPs (average spacing = 13 Kb) was developed by The LUPA Consortium (http://www.eurolupa.org/) including SNPs from the 2.8 million SNPs from the canine genome project and additional SNPs discovered by resequencing gaps in the dog genome (**Vaysse** *et al.* **2011**). The SNP array is commercialized by Illumina [CanineHD BeadChip panel] and it is increasingly used in canine genetic studies (**Table 2** and **Table 3**).

Table 2. Chronology of canine genome-wide association studies for simple and complex traits and diseases. Breeds used for the initial GWAS are shown, yet other breeds may have been additionally used for replication and/or fine-mapping. When possible, the actual number of SNPs used for analysis after quality control filters are presented; the canine SNP array is also indicated: [A]27K Canine Affymetrix v1 (~27,000 SNPs), [B]50K Canine Affymetrix v2 (~50,000 SNPs), [C]CanineSNP20 BeadChip panel, Illumina (~22,000 SNPs) and [D]CanineHD BeadChip panel, Illumina (~174,000 SNPs).

Trait	Type of trait	Breed	# dogs		# markers	Reference
			Cases	Controls		
Within-breed GWAS: simple traits						
Dorsal ridgeless	Autosomal recessive (breed-defining)	Rhodesian ridgeback	9	12	26,625[A]	(**Karlsson** *et al.* **2007**)
White spotting	Semi-dominantly inherited	Boxer	10	10	26,625[A]	(**Karlsson** *et al.* **2007**)
Canine ectodermal dysplasia (Hairless dogs)	Monogenic autsomal semidominant (breed-defining)	Chinese crested dog	20	19	12,355[B]	(**Drogemuller** *et al.* **2008**)
Degenerative myelopathy	Monogenic (disease)	Pembroke Welsh corgi	38	17	49,663[B]	(**Awano** *et al.* **2009**)
Furnishing	Simple (breed-defining)	Daschund[1]	33	53	51,071[B]	(**Cadieu** *et al.* **2009**)
Hair length	Simple (breed-defining)	Daschund[1]	29	57	51,071[B]	(**Cadieu** *et al.* **2009**)

Table 2 (*Continued*)

Curly coat	Simple (breed-defining)	Portuguese water dog[1]	35	45	56,395[B]	(**Cadieu** *et al.* **2009**)
Neuronal ceroid lipofuscinoses	Autosomal recessive (disease)	American Staffordshire terrier	12	10	174,000[D]	(**Abitbol** *et al.* **2010**)[2]
Primary ciliary dyskinesia	Autosomal recessive (disease)	Old English sheepdog	5	15	50,000[B]	(**Merveille** *et al.* **2011**)
Benign familial juvenile epilepsy	Autosomal dominant (disease)	Lagotto Romagnolo	28	112	17,273[A]	(**Seppälä** *et al.* **2011**)
Within-breed GWAS: complex traits						
Atopic dermatitis	Complex (disease)	Golden retriever	25	23	22,591[C]	(**Wood** *et al.* **2009**)
Canine SLE-related disease complex	Complex (disease)	Nova Scotia duck tolling retriever	81	57	22,000[C]	(**Wilbe** *et al.* **2010**)
Canine cone-rod dystrophy 3	Undetermined inheritance (disease)	Glen of Imaal Terrier	21	22	50,000[B]	(**Goldstein** *et al.* **2010**)
Compulsive disorder	Complex (disease)	Doberman pinscher	92	68	14,700[B]	(**Dodman** *et al.* **2010**)
Familial Shar-Pei Fever	Complex (disease)	Shar-pei	24	17	17,227[A, B]	(**Olsson** *et al.* **2011**)

Table 2 (*Continued*)

Congenital megaesophagus	Several loci responsible (disease)	German shepherd dog	19	177	48,415[B]	(**Tsai** *et al.* **2012**)
Pancreatic acinar atrophy	Polygenic (disease)	German shepherd dog	100	79	48,415[B]	(**Tsai** *et al.* **2012**)
Intervertebral disc calcification	Multifactorial etiology (disease)	Dachshund	48	46	109,000[D]	(**Mogensen** *et al.* **2011**)
Abnormal immunoglobulin E serum levels in response to *D. farina*	Multifactorial (disease)	West Highland white terrier	31	24	48,157[B]	(**Barros Roque** *et al.* **2011**)
Myxomatous mitral valve disease	Polygenic threshold trait (disease)	Cavalier King Charles Spaniel	139	102	84,693[D]	(**Madsen** *et al.* **2011**)
Necrotizing meningoencephalitis	Non-Mendelian heritance (disease)	Pug	30	68	86,692[D]	(**Barber** *et al.* **2011**)
Progressive retinal atrophy	Multiple genes (disease)	Golden retriever	27	19	14,389[C]	(**Downs** *et al.* **2011**)
Across-breed GWAS						
Chondrodysplasia (short-limbed)	Dominant (breed-defining)	76 breeds	95	702	41,635[B]	(**Parker** *et al.* **2009**)

Table 2 (*Continued*)

Chondrodysplasia	Dominant (breed-defining)	17 breeds (+crossbred dogs)	18 (6 breeds)	27 (11 breeds)	50,000[B]	(**Bannasch** *et al.* **2010**)
White spotting	Semi-dominantly inherited (breed-defining)	28 breeds	31 (11 breeds)	31 (14 breeds)	50,000[B]	(**Bannasch** *et al.* **2010**)
Hyperuricosuria	Simple (disease)	28 breeds	10 (3 breeds)	59 (25 breeds)	50,000[B]	(**Bannasch** *et al.* **2010**)
Brachycephaly	Complex (breed-defining)	24 breeds	20 (10 breeds)	33 (14 breeds)	50,000[B]	(**Bannasch** *et al.* **2010**)
		16 breeds	28 (3 breeds)	120 (13 breeds)	50,000[B]	
55 morphological features	Multiple loci (breed-defining)	80 breeds (915 dogs)	Depending on trait		60,968[B]	(**Boyko** *et al.* **2010**)
Furnishing, drop ear, curly tail, boldness	Multiple-loci (breed-defining)	46 breeds (509 dogs)	Depending on trait		157,393[D]	(**Vaysse** *et al.* **2011**)

[1]The second part of the study consisted of an across-breeds GWAS using 903 dogs from 80 breeds (50K Canine Affymetrix v2, 40,812 SNPs) in order to confirm the associations found for the three phenotypes in each breed.

[2]In a first step, Abitbol *et al.* (**Abitbol** *et al.* **2010**) performed genome-wide association analysis using 247 microsatellite markers genotyped in 39 affected and 38 non-affected as well as linkage analysis with 48 affected dogs genotyped at 315 microsatellites. Dogs were genotyped with the CanineHD Beadchip in order to identify additional informative markers in the region associated in the first step.

Table 3. Selection mapping in dogs. The method used to detect the genetic signature of selective sweeps in the dog genome is presented. When possible, the actual number of SNPs used for analysis after quality control procedures are presented; the canine SNP array is also indicated: [A]27K Canine Affymetrix v1 (~27,000 SNPs), [B]50K Canine Affymetrix v2 (~50,000 SNPs), [C]CanineSNP20 BeadChip panel, Illumina (~22,000 SNPs) and [D]CanineHD BeadChip panel, Illumina (~174,000 SNPs).

Genetic signature	Method	# Dogs	# Breeds	# markers	Reference
Locus-specific population differentiation	Sliding-window pairwise F_{ST} (d_i statistic)	275	10	21,114[C]	(**Akey** *et al.* **2010**)
Locus-specific population differentiation	Single-SNP pairwise F_{ST}	915[1]	80	60,968[B]	(**Boyko** *et al.* **2010**)
Decrease in genetic diversity	Sliding-window relative heterozygosity	53 / 148[2]	24 / 16[2]	50,000[B]	(**Bannasch** *et al.* **2010**)
Decrease in genetic diversity	Sliding-window relative heterozygosity	280[3]	25	50,000[B]	(**Olsson** *et al.* **2011**)
Locus-specific population differentiation Selective sweeps Fixation of long halpotypes	Single-SNP F_{ST} and d_i statistics Sliding-window relative heterozygosity (S_i) XP-EHH statistic	471[4]	30	>174,000[D]	(**Vaysse** *et al.* **2011**)

[1]In addition, 83 wild canids (wolves, jackals and coyotes) and 10 Egyptian shelter dogs were included.

[2]Two datasets were used and the specific numbers of breeds, cases and controls are those presented in **Table 2** for the same study.

[3]Specifically, data from 50 Shar-pei dogs and 230 dogs from other 24 breeds were used in order to search selective sweeps in the Shar-pei breed.

[4]In addition, 15 wolves were included.

Applying SNP arrays

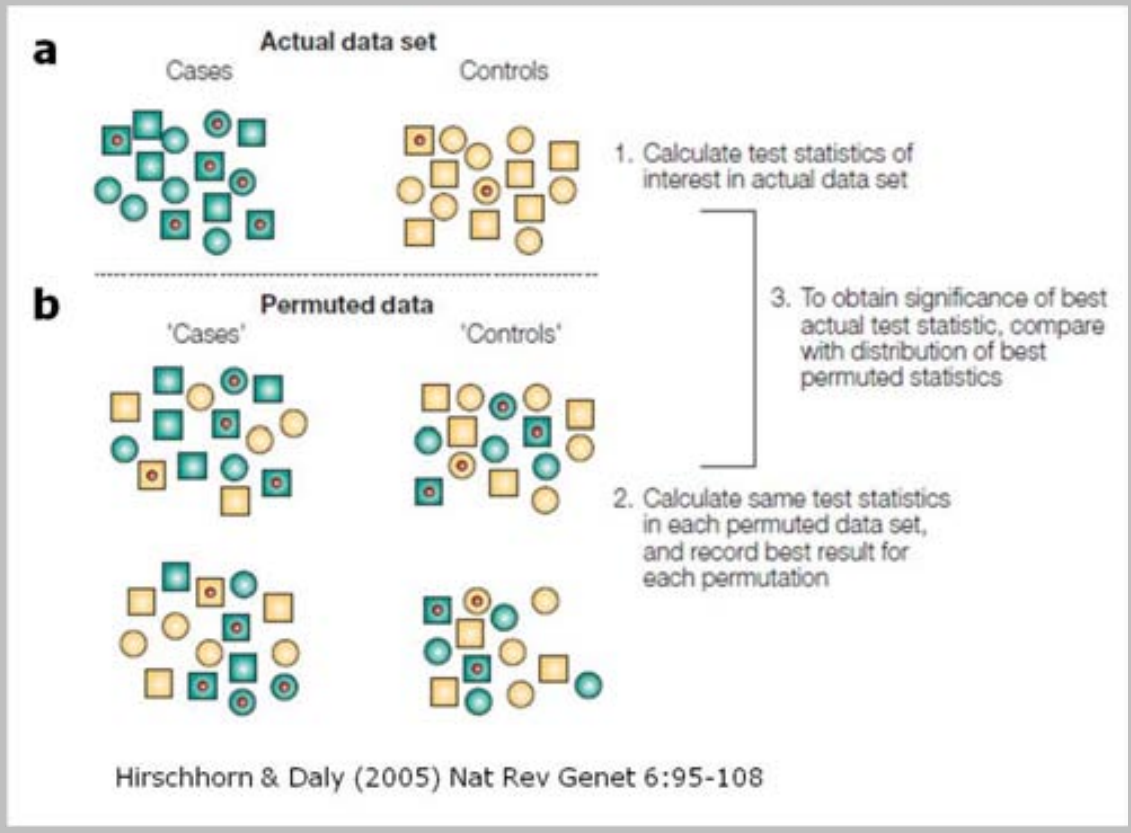
The applications of SNP arrays are numerous (e.g. structural and copy number variation profiling and association, detection of population ancestry, estimation of recombination rates) and they can be currently used in the species for which commercially SNP arrays are nowadays available (e.g. **Table 1**). The aim of this section is, however, to introduce those applications that are generally more popular in canine genetic studies and specifically those applied in the works presented in this thesis.

Genome-wide association studies (GWAS)

In association studies, unrelated individuals from a population with (cases) and without (controls) a trait of interest are compared in order to determine if there are features that differ statistically between both groups. In GWAS the features tested are genetic variants (e.g. SNPs) so that the genomic locations of the trait-causing loci can be mapped. In spite of the high-density of SNP arrays, that the causal variant is actually interrogated is unlikely (direct association) (**Figure 2a**). Instead, it is assumed that at least one of the interrogated SNPs will be in linkage disequilibrium (LD) with the causal variant (indirect association) (**Figure 2b**). This assumption requires two conditions: (i) most cases of the trait are due to relatively few distinct ancestral mutations at the trait-causing locus, and (ii) the marker allele was present on one of the ancestral chromosomes and lies close enough to the trait-causing locus so that the correlation has not yet been eroded by recombination during the population history (**Lander and Schork 1994**). A special consideration inherent to association studies, not only genetic ones, is the presence of confounding factors (e.g. sex, age, weight) beyond the tested feature, which may cause false positive associations. In GWAS a major confounding effect can be stratification of cases and controls (**Box 1**) and may result in a distribution of test statistics deviating from the distribution expected by chance, which is detected by an

Box 1 | **Critical considerations in GWAS**

False positive associations due to **genetic stratification** can be illustrated if we imagine that the different colors of cases and controls (**a**) represent also different, for instance, population ancestry and that this is reflected in different allelic frequencies genome-wide. Then, any potential true association (red circles) will be masked by a myriad of false positives.



The most common **multiple-testing correction** is permutations (**b**). *N* permuted sets (typically 10,000–100,000) in which the labels of cases and controls are randomly re-assigned to samples are generated (i.e. the relationship between phenotype and genotypes is broken). Then the test statistic of choice calculated in the actual dataset is calculated for each permuted dataset. To obtain the empirical p-value for each SNP (often denoted as P_{genome}) its test statistic in the actual set is compared with the distribution of best permuted test statistics.

increased genomic inflation factor or lambda (λ) (**Devlin and Roeder 1999**); note that in the presence of true SNP associations with the trait and no stratification at all, true associations will represent a very small fraction of the markers tested and thus the observed and expected test statistic distributions will largely resemble. Sources of stratification can be differential bias (e.g. differences in sample collection, sample preparation and/or genotyping assay procedures), diverse population or geographical provenance, family structure or cryptic relatedness, which can all co-occur. Another issue, this specific to GWAS, is multiple testing, namely, the fact that the accepted statistical significance levels (e.g. 0.05 or 0.01) applied to single tests are not valid when many SNPs are tested. At these levels, several SNPs will be significantly associated by mere chance. Methods have been developed to correct for stratification (**Price** *et al.* **2010**) and multiple-testing (for example, see **Box 1**), although efforts to avoid stratification should be considered during the collection of samples and genotyping (e.g. collecting unrelated samples from a population as broad as possible and geographically matched, uniform phenotyping criteria in the selection of samples, case-control matching in DNA extraction and genotyping batches).

Almost 1,000 human GWAS have been published (Published GWAS report 2005–6/2011: http://www.genome.gov/gwastudies/) and about a hundred of such studies have been performed in domestic animals. For instance, GWAS have been recently published in cattle for milk production (**Meredith** *et al.* **2012**) and susceptibility to tuberculosis (**Finlay** *et al.* **2012**), in pig for body composition (**Fan** *et al.* **2011**) and *Escherichia coli* susceptibility (**Fu** *et al.* **2012**), morphological malformations in sheep (**Zhao** *et al.* **2011**) and for several phenotypes in the dog (**Table 2**).

In the case of canine GWAS, Lindblad-Toh *et al.* (**Lindblad-Toh** *et al.* **2005**) first suggested a two-stage mapping strategy: an initial genome-wide scan to map associated regions by genotyping tens of thousands SNPs in a single dog breed, and a second stage of fine-mapping of the associated regions using multiple breeds. Simulation-based estimates of the required sample sizes for the initial stage indicated that 10–20 cases and 10–20 controls or 20–50 cases and 20–50 controls would be sufficient to map

genes underlying respectively monogenic recessive and dominant traits; for complex traits, 100–300 cases plus 100–300 controls from one breed would provide adequate power to detect alleles conferring 2- to 5-fold multiplicative risks (**Lindblad-Toh** *et al.* **2005**).

It is worthwhile to note that the within-breed GWAS approach cannot be applied to traits that have been driven to fixation by drift or artificial selection within a dog breed (i.e. there is no phenotypic variation). An alternative is across-breed association studies, in which multiple breeds with and without the trait of interest are used. Although any pair of breeds will have countless genetic differences, with enough breeds only trait-related alleles will consistently differ between cases and controls (**Karlsson and Lindblad-Toh 2008**) (**Table 2**).

Whole-genome marker-enabled prediction (WGP)

WGP methods use LD between markers and trait-causing loci in order to estimate the joint contribution of loci across the genome to the trait. Crucially, essential to these methods is the prediction of genetic values and phenotypes, instead of the identification of specific genes, which has been the central focus of GWAS. WGP methods, which lay on the Quantitative genetic theory (**Falconer and Mackay 1996**), are appropriate to study complex traits because such methods assume that traits may be affected by a large number of small-effect, possibly interacting, genes. Conversely, GWAS assume that few loci with a relatively high effect explain the phenotype.

WGP methods stem from the pioneer work of Meuwissen *et al.* (**Meuwissen** *et al.* **2001**). Where pedigree-based relationships between individuals were used to estimate the heritability of a trait of interest, Meuwissen *et al.* replaced pedigree information with resemblance between individuals inferred from genome-wide SNP data (**Meuwissen** *et al.* **2001**). Specifically, the method proposed to regressing phenotypes on markers genotypes using a linear regression model. An advantage of dense markers maps is that as the number of markers largely exceeds the number of samples, predictions can be accurate even when the estimated effect of

each marker is subjected to large uncertainty (**de los Campos** *et al.* **2010**). On the other hand, an undesired consequence is that the estimation of markers effects is not feasible with ordinary least squares methods. Penalized and Bayesian estimation methods, in which markers effects estimates are shrunk typically to zero, have been developed to overcome this and other problems, and such methods have been reviewed elsewhere (**de los Campos** *et al.* **2010**). For instance, the BayesB method uses the prior knowledge that a high proportion of SNPs (typically denoted as π) have a zero effect and that the effects of the remaining fraction of markers (1–π) will follow the chosen distribution (**Meuwissen** *et al.* **2001**).

If the genetic value of an individual for a trait of interest can be accurately predicted based on genome-wide SNP data, individuals with higher genetic value will be selected for breeding. Because of their initial development in agriculture research, WGP methods have been largely applied with commercial breeding purposes. Indeed, WGP methods are often referred as genomic selection for their initial and most common application to the selection of individuals for traits, typically continuous, of economical interest. For instance, genomic selection has been utilized in cattle for dairy traits, meat quality and other phenotypes of economic interest (**Harris** *et al.* **2009**; **Hayes** *et al.* **2009**; **VanRaden** *et al.* **2009**; **Weigel** *et al.* **2009**; **Garrick 2010**), in chicken for body weight as well as food consumption and mortality rates (**González-Recio** *et al.* **2008**; **Long** *et al.* **2010**), in wheat for grain yield (**de los Campos** *et al.* **2009**; **Crossa** *et al.* **2010**) and in maize for flowering and grain yield phenotypes (**Crossa** *et al.* **2010**). On the other hand, WGP methods have been also used in mice to predict body mass index (BMI) (**de los Campos** *et al.* **2009**). Studies focusing on species of economic interest typically assess the accuracy or reliability of WGP methods as the correlation between the true value of the phenotype and the genomic prediction (commonly called genomic estimates of breeding values or GEBV), and values in the 0.02–0.70 range have been published in the cited studies.

In humans, genetic markers have been used in order to predict children and adult obesity using BMI as proxy (**Willer** *et al.* **2009**; **Zhao** *et al.* **2009**; **Holzapfel** *et al.* **2010**), pigmentation (**Valenzuela** *et al.* **2010**), type 2

diabetes (**van Hoek** *et al.* **2008**) and Alzheimer's disease (**Seshadri** *et al.* **2010**). However, only genome-wide significant SNPs associated in previous GWAS or meta-analyses (3–75 SNPs, depending on the study) were used to build up the predictive models. With the exception of pigmentation, probably the least complex trait amongst them, these studies have shown little success to explain the observed phenotypic variance. Only Purcell *et al.* (**International Schizophrenia Consortium** *et al.* **2009**) allowed more relaxed significance thresholds to include tens of thousands of common variants contributing to the risk to develop schizophrenia. Although the selected variants explained as little as ~3% of the phenotypic variance, simulations consistent with real data suggested that a polygenic basis to schizophrenia explained at least one-third of the total variation liability. Still, in these examples, which markers are included in the model is assessed one marker at a time by imposing significance thresholds, whereas in the WGP methods the effects of all markers are jointly inferred. More recently, Yang *et al.* fitted a linear model to human height data and used restricted maximum likelihood (REML) to estimate the variance explained by 294,831 SNPs genotyped on 3,925 unrelated individuals and with no pre-selection of SNPs based on statistical significance (**Yang** *et al.* **2010**). This approach greatly increased the 5% of variance in height explained by ~50 loci associated in previous GWAS up to 45% and evidenced that the missing heritability (human height heritability is known to be as high as 80%) can be explained by the incomplete LD between the genotyped and causal variants and by the lower minor allele frequency (MAF) of the latter. Later, this approach has been applied to estimate the genetic contribution of common SNPs to disease liability to Crohn's diseases, bipolar disorder, type I diabetes (**Lee** *et al.* **2011**), Parkinson's disease (**Do** *et al.* **2011**), venous thrombosis (**Germain** *et al.* **2011**) and schizophrenia (**Lee** *et al.* **2012**) as well as to estimate the risk to cardiovascular disease (**Simonson** *et al.* **2011**) and the heritability of half a dozen of metabolic traits (e.g. BMI, systolic blood pressure) (**Vattikuti** *et al.* **2012**). To note, population structure can inflate SNP-based heritability estimates (**Browning and Browning 2011**).

In dogs, there are few examples in which pedigree-based heritability estimates of certain disease-related phenotypes have favored the selection against such diseases. These examples include the genetic evaluation of the hip score, related to hip dysplasia, in Labrador retrievers (**Lewis** *et al.* **2010a**, **2010b**) and the heritability of premature mitral valve disease in Cavalier King Charles spaniel (**Lewis** *et al.* **2011**). Likewise, a method for predicting the risk of joint dysplasia, osteoarthritis and secondary related-conditions based on genotyping polymorphisms in the *CHST3* gene has been developed for dogs (**Martinez** *et al.* **2012**).

Selection mapping

Selection mapping consists in the identification of trait loci through the detection of selective sweeps, which are the result of an increase in frequency, near to fixation, of favorable mutations owing to selection for specific traits (**Box 2**). The genetic signatures of selective sweeps include reduction in the genetic diversity in the nearby vicinity, population differentiation (particularly if the sweep has not spread to all populations within the species), increased frequency of derived and rare alleles and increased LD and long-range haplotypes. These signatures and the specific statistical tests that have been proposed for their detection have been excellently reviewed elsewhere (**Nielsen 2005**; **Sabeti** *et al.* **2006**; **Nielsen** *et al.* **2007**). These reviews give numerous examples of detection of selection in the human lineage through genome-wide scans; for instance, two selective sweeps at *LCT* and *G6PD* loci related respectively to lactose tolerance in European populations and resistance to malaria in many African populations.

Whilst in the case of humans positively selected beneficial traits likely include adaptations to new and diverse environments (e.g. bipedalism, speech, resistance to infectious disease) (**Sabeti** *et al.* **2006**), in domestic animals the major driving selective force has been the strong artificial selection imposed by breeding. Moreover, the detection of selective sweeps in these species is influenced by the strength of the selection and their recent evolutionary history. This is because, in general, the size of the

selective sweep will depend on (i) the rate at which the favorable haplotype becomes fixed, which in turn is a product of the strength of the selection and the effective population it acts on, and (ii) the rate at which new haplotypes arise from recombination and new mutations. Then, because of both the strong directional selection that domestic animals have been subjected to and their short evolutionary history, the genomic footprints of major selective sweeps should largely remain (**Andersson and Georges 2004**; **Rubin** *et al.* **2010**; **Elferink** *et al.* **2012**).

In dogs, selection mapping has been proven to be a successful strategy to identify genomic regions that govern traits that are specific to one or a group of breeds (**Table 3**). In this sense, selection mapping is another alternative, together with across-breeds GWAS, for the mapping of traits that have been driven to fixation within a dog breed. Across-breed GWAS, however, may be underpowered in the case of rare selective sweeps corresponding to regions of the genome under selection only in one or a small number of breeds and selection mapping can be then more appropriate (**Vaysse** *et al.* **2011**). Moreover, this approach is attractive because it does not require a prior knowledge of the trait that is being investigated; that is, one can detect a selective sweep in a given breed or group of breeds and then pinpoint the phenotypes that are characteristic in that breed or breeds and hypothesize about the function of the selected loci with regard to the trait. There are at least two caveats to this approach however. First, it assumes that breeds with the same trait have inherited the same causal variants from the ancestral dog population (**Karlsson and Lindblad-Toh 2008**); the same problem applies to across-breeds mapping. If that happens not to be the case, the sweeps will not be shared across breeds and association will not be found in selection mapping and across-breeds GWAS. Second, it may be difficult to determine whether a signature is due to selection or to genetic drift; however, Vaysse *et al.* suggested that extended blocks of homozygosity on the megabase (Mb) scale appear to be best explained by selection in the dog (**Vaysse** *et al.* **2011**).

Box 2 | **Genetic signatures of selection: selective sweeps**

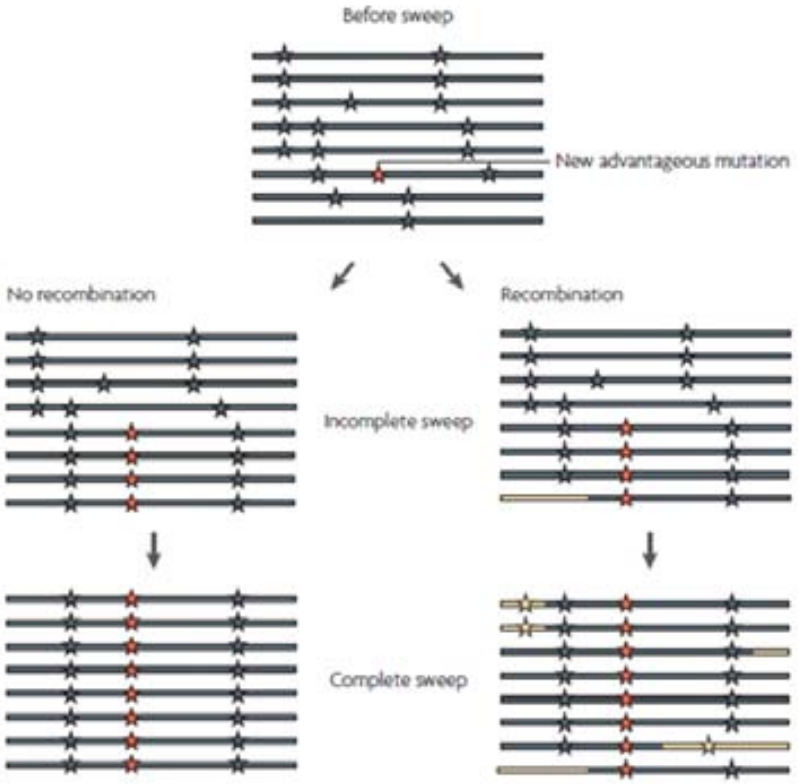
Formation of selection sweeps (a)

A new advantageous mutation (indicated by a red star) appears initially on one of the haplotype (lines). In the absence of recombination, all neutral SNP alleles (gray stars) on the chromosome in which the advantageous mutation first occurs will also reach a frequency of one as the advantageous mutation becomes fixed in the population. Likewise, SNP-alleles that do not occur on this chromosome will be lost, so that all variability has been eliminated in the region in which the selective sweep occurred. However, new haplotypes can emerge through recombination and mutation, allowing some of the neutral mutations that are linked to the advantageous mutation to segregate after a completed selective sweep. Chromosomal segments that are linked to advantageous mutations through recombination during the selective sweep are colored yellow. Data that are sampled during the selective sweep at a time point when the new mutation has not yet reached a frequency of one represent an incomplete selective sweep (text modified from **Nielsen** *et al.* **2007**).

Spatial pattern of selective sweeps (b)

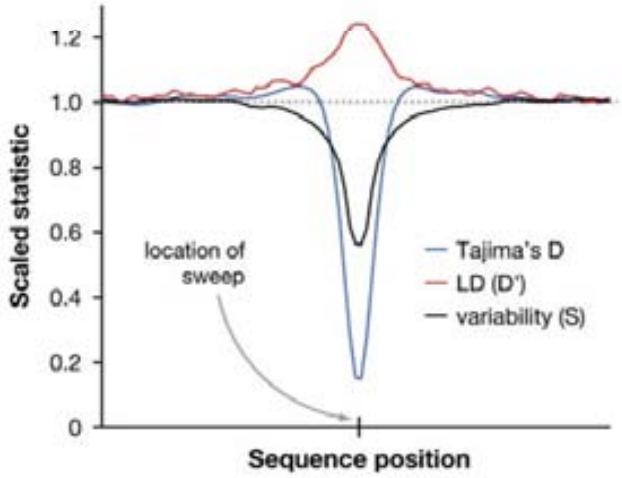
The loss of variability will be strongest in the immediate vicinity of the selected mutations, and will diminish with increasing genetic distance from them because of recombination and new mutations. Likewise, LD will be maximal between the selected and closer variants and will decay as distance away from the advantageous mutation increases (text modified from **Nielsen 2005**).

a



Nielsen et al. (2007) Nat Rev Genet 8(11):857-868

b



Nielsen (2005) Annu Rev Genet 39:197-218

Genetic mapping of complex traits

Dissecting human traits is complicated

Genetic mapping of trait-causing genes to chromosomal locations dates back to the work of Sturtevant in *Drosophila* in 1913 (**Sturtevant 1913**) and its use in order to understand the genetic basis of human disease has been one of its greatest applications. Yet medical genetics was revolutionized during the 1980s by the genetic mapping of human rare diseases having simple Mendelian inheritance, the fact that most traits of medical relevance are complex and do not follow a simple monogenic Mendelian pattern posed the need to develop new approaches to study human complex traits (e.g. linkage analysis, allele-sharing methods and association studies) (**Lander and Schork 1994**). However, there are significant difficulties inherent to human-based studies (e.g. genetic heterogeneity, long generation periods) and these can be overcome by using appropriate model organisms.

Model organisms for complex traits

Since the beginning of the twentieth century, the foremost model for genetic studies in mammals has been the mouse (**Paigen 1995**). Nonetheless, the mouse is not the most suitable model organism to study complex traits: in many cases traits studied in the mouse are not spontaneously occurring but, instead, trait-causing mutations are induced in the laboratory and these solely test the effect of single loci of major effect. On the other hand, it has been advocated to using domestic species as animal models to decipher the genetics of complex traits (**Andersson and Georges 2004**; **Karlsson and Lindblad-Toh 2008**; **Andersson 2009**). Indeed, domestic animals have been used successfully as model organisms to map complex phenotypes; for instance, susceptibility to melanoma in horse (**Pielberg** *et al.* **2008**), bone strength in chicken (**Rubin** *et al.* **2007**) and muscle growth in pig (**Van Laere** *et al.* **2003**).

The domestic dog: leader of the pack

Amongst the domestic animals mentioned above, the dog has numerous advantages that have made it very appealing as a model for mapping complex traits, especially complex diseases. These valuable characteristics of the domestic dog are determined by its particular evolutionary history, shaped by two bottleneck events: the domestication and the creation of breeds (**Box 3**). As a proof of principle, these advantages (**Table 4**) already partly contributed to the early mapping of diseases occurring in humans and dogs with a simple pattern of heritance (**Ostrander** *et al.* **2000**) and are now widely used in canine studies aiming to unravel the genetic basis of complex traits (**Table 2** and **Table 3**).

In first place, there are >360 canine genetic disorders that are shared with humans, constituting the largest catalogue of naturally occurring genetic disorders that are known in non-human species. Of these, >200 disorders have clinical and laboratory abnormalities that closely resemble a specific human genetic disease and in ~40 disorders the same gene product is abnormal in the corresponding human disease (**Ostrander** *et al.* **2000**). Complex diseases homologous to human diseases include cardiovascular, neurological, inflammatory, immune and metabolic diseases as well as numerous cancers.

Another important benefit is the reduced trait complexity in the dog, with few alleles causing the trait in one or few breeds. This is supported by the high susceptibility to specific diseases or presence of certain physical traits in particular breeds, together with its much lower incidence or complete absence in other breeds. Indeed, 46% of genetic diseases reported in dogs are believed to occur predominantly or exclusively in one or a few breeds (**Ostrander** *et al.* **2000**). This scenario, caused by small founding populations, bottlenecks, and popular sire effects, will only occur when the number of risk alleles is small, and they are relatively rare in the overall dog population.

Table 4. Advantages of the dog as animal model.

	Advantage
Disease	• Spontaneously occurring • Spectrum of disease similar to humans • Often similar clinical manifestations • In some cases the same genes are involved
Genome content	• The dog genome is less diverged from the human than the mouse genome • Relatively compact genome, with less overall repeat insertion and segmental duplication than many mammals • Approximately the same number of genes as in humans, most of which are 1:1 orthologues
Haplotype structure	• Within breeds, long LD and haplotype blocks (40–100 times longer than in humans) • Across the whole domesticated dog population, LD is shorter than in humans, facilitating fine-mapping
Genetic predisposition	• High prevalence of particular diseases or traits in one or few breeds • Few trait-causing loci each with strong effect
Medical aspects	• High medical surveillance • Well documented diseases • Same treatment practices as in humans • Seven times shorter lifespan than humans • Exposed to the same westernized environment as humans

Furthermore, the haplotype structure of the dog genome makes very appropriate its use for LD-mapping. Within breeds, extensive LD created by bottlenecks during breeds creation results in long haplotypes blocks (0.5–1 Mb and 3–6 breed-specific haplotypes (**Lindblad-Toh** *et al.* **2005**), implying that fewer markers and individuals are required for mapping compared to human studies. In the whole dog population, haplotype blocks and LD are much shorter than in humans (~10 Kb and 3–5 common ancestral haplotypes), thus facilitating fine-mapping using multiple breeds.

In addition, the level of medical surveillance and care that pet dogs receive is second only to that to which we treat ourselves, which provides well-

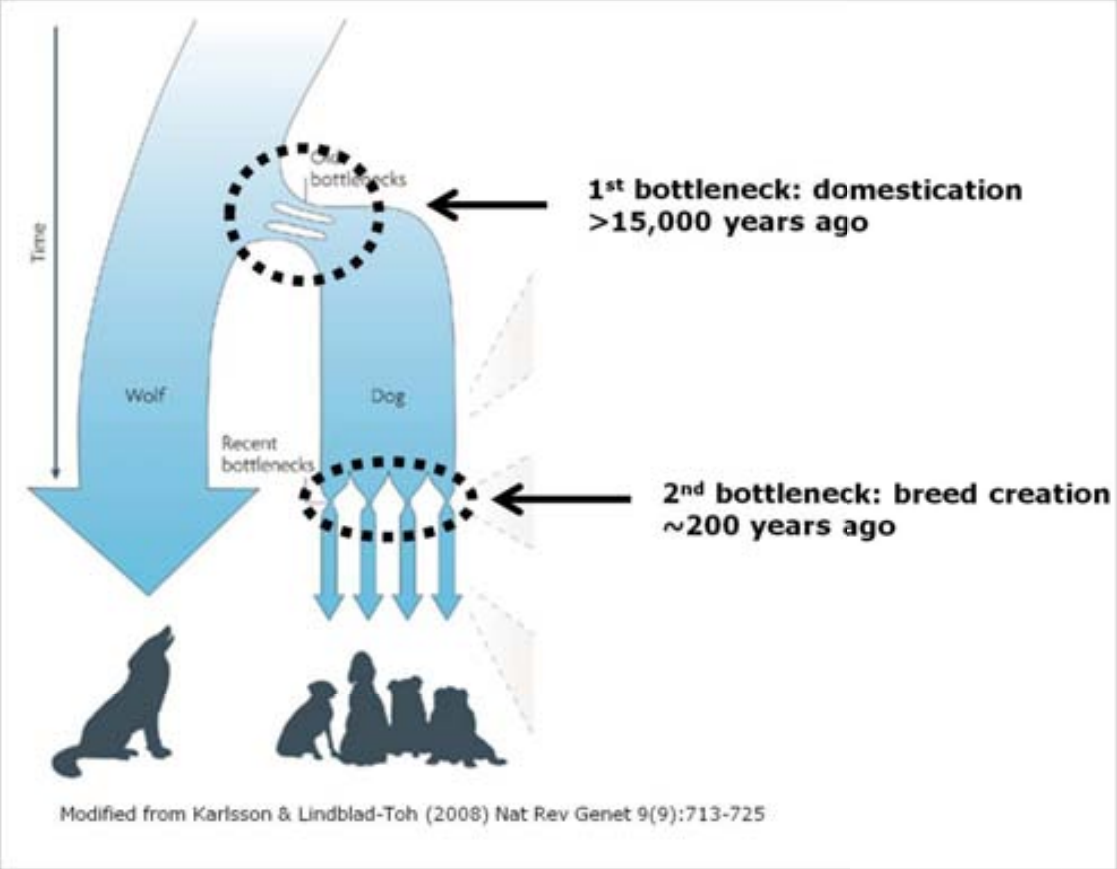
documented medical records, clinical resources and availability of non-experimental clinical samples.

Considering these advantages, it is not a coincidence that the domestic dog has been long proposed as model organism for biomedical research of human diseases (**Patterson** *et al.* **1982**, **1988**; **Ostrander and Giniger 1997**; **Galibert** *et al.* **1998**; **Ostrander** *et al.* **2000**; **Ostrander and Kruglyak 2000**; **Sutter and Ostrander 2004**; **Parker and Ostrander 2005**; **Karlsson and Lindblad-Toh 2008**). Nevertheless, it should not be overlooked that there are certain limitations in the use of the dog as animal model, such as a limited knowledge of the recorded histories of dog breeds, and more importantly, the fact that geneticists somewhat rely on breed clubs and owners for access to phenotypes and DNA samples (**Sutter and Ostrander 2004**). It should also be noted that the extent to which these advantages hold in other domestic animals was out of the scope of this thesis and has not been so extensively explored by the author.

The LUPA Consortium

Probably a landmark example of the use of the dog as animal model has been The LUPA Consortium (http://www.eurolupa.org/) and the research collaborations that have derived from it. Taking advantage of the lower number of cases and controls required for GWAS in dogs compared to humans, together with additional benefits of the dog as animal model (**Table 4**), The LUPA Consortium has carried out GWAS in the dog for the identification of genes relevant to common complex diseases found in humans and dogs. Principally, The LUPA Consortium has focused on neurological, cardiovascular and inflammatory disorders, cancer as well as some monogenic disorders. In addition to GWAS, the availability of large canine SNP datasets generated during the project has enabled other analyses that have resulted, for instance, in a finer picture of the recombination landscape of the dog genome (**Axelsson** *et al.* **2012**) and of the evolutionary relationship between dog breeds as well as the identification of genomic regions under selection (**Vaysse** *et al.* **2011**).

Box 3 | **Evolutionary history of the domestic dog**



Domestication

It has been well established that the dog was domesticated from the gray wolf (*Canis lupus*) (**Wayne 1993**; **Clutton-Brock 1995**) but some specific details have been elusive, such as when and where domestication occurred. Various estimations of the time of domestication have been published in studies using mitochondrial DNA (mtDNA) analyzed in hundreds of samples from wolves and dog breeds from varied geographical locations (~100,000 (**Vilà** *et al.* **1997**); ~15,000 (**Savolainen** *et al.* **2002**); ~16,300 (**Pang** *et al.* **2009**) years ago) as well as in studies comparing nuclear genomic data from wolves, dogs and coyotes (18,000–27,000 years ago (**Lindblad-Toh** *et al.* **2005**). Similarly, diverse most-probable geographic origins have been supported in different studies. Archeological data first indicated an origin of domestication in Europe or Southwest Asia. Conversely, mtDNA and Y-

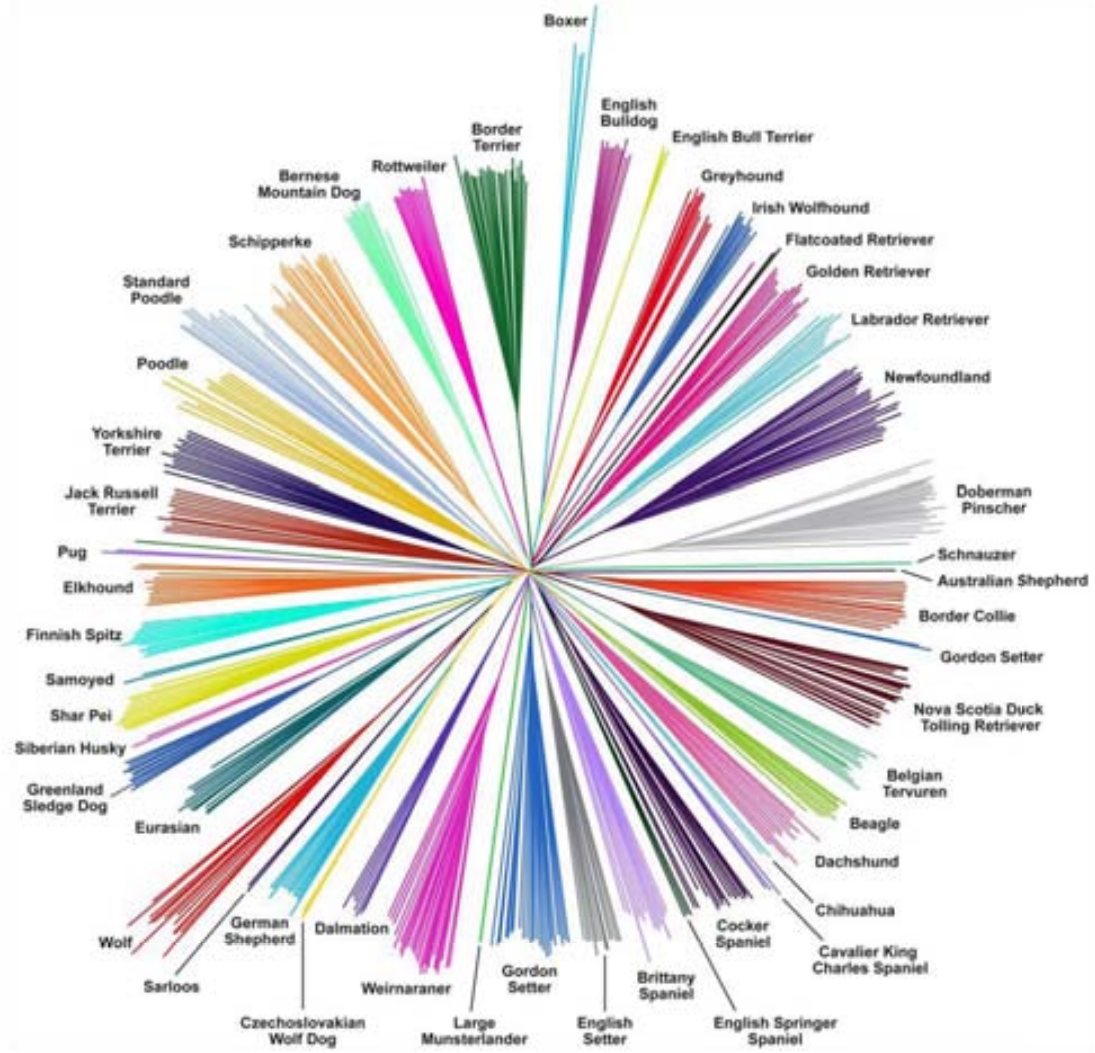
chromosome data agree to place domestication in South East Asia based on higher levels of diversity (**Savolainen** *et al.* **2002**; **Pang** *et al.* **2009**; **Ding** *et al.* **2011**). The fact that similar levels of mtDNA diversity were observed in African and East Asian village dogs (**Boyko** *et al.* **2009**) questioned the hypothesis of a single East Asian domestication event, although this claim has been later shown to be incorrect (**Pang** *et al.* **2009**; **Ding** *et al.* **2011**). Recently, VonHoldt *et al.* (**VonHoldt** *et al.* **2010**), using autosomal SNP data, suggested that domestication took place in Middle East but with minor posterior contributions from local wolf populations.

Breed creation

Breed creation occurred ~200 years ago as a result of a concerted breeding scheme with the objective to breed and select dogs for certain physical appearance (**Alderton and Bailey 2006**). Prior to that time, dogs were bred focusing on their working ability (e.g. guarding, hunting and chasing). This breeding scheme included the creation of breed clubs, or Kennels, that established that registering a dog with a breed club required that both parents of the dog were registered members. In addition, competitions were developed to reward dogs with physical appearance meeting well-documented breed standards (e.g. specific size, shape, color and sometimes behavior). Diversity in some breeds has been further reduced by the presence of popular sires. These dogs have physical features that make them particularly successful in the show ring and hunting or performance events, and as a result, they may produce >100 litters in their lifetime (**Ostrander and Kruglyak 2000**). As a result, each dog breed represents an isolated breeding population with little genetic within-breed variation (**Figure 3**), and with a constellation of traits maintained under strong selection that define each breed (**Ostrander and Kruglyak 2000**; **Ostrander 2005**; **Vaysse** *et al.* **2011**).

Figure 3. Evolutionary relationships between dog breeds. Three main features are obvious: 1) Dogs from the same breed almost invariably cluster together, reflecting the notion that modern breeds are essentially closed gene pools that originated via population bottlenecks. 2) Little structure is obvious in the internal branches that distinguish breeds. This is consistent with the suggestion that all modern dog breeds arose from a common population within a short period of time and that only a very small proportion of genetic variation divides dog breeds into subgroups. 3) The longer boxer branch can be explained by the fact that a large proportion of the SNPs were assayed by comparing boxer with other breeds, which implies that the dataset is enriched for SNPs that differ between boxer and other breeds. The longer wolf branch probably reflects more distant relatedness (text modified from **Vaysse** *et al.* **2011**).

Figure 3 (*Continued*)



Vaysse et al. (2011) PLoS Genet 7(10):e1002316

AIMS OF THIS THESIS

The main objective of this thesis was to use genome-wide SNP data to study specific questions of canine disease and evolution.

The specific aims were to:

(1) Investigate on the host genetic control underlying susceptibility to develop canine leishmaniasis from *Leishmania* infection (Paper I).

(2) Identification of selective sweeps in the genome of the Boxer breed (Paper II).

PRESENT STUDIES (PART 1)

Genome-wide dense SNP data to study host genetic control of canine leishmaniasis (PAPER I)

Background

Leishmaniasis is a vector-borne disease affecting humans and animals, caused by parasitic species of the genera *Leishmania* (protozoa) and transmitted by the bite of phlebotomine sand flies. Human leishmaniasis is an important medical condition both in terms of people at risk (350 millions) and geographic distribution, being endemic in ~100 countries across five continents. Moreover, leishmaniasis is emerging worldwide due to diverse factors (e.g. higher flow of infection carriers, expansion of the geographical area where the vector can live due to the increase of global temperatures).

Some *Leishmania* species cause visceral leishmaniasis (VL), which is the most severe and often fatal clinical form of leishmaniasis (**Kedzierski 2010**). VL has the heaviest disease burden amongst the forms of human leishmaniasis, with an estimated incidence of 0.5 million new cases every year and over 50,000 deaths annually, ranking second in mortality after malaria among parasitic infections (**Bern** *et al.* **2008**). VL can be anthroponotic or zoonotic depending on whether the principal source of infection is human or animal, respectively. Importantly, in many geographic regions VL is caused by *L. infantum*, which finds its principal reservoir host in the dog and causes canine leishmaniasis (CanL) (**Figure 4** and **Figure 5**). Transmission of parasites is thus facilitated by the close presence of dogs in human settlements and adds difficulty to the epidemiological control of VL. Furthermore, the fact that dogs are not mere reservoirs of the parasite but do suffer the disease is a veterinary health problem *per se*. According to the Canine Vector-Borne

Figure 4. Reservoirs of *Leishmania* infection and disease model. *Leishmania* parasites are transmitted by the bites of phlebotomines female sand flies. Man can be the sole source of infection for the vector (anthroponotic) or alternatively wild and domestic animals can act as parasite reservoir hosts (zoonotic). In both cases there are one or few hosts that guarantee the long-term maintenance of the system. In the case of VL caused by *L. infantum*, dogs are the principal animal reservoir for human infection and both species suffer from the disease. The infection process begins with the injection of *Leishmania* parasites into the skin of the host by the vector and with the subsequent infection of, principally, macrophagues. Infection *per se* does not imply the presence of the disease and infected individuals may control or even clear the parasite and remain healthy (red dashed lines). Conversely, some infected individuals will develop clinical leishmaniasis.

Disease (www.cvbd.org), CanL is endemic in more than 70 countries in the world (southern Europe, Africa, Asia, South and Central America and it has recently appeared in the United States). It has been estimated that 2.5 million dogs in France, Italy, Portugal and Spain are infected. The number of infected dogs in South America is also estimated in millions with high infection rates in some areas of Brazil and Venezuela.

A key point in the disease model for leishmaniasis is that only a certain proportion of infected individuals are susceptible to develop clinical disease (**Figure 4**) and epidemiological data suggest that this is partially determined by the genetic background of the host. For instance, in humans there are ethnic differences in developing either subclinical infections or VL (**Ibrahim** *et al.* **1999**) and disease cases frequently cluster in families (**Peacock** *et al.* **2001**). Similarly, in dogs the prevalence of *Leishmania* infection has been estimated to reach ~70% whereas a much lower percentage of dogs (13%) do progress towards clinical disease (**Solano-Gallego** *et al.* **2001**).

Figure 4 (*Continued*)

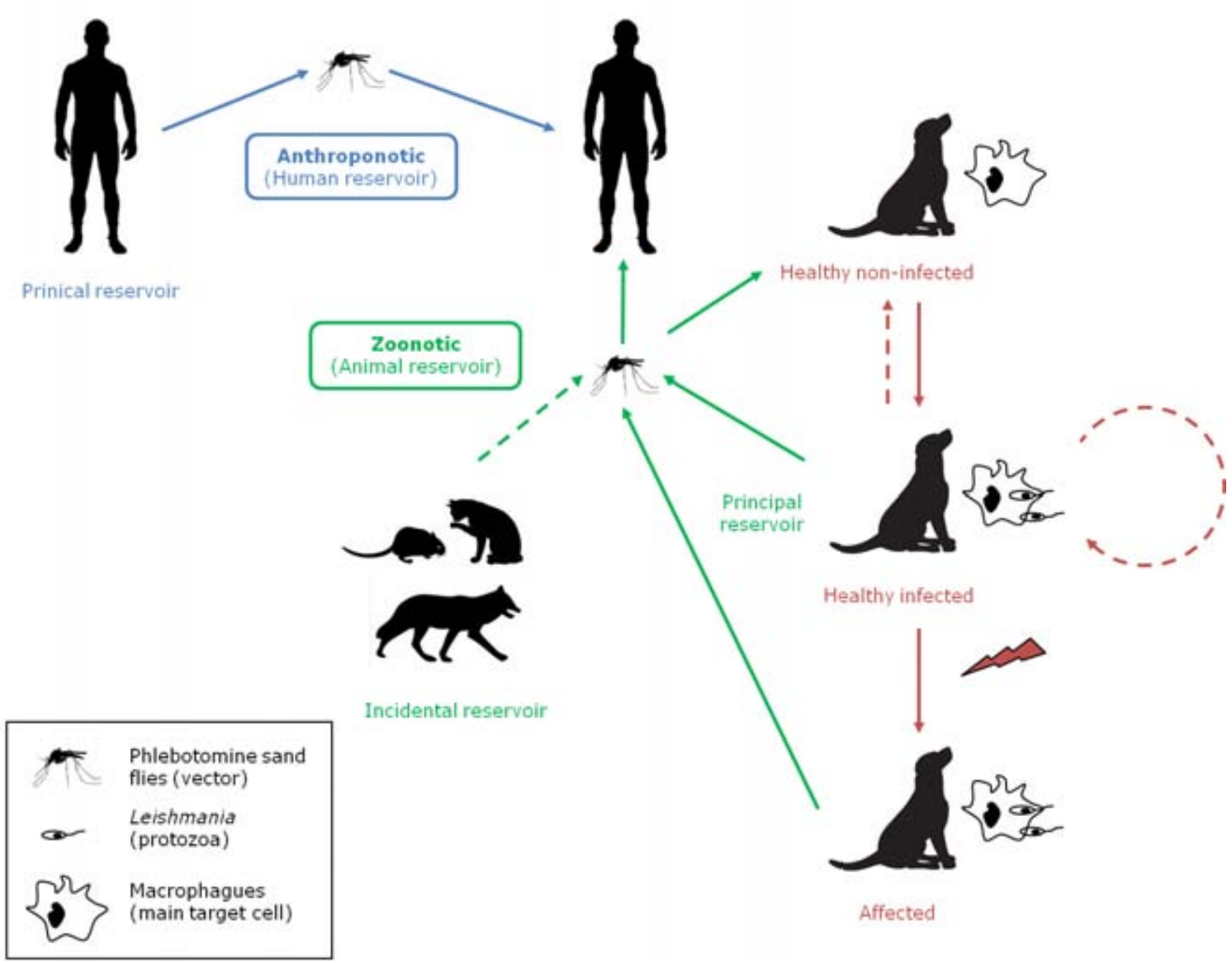


Figure 5. Countries where VL is caused by *L. infantum* with suspected or proven implication of the dog as animal reservoir. In some of the countries colored black, infection with other *Leishmania* species and alternative reservoir hosts can co-occur. Information used to generate this map was extracted from the World Health Organization (**WHO Expert Committee on the Control of the Leishmaniases**).

In addition, CanL has a higher prevalence in certain dog breeds such as Boxer, German shepherd, Doberman and Cocker Spaniel (**Abranches** *et al.* **1991**; **Miranda** *et al.* **2005**; **Sanchez-Robert** *et al.* **2005**), and conversely, it has been suggested that the Ibizan hound breed may be resistant to CanL (**Solano-Gallego** *et al.* **2000**).

Therefore, an important question in the biology of the disease is how host genetics contributes to determining whether individuals progress to clinical disease following *Leishmania* infection. Most previous human (**Meddeb-Garnaoui** *et al.* **2001**; **Karplus** *et al.* **2002**; **Jeronimo** *et al.* **2007a**) and canine (**Altet** *et al.* **2002**; **Quinnell** *et al.* **2003**; **Sanchez-Robert** *et al.* **2008a**, **2008b**) genetic studies have addressed this question using a candidate gene approach targeting at loci like *IFNG*, *TNF*, *SLC11A1* or HLA (DLA in dog). Only, Jeronimo *et al.* (**Jeronimo** *et al.* **2007b**) have published a study in humans focusing on the progression of leishmaniasis following infection using a genome-wide linkage approach based on a few hundred microsatellite markers, and three chromosomes showed linkage with VL or a related disease phenotype. Altogether, no gene has been proposed as a major determinant of the disease progression. Currently, the Wellcome Trust Case Control Consortium is conducting population and family-based GWAS for VL (www.wtccc.org.uk/ccc2/projects/ccc2_vl.shtml).

The aim of Paper I in this thesis was to use the dog as animal model to study leishmaniasis. Cases (affected) and controls (healthy infected) were selected so that the studied phenotype was the progression towards clinical disease from *Leishmania* infection. The numbers of cases and controls (~100 dogs each) and the use of a single breed were decided according to the suggested mapping strategy for canine GWAS (see **INTRODUCTION: Applying SNP arrays**). The Boxer was chosen because this breed appears more predisposed to overt CanL than others. We put especial effort in considering two confounding factors: the stratification of cases and controls and the effect of dogs living indoors/outdoors. Firstly, we performed a GWAS in order to identify loci affecting the studied phenotype and, secondly, a WGP approach was applied with the aim to estimate the genetic variance in the phenotype and to assess the capability of SNP data to predict the trait.

Results and discussion

Brief summary of the genetic models

Throughout Paper I we refer to three genetic models which differ in the confounding factors that were accounted for (**Table 5**). Stratification of cases and controls was present in our dataset as indicated by λ=1.29. A value of λ larger than one indicates that overall the genotypes of cases and controls notably differ and this will be detectable for instance in an *n*- by *n*-individuals matrix of allelic similarities. Statistical techniques such as multidimensional scaling (MDS) are helpful in order to reduce the high-dimensionality of this matrix into a lower-dimensionality matrix that may allow the observation of the differences between cases and controls by plotting the first and most informative MDS dimensions. In addition, fitting MDS dimensions as covariates into the analysis allows correcting for these differences, which are very unlikely to result from true associations with the phenotype genome-wide and reflect instead stratification of cases and controls. Sources of stratification are presented in the **INTRODUCTION** and their presence in our study is evaluated later. Stratification was accounted for by fitting the eigenvalues for the first two MDS dimensions (C1 and C2) as continuous covariates. Only C1 and C2 were used because the fraction of additional genetic variance explained by each of the subsequent MDS dimensions was minimal (see **PAPER I: Figure S2A**). Furthermore, fitting additional MDS dimensions did not further reduce λ (see **PAPER I: Table S2**) or changed the results of the analyses in which an additional MDS dimension was included (see **PAPER I: Table 1**). Lifestyle (defined as whether dogs lived indoors, outdoors or both) was included as a categorical covariate for the models indicated; dogs living outdoors, more exposed to infection, are believed to more frequently develop the disease.

Table 5. Summary of the genetic models and principal findings. The covariates that were fitted into the model in order to correct for the indicated confounding factors are presented (Abbreviations: bp = base pai; h^2 = heritability; s.e. = standard error; CI = confidence interval).

	Model 1	Model 2	Model 3
Confounding factor			
Genetic stratification	__	C1, C2 (continuous)	C1, C2 (continuous)
Lifestyle	__	__	Lifestyle (categorical)
GWAS (strongest associations, P_{raw})			
CFA 1: 39,058,553 bp	1.0×10^{-5}	2.1×10^{-4}	2.7×10^{-4}
CFA 4: 68,238,371 bp	1.1×10^{-5}	2.5×10^{-4}	3.4×10^{-4}
CFA 20: 30,132,329 bp	2.5×10^{-5}	[1]4.4×10^{-4}	[1]6.2×10^{-4}
Genetic variance estimation			
BayesB			
1−π (%)	1.65	1.57	1.54
h^2	0.61	0.63	0.64
GCTA			
h^2 (s.e.)	0.53 (0.18)	0.55 (0.18)	0.59 (0.17)
Accuracy			
Full model			
r	0.18	0.20	0.29
95% CI	(0.05, 0.30)	(0.07, 0.32)	(0.16, 0.41)
Empirical significance	<0.01	0.04	0.03
Permuted genotypes			
r	−0.14	0.02	0.15
95% CI	(−0.27, −0.01)	(−0.11, 0.15)	(0.02, 0.28)
Covariates alone			
r	__	0.11	0.22
95% CI	__	(−0.02, 0.24)	(0.09, 0.35)

[1]For Models 2 and 3 the P_{raw} values correspond to the nearby SNP at 20,126,633 bp.

Sources of stratification and comparison with other canine GWAS

Correcting for stratification by fitting MDS dimensions as covariates only supposed a reduction of λ from 1.29 to 1.17. Roughly half of the dozen published GWAS in dogs provided information with regard to stratification. Three studies (**Mogensen** *et al.* **2011**; **Madsen** *et al.* **2011**; **Tsai** *et al.* **2012**) observed good clustering of cases and controls when plotting the first two MDS dimensions, in spite of different geographical origin of the samples in the study from Madsen *et al.*. Barber *et al.* (**Barber** *et al.* **2011**) also used MDS in order to detect stratification and excluded a good number of outlier samples. Wilbe *et al.* (**Wilbe** *et al.* **2010**) and Downs *et al.* (**Downs** *et al.* **2011**) reported inflation factor values, before correction, of 1.3 and 1.4, respectively. Both studies observed clustering of either samples with similar geographic provenance (**Wilbe** *et al.* **2010**) or known to be related (**Downs** *et al.* **2011**) and performed a Cochran-Mantel-Haenszel stratified analysis within the clusters as a measure of correction. However, no value of the inflation factor after the correction was presented. Only Olsson and colleagues reported an inflation factor of 1.2 after removing two outliers in the MDS plot (**Olsson** *et al.* **2011**).

We consider that it is unlikely that our lambda value was inflated due to population stratification because we neither observed geographical clustering of samples within Spain (the majority of the samples were collected from different areas in the country) nor differentiation of samples collected in other countries (i.e. Italy, Greece and Portugal). It is reasonable to think that geographical stratification would have been noticed if present, as it has happened with some other canine GWAS cited above. Although population or geographical stratification is a common cause of increased inflation factor, there are other confounder effects that can produce the same results (**Price** *et al.* **2010**). We tried to avoid differential bias by following the same procedures in the collection of samples and clustering of samples that went through different DNA extraction protocols or genotyping batches was ruled out. Although we tried to avoid family structure by not including members of the same family, cryptic relatedness might have certainly inflated the lambda value. Nonetheless, we note that lambda values >1.05 are typically considered to denote stratification in human studies (**Price** *et al.* **2010**). Although this is

a statistical rule-of-thumb and it should be the same regardless of the species, we wonder if certain degree of relatedness owing to founder effects, inbreeding, popular sire effects and repeated mating might be inherent to GWAS in dogs in spite of a careful study design.

Genome-wide scan of loci affecting disease progression

Markers in the whole dataset were tested individually for association with the phenotype. Three genomic locations on *Canis familiaris* autosome (CFA) 1, 4 and 20 showed the strongest associations with the phenotype (**Table 5**). Genetic stratification was likely to explain part of each association as P_{raw} values were about an order of magnitude higher when C1 and C2 were fitted into the model; inclusion of dog lifestyle did not have such notable effect on the significance (**Table 5** and **Figure P6a**). None of these associations reached genome-wide significance (see **PAPER I: Figure S1**).

The lack of genome-wide significance at the individual SNP level may indicate that our study was underpowered for GWAS due to small sample size; our dataset consisted of 104 cases and 115 controls and was therefore at the lower end of the range of 100–300 cases and 100–300 controls that has been suggested to map complex diseases through GWAS in dogs (**Lindblad-Toh** *et al.* **2005**). The lack of genome-wide significance may also be evidence of a complex genetic nature of the phenotype, with loci affecting it having a weaker effect than required to pass the stringent significance thresholds imposed in GWAS. Future steps to confirm associations on CFA 1, 4 and 20 would then require replication in an independent sample.

Figure 6. Association on CFA 4. (a) Chromosome-wide Manhattan plot for the three models. The grey line indicates the level of nominal significance that would have reached genome-wide significance in our dataset. On the top, white boxes indicate the non-overlapping SNP sets containing genes that have been previously associated with leishmaniasis in mice (blue lines). Four adjacent SNPs showed the strongest association (black rectangle). **(b)** These four SNPs (black rectangles) lay in two separated haplotype blocks for which haplotypes and their frequencies were estimated using all samples with the Four Gamete rule in Haploview (**Barrett** *et al.* **2005**). An observed frequency >0.003 for the fourth gamete was required so that it was seen in at least one chromosome (1 / 2 x 219 ≈ 0.003). The numbers in the crossing areas indicate the level of recombination between two adjacent blocks. Arbitrary SNP indexes are shown on the top.

Figure 6 (*Continued*)

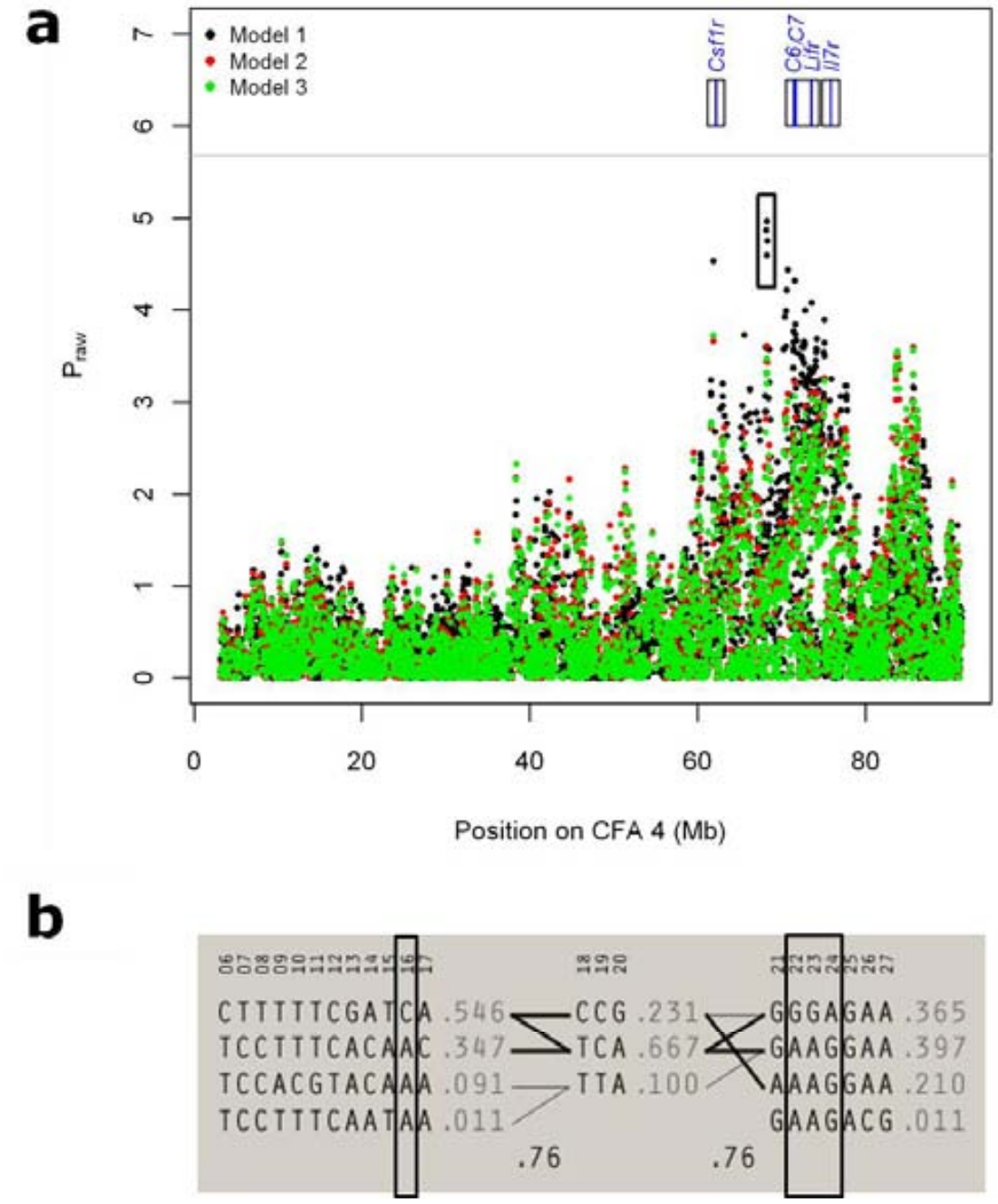


Table 6. Haplotype structure flanking the strongest SNP associations on CFA 1, 4 and 20. The best SNP hits on each association peak (underlined bold) lay in one or two haplotype blocks for which chromosome positions and the number of SNPs that form them are given. For each block, haplotypes were fitted simultaneously into a logistic regression model in order to test for haplotype association with the phenotype and its estimated effect. Covariates according to Model 3 were also fitted. The estimated effect, depending on whether the haplotype is more frequent in affected (risk) or healthy infected (protective), is presented only for significant associations at the <0.05 level after 10,000 permutations (P_{perm}) in order to correct for the multiple haplotypes tested (Abbreviations: OR = odds ratio).

Haplotype	Affected	Healthy infected	OR	P_{perm}	Effect
CFA1_H1: 39,058,553–39,100,429 bp (5 SNPs)					
G A **T** G T	0.029	0.145	0.19	3.8×10^{-3}	Protective
A A **C** G T	0.029	0.018	2.27	0.98	
A G **C** G T	0.644	0.573	1.26	0.98	
A A **C** A C	0.289	0.260	1.16	1	
A A **C** G C	0.010	0.004	3.46	1	
CFA4_H2: 68,058,939– 68,166,961 bp (12 SNPs)					
T C C T T T C A C A A **C**	0.464	0.245	2.22	2.1×10^{-3}	Risk
C T T T T T C G A T C **A**	0.464	0.625	0.59	0.07	
T C C A C G T A C A A **A**	0.072	0.109	0.80	1	
T C C T T T C A A T A **A**	0	0.022	1.0×10^{-9}	1	

Table 6 (*Continued*)

CFA4_H3: 68,225,888–68,356,541 bp (7 SNPs)

G	**G**	**G**	**A**	G	A	A	0.478	0.268	2.20	2.8×10^{-3}	Risk
G	**A**	**A**	**G**	G	A	A	0.353	0.443	0.71	0.70	
A	**A**	**A**	**G**	G	A	A	0.164	0.254	0.67	0.76	
G	**A**	**A**	**G**	A	C	G	0.005	0.018	0.39	1	
G	**A**	**A**	**G**	G	C	G	0	0.018	1.1×10^{-9}	1	

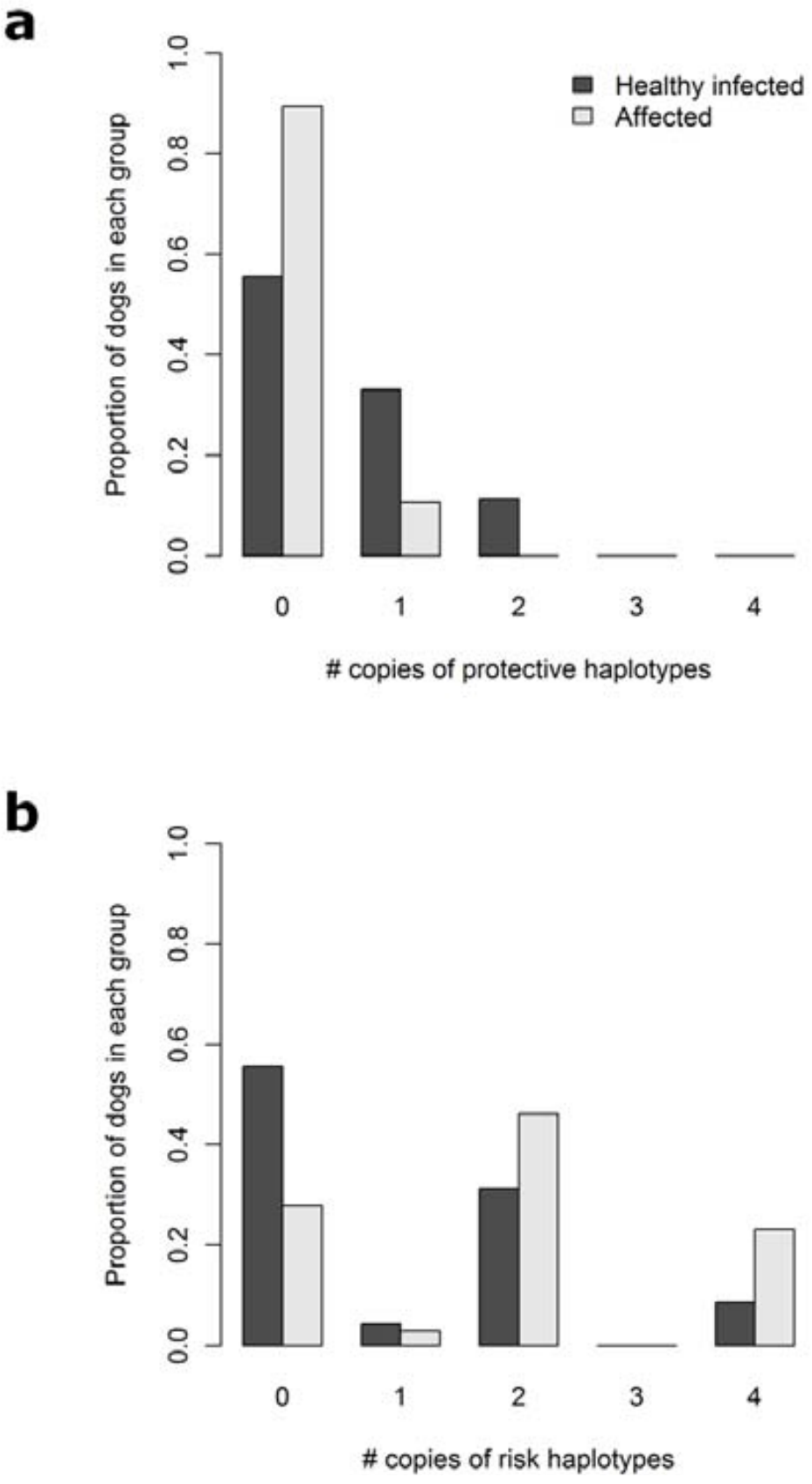
CFA20_H4: 28,638,809–28,684,765 bp (4 SNPs)

T	A	**C**	C				0.024	0.137	0.17	4.0×10^{-3}	Protective
C	G	**T**	T				0.797	0.668	1.79	0.11	
T	G	**C**	C				0	0.007	1.1×10^{-15}	1	
T	A	**T**	C				0.174	0.182	0.93	1	
T	G	**T**	C				0.005	0.006	1.18	1	

CFA20_H5: 28,638,809–28,684,765 bp (4 SNPs)

T	**G**	A	A				0.087	0.2	0.43	0.09	
C	**A**	T	C				0.433	0.344	1.64	0.23	
T	**G**	T	A				0	0.035	8.2×10^{-10}	1	
C	**A**	T	A				0.481	0.422	1.08	1	

Figure 7 (*Figure legend on the next page*)



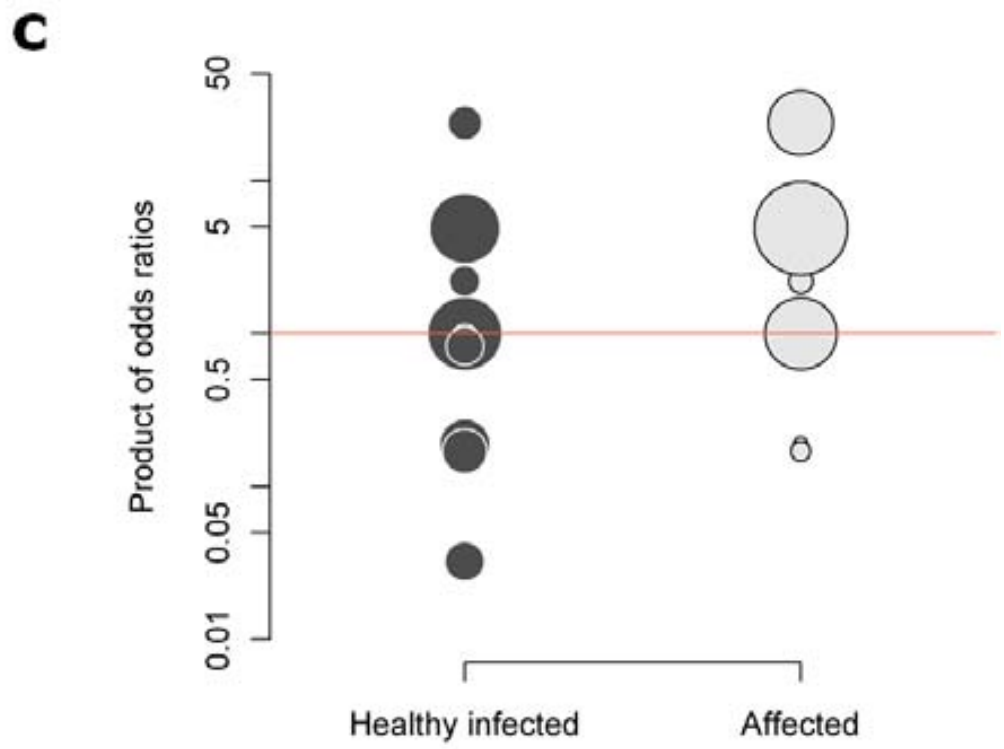


Figure 7. Correlation between protective/risk haplotypes and the phenotype. Distribution of the number of copies of protective **(a)** and risk **(b)** haplotypes in healthy infected and affected individuals. **(c)** Correlation between the affection status and the product of OR. For each individual, this was calculated using the estimated OR if the protective/risk haplotype was carried and one otherwise. Only the four significantly associated haplotypes were used. Due to the limited number of combinations of haplotypes many samples had the same product value and hence the area of the plotting symbols is proportional to the number of samples with a given y-axis value. The red line indicates a product of OR equal to one.

Further analysis of the strongest associations on CFA 1, 4 and 20

Haplotype structure

The best SNP hits within each association peak on CFA 1, 4 and 20 lay on one or two nearby haplotype blocks, as it can be seen for CFA 4 (**Figure 6**). For most blocks one of the inferred haplotypes was significantly associated with the phenotype and had a notable either protective or risk estimated effect (**Table 6**). Specifically, there were two protective haplotypes, one on CFA 1 and the other on CFA 20, and two risk haplotypes both on CFA 4, meaning that each sample can have from zero to four haplotype copies of each class. Healthy infected individuals had 0–2 copies of the protective haplotypes whereas affected individuals had no more than one copy (**Figure 7a**). The larger the number of protective haplotypes the higher the proportion of healthy infected respect to affected individuals (p-value=1.3×10^{-8}, Chi-squared test for trend in proportions). Similarly, the number of copies of the risk haplotypes significantly correlated with the proportion of affected samples (p-value=8.4×10^{-6}). All samples with two copies of risk haplotypes regardless of their affection status (n=84) had a copy from each haplotype block (CFA4_H2 and CFA4_H3) and, with the exception of one sample, these happened to be in the same phase. Thus, CFA4_H2 and CFA4_H3 define a longer-range haplotype, although recombination has happened between each of them and the intermediate haplotype block defined by SNPs 18–20 in **Figure 6b**. This would explain why there were almost no samples with an odd number of risk copies (**Figure 7b**), i.e. most individuals have zero, one or two copies of the long-range haplotype defined by CFA4_H2 and CFA4_H3. Intuitively, the affection status should correlate with the combined product of the number of copies of protective and risk haplotypes and the magnitude of the effect caused by each haplotype, and this correlation was observed in our data (r=0.30, p-value=6.2×10^{-6}, Pearson's product-moment correlation) (**Figure 7c**). The percentage of individuals with product of odds ratio larger than one was notably higher in affected (67%) than in healthy infected (35%) individuals.

Genetic content

Haplotypes CFA1_H1 and CFA4_H3 lay within two genes (**Figure 8**). The region flanking the association on CFA 4 is especially interesting because it is syntenic to a locus in the mouse genome that mediates host response to *L. major* infection (**Figure 6a**). In fact, SNPs nearby dog genes homologous to murine *Csf1r*, *C6*, *C7*, *Lifr* and *Il7r* showed higher levels of association than other candidate loci previously associated with host response to *Leishmania* infection and susceptibility to leishmaniasis in *H. sapiens* and *M. musculus*. For this analysis, SNPs from the whole dataset contained within regions in the dog genome syntenic to candidate human and murine regions were assigned to independent non-overlapping sets, which were tested one at a time for association with the phenotype correcting for multiple-testing and within-SNP-set LD.

In the region on CFA 4 flanking the gene homologous to *Il7r*, three nearby SNPs (CFA 4:75.7–75.9 Mb) significantly deviated from Hardy-Weinberg equilibrium (HWE) in both affected and healthy infected samples (**Figure 9**). Extended patterns of markers deviating from HWE may indicate variation in copy number, and differences in copy number between cases and controls may affect the phenotype. In fact, structural variations have been described for the syntenic region in the human genome (**Korbel** *et al.* **2007**; **Jakobsson** *et al.* **2008**; **Kidd** *et al.* **2008**), which encompasses *SPEF2*, *CAPSL*, *UGT3A1* and *UGT3A2* in addition to *IL7R*. In mice, structural variation has also been reported for a shorter region overlapping *Ugt3a1* (**Cutler** *et al.* **2007**). However, preliminary results indicate that affected and healthy infected dogs do not significantly differ in the number of copies for this region (data not shown); this analysis was performed with Golden Helix Copy Number Analysis software package (Golden Helix, Inc. Bozeman, MT, USA; http://www.goldenhelix.com). Alternatively, the extreme deviation from HWE seen in these three SNPs might result from genotyping errors. Although these SNPs passed the quality control filters, erroneous genotype calling might have occurred. Unfortunately, access to cluster plots of genotypes in order to manually rule out this possibility was not possible.

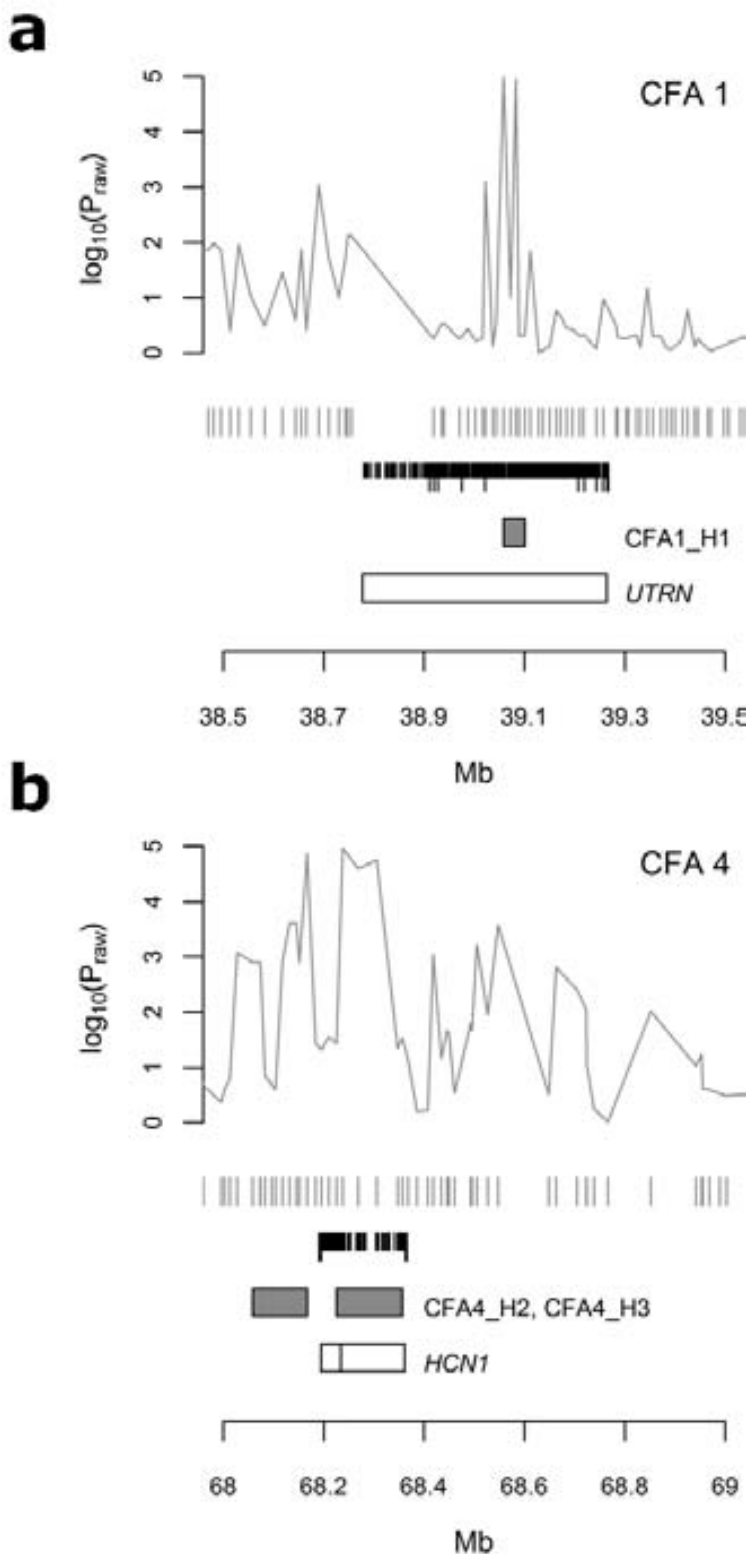
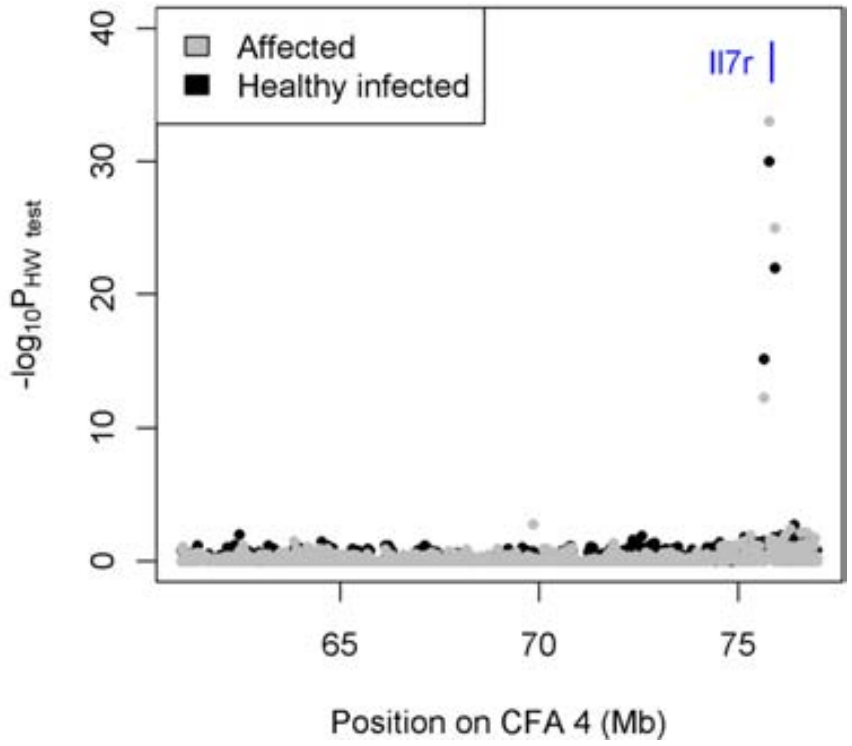


Figure 8. Genetic content in the vicinity of haplotypes CFA1_H1 (a), CFA4_H2 and CFA4_H3 (b). The top panels show the significance in the GWAS (Model 1). The bottom panels show the genes (white boxes) where the haplotypes (gray boxes) are present; additionally, the black track indicates variation in the dog genome (shorter bars correspond to intronic variation) and the gray track indicates SNPs in our dataset. CFA20_H5 was farther than 150 Kb from any gene. A non-coding RNA gene (*U6*) is embedded within *HCN1*.

Figure 9. Three SNPs nearby dog homologous to murine *Il7r* greatly deviate from HWE.

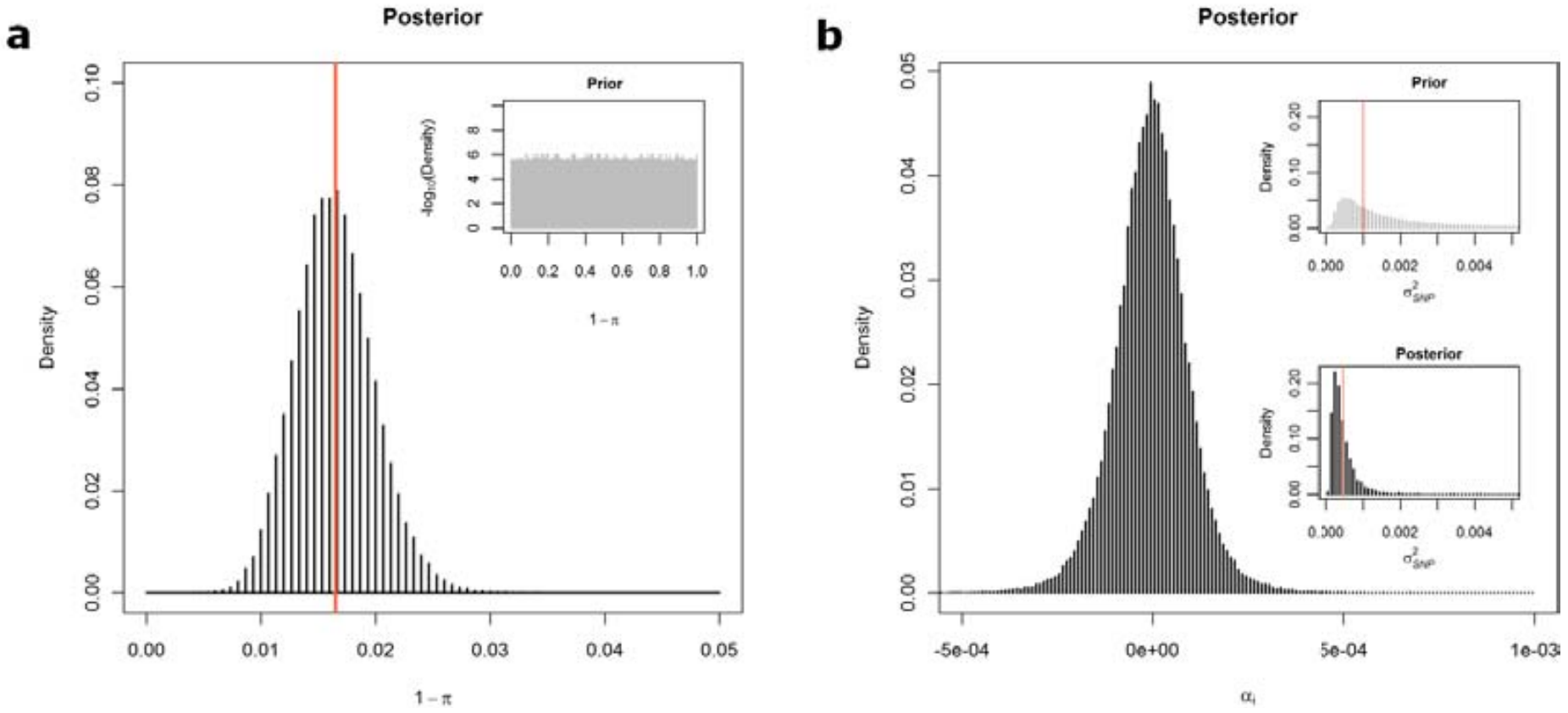


Brief summary of the BayesB method used

As explained in the **INTRODUCTION**, WGP methods estimate simultaneously the effects of all markers on the phenotype and these can be combined in order to accurately predict individual genetic values and phenotypes. The BayesB method originally developed by Meuwissen *et al.* (**Meuwissen** *et al.* **2001**) is one of the WGP methods currently used, in part because in most simulated published data the accuracy of the BayesB method outperformed that of other popular methods (e.g. GBLUP) (**Meuwissen** *et al.* **2001**; **Habier** *et al.* **2007**; **Lund** *et al.* **2009**); although empirical results in cattle have shown similar accuracies for most traits (**Hayes** *et al.* **2009**; **VanRaden** *et al.* **2009**). Specific of the BayesB method is the assumption of a realistic scenario in which only a certain fraction of the markers will have a non-zero effect on the phenotype whilst the remaining fraction will have no effect at all.

Figure 10. Prior and posterior distributions of parameters. (a) a flat prior distribution for the proportion of markers with non-zero additive effect (1–π) was set to follow a beta distribution with shape parameters α=1 and β=1 (inner graph). σ_e^2 was also set to follow a flat prior distribution. The mean of the posterior distribution (~1.6 %) is indicated by the vertical red line. Note the difference in the scale of both axes. **(b)** For those SNPs affecting the trait, the effect is distributed N(0, σ_{SNP}^2), where σ_{SNP}^2 was set to follow an inverse chi-squared distribution with two degrees of freedom (ν=2) and a scale parameter (*S*) of 0.001 (weak informative prior distribution), represented in the top inner graph. The vertical red lines designates *S*.

Figure 10 (*Continued*)



We used the BayesB method (**Meuwissen** *et al.* **2001**) with some modifications previously published (**Pong-Wong and Hadjipavlou 2010**). The model assumed in the method is:

$$\mathbf{y} = \mathbf{Xb} + \sum_{i=1}^{m} \mathbf{g}_i \alpha_i + \mathbf{e}$$

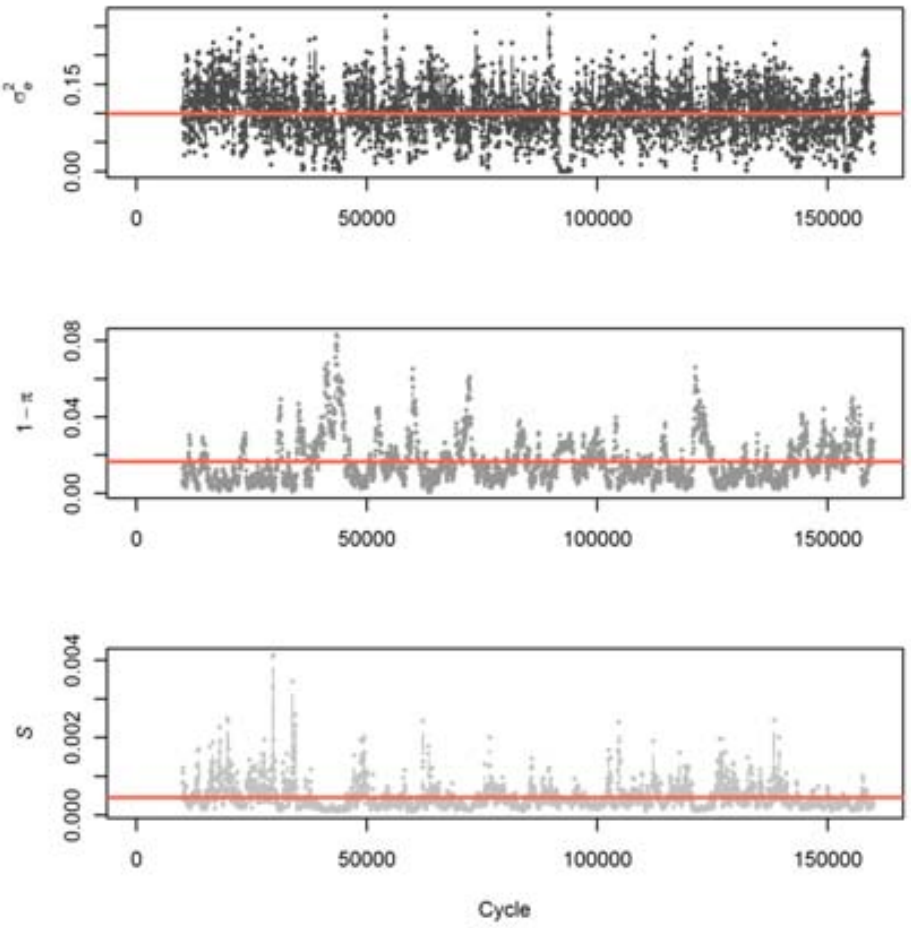
Where **y** is the vector of phenotypes; **b** contains the fixed effects and **X** is its incidence matrix; α_i is the allelic substitution effect for SNP *i*; $\mathbf{g}_i$ is the vector of genotypes for SNP *i*; and **e** the vector of residuals distributed N(0, σ_e^2). The allelic substitution effects α for each SNP are assumed to be from a mixture distribution:

$\alpha_i = 0$ with probability π,

$\alpha_i \sim N(0, \sigma_{SNP}^2)$ with probability (1–π),

And the corresponding prior distributions were chosen for the parameters π, σ_e^2 and σ_{SNP}^2 (**Figure 10**). In the BayesB method, the posterior distributions of the parameters cannot be calculated directly. Instead, Gibbs sampling, a particular Markov chain Monte Carlo (MCMC) algorithm, was used. In essence, the Gibbs sampler produces a sequence of random samples from the target distribution, in this case the posterior distributions of the model parameters. This sequence can be used to approximate each posterior distribution and the posterior mean used as the estimate for each parameter of interest. Inherent to MCMC algorithms, firstly, nearby samples generated by the Markov chain are correlated and, secondly, samples from the beginning of the Markov chain may not represent the target distribution. Therefore, realization of the sampling is done very *n*th value of the chain and the initial samples are discarded as burn-in period (**Figure 11**).

Figure 11. Sequence of samples generated with the MCMC. The MCMC was run for 160,000 cycles, the first 10,000 cycles were discarded as burn in (see that no data is plotted for the initial cycles) and 3,000 realizations of sampling were performed with 50 cycles between realizations; note that data points correspond to realizations of sampling and not to cycles. For each parameter of the model, samples were used to approximate the posterior distribution and the mean (red line) as the estimate of the parameter.



Estimating genetic variance in the phenotype

Two methodologies were adopted in order to estimate the genetic variation relating to the leishmaniasis phenotype: the first the explained BayesB methodology and the second a REML methodology implemented within the

GCTA package (**Yang** *et al.* **2011**). The estimate of heritability was in the range of 0.53–0.64 and was fairly consistent across the different methodologies and models used for its estimation (**Table 5**). Note that these estimates are likely to be biased upwards. For dichotomous traits such as a disease, the observed heritability increases with the prevalence of the trait. But in case-control studies the proportion of affected individuals, aimed to be ~50%, is typically higher than the actual prevalence. Consequently, heritability estimates of binary traits based on case-control data are biased upwards. Importantly, this is the first clear evidence that there is a significant genetic component to leishmaniasis in dogs within breeds. In addition, it is likely the first heritability estimate for progression of clinical leishmaniasis from *Leishmania* infection in any species, although an estimate of heritability for a marker of healed *Leishmania* infection and protection against subsequent reinfection has been reported (**Jeronimo** *et al.* **2007b**).

Using BayesB it was estimated that ~1.6% of markers (1–π) would contribute to the genetic variance (**Table 5**), which suggests that many different genomic segments contribute to the complex phenotype. However, experience with such methods suggests that this figure is sensitive to the distribution of alleles that it is assumed.

Genome-wide plots of the estimated marker effects were similar for Models 1–3 and had a most detectable peak on CFA 4:61–77 Mb (see **PAPER I: Figure S4**). Interestingly, this region overlapped with the association on CFA 4 in the GWAS (**Figure 6**).

Prediction of the phenotype

Finally, we tested the predictive potential of the SNP data. When predicting individual genetic risk to a disease with a polygenic basis, it is important to understand that whereas the outcome is binary (i.e. affected or unaffected) the risk prediction is a continuous measure. This derives from the fact that developing the disease is affected by many loci each with a variable effect. Thus, the combinations of risk genotypes seen in individuals are numerous and only individuals with an aggregated genetic risk beyond a certain threshold will suffer from the disease. Similarly, continuous predictions of the

leishmaniasis disease phenotype were produced in our study with the BayesB method (see **PAPER I: Text S1**). Briefly, the combination of individual SNP genotypes with the respective estimated effects were used to produce predictions for each sample (GEBV or SNP component). Depending on the model, the contribution of confounding factors to the phenotype was added to the prediction.

Cross-validation was used in order to obtain measures of the prediction capability that were not biased by the use of samples in making and testing the predictions. The data was divided so that approximately 4/5 of the samples were used to estimate the SNP effects (training set) and these were used to predict the phenotype in the remaining one fifth (cross-validation testing set). This was repeated five times so that each time totally different individuals constituted the testing set; each time, proportions of affected and healthy infected as well as lifestyle status were kept as in the original full dataset. In this way, we obtained phenotype predictions for each of the samples in the initial dataset (n=219) without each sample having been involved in the estimation of the SNP effects. Otherwise, over-fitting can occur if samples in the training set are also included in the testing set: the predictive model can perform well with this initial data but will perform poorly in an independent collection of subjects.

Accuracy

The correlation between the predictions and the respective known actual phenotypes was calculated as a measure of accuracy (r). The accuracy was between 0.18 and 0.29 and substantial gain was achieved by including lifestyle (**Table 5**, 'Full model'). Still, the key question was whether the genomic data added accuracy and this was assessed in different ways.

The aim of the first was to produce estimates of the SNP effects that were based on random combinations of genotypes and phenotypes. This was done by permuting genotypes with respect to both phenotypes and covariates (the link between phenotypes and covariates was maintained) before running the BayesB method. In general, accuracy values were notably lower with permuted data than with the actual data, regardless of which of Models 1 to 3

were fitted (**Table 5**). Statistical significance was observed only when lifestyle was included (Model 3), which confirms the earlier result that lifestyle supposes a gain in the predictive value.

In the second analysis, the phenotype predictions were decomposed into the component from the covariates and the component from the SNPs. Within each cross-validation testing set, the SNP component was permuted whilst the link between the predictor from covariates and the phenotypes was maintained. One hundred permutations were carried out and the empirical significance of the accuracy estimate was calculated as the fraction of permutations where the p-value of the accuracy was greater than in the true data. **Table 5** shows that the accuracy from the observed data with the true link between phenotypes and genotypes was in the upper tail of the distribution of accuracies ($P<0.05$).

In the third analysis, we analyzed the performance of predictions that only incorporated the component from the covariates. Overall accuracy obtained by C1 and C2 as explanatory factors alone was not significant. Accuracy using only covariates in Model 3 was significant although the accuracy achieved was only half the value obtained from the full Model (**Table 5**). Altogether, these three ways to look at the data proved that prediction of the phenotype was more accurate when genetic markers were included.

Classification into affected or healthy infected

Individuals were classified as either affected or healthy infected for increasing thresholds of the continuous predictions of the phenotype. Values of sensitivity (as the fraction of true affected individuals classified as such) and specificity (as the fraction of true healthy infected individuals classified as such) for different thresholds were used to generate receiver operating characteristic (ROC) curves for each model (**Figure 12**). The area under the ROC curve (AUC) was calculated as indicative of the balance between sensitivity and specificity in classifying individuals: ideally, sensitivity will remain one for increasing specificity and the resulting AUC will be one; in random classification, the AUC will equal to one half. With our data, AUC values were notably higher than randomness and Model 3 achieved the best performance.

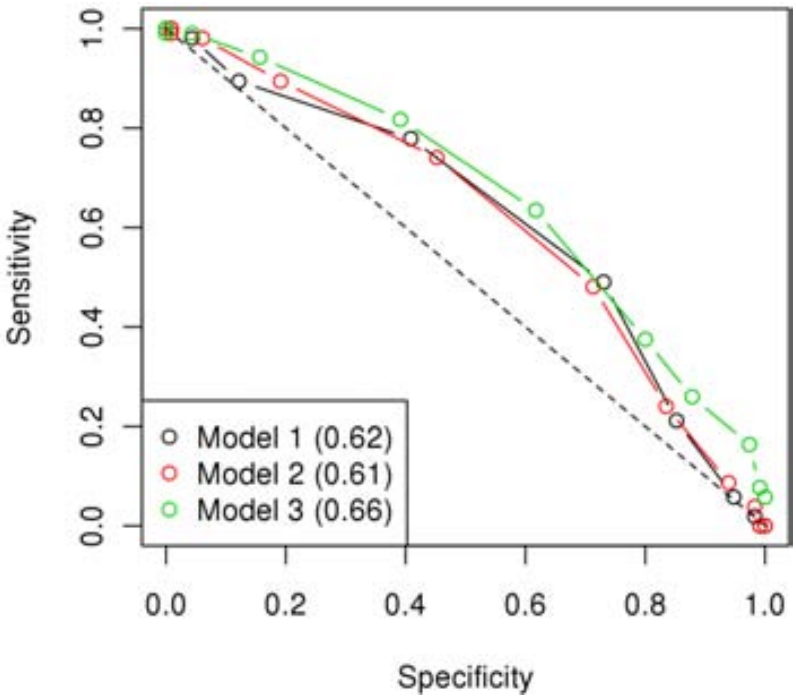


Figure 12. ROC curves. In the legend, the values for the AUC are indicated in parenthesis for each model. AUC can range between 0.5 (randomness, dashed line) and 1.0 (ideally) (figure modified from Paper I).

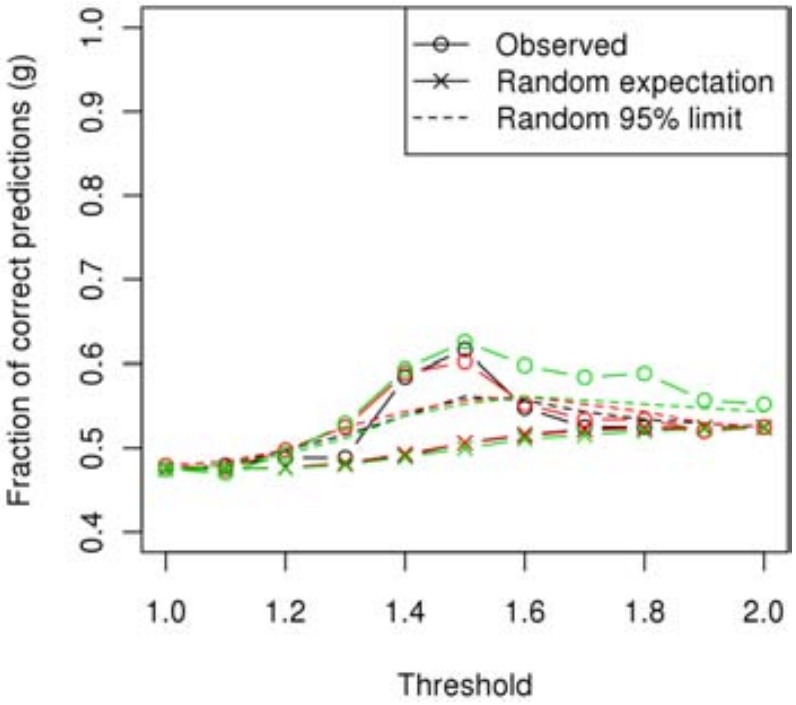


Figure 13. Fraction of correct predictions. For increasing classification thresholds percentages of correct classifications were compared to those expected by chance. Calculations for the random expectation and the random 95% limit were drawn from a hypergeometric distribution and are detailed in **Paper I: Text S1** (figure modified from Paper I).

Regardless of the model, a threshold of 1.5 to diagnose individuals would reach the highest fraction of correct predictions (g), notably higher than the expected by chance alone (for Model 3, $g_{1.5}$=0.63; $g_{95\%limit}$=0.55) (**Figure 13**).

Concluding remarks

We had limited success in identifying loci associated with the phenotype through GWAS: association was found on CFA 1, 4 and 20 but genome-wide significance was not reached. Within the haplotype blocks defined by these associations the presence of certain protective or risk haplotypes correlated with the affection status. Moreover, the association on CFA 4 corresponded to a genomic region which maps to a locus that has been previously associated with host susceptibility to human and murine leishmaniasis, and larger effects for SNPs in this region were estimated with BayesB. Nonetheless, future steps to confirm these associations would require replication in an independent sample. We provided the first heritability estimate for progression of clinical leishmaniasis from *Leishmania* infection in any species. Furthermore, we assessed the capability of genomic information to predict the phenotype, which in the future might be applied in dog breeding and veterinary diagnostics.

PRESENT STUDIES (PART 2)

Identification of selective sweeps in the genome of the Boxer breed (PAPER II)

Background

The history of the domestic dog has been shaped by two chronological population bottlenecks that occurred first at the domestication from wolf and second at the creation of breeds (see **INTRODUCTION: Box 3**). The reduction in the effective population size from the pre-existing dog population into isolated dog populations (i.e. dog breeds) was dramatic, exacerbated by the use of popular sires for breeding and occurred in parallel with strong artificial selection for behavioral and physical characteristics favored by humans. At the phenotypical level, this process resulted in the enrichment of specific traits in single breeds or groups of few breeds; at the genetic level, variation was greatly reduced across the genome.

The same bottlenecks and artificial selection that generated these breed-specific features have, in some instances, provoked undesired health effects (**Clark** *et al.* **2006**; **Karlsson** *et al.* **2007**; **Bannasch** *et al.* **2010**; **Olsson** *et al.* **2011**). For instance, random fixation of detrimental variants can occur during bottlenecks. Similarly, risk alleles may be in LD with selected phenotypic variants or these may have pleiotropic affects (**Patterson** *et al.* **1988**; **Sargan 2004**).

Traits that are fixed in one or few breeds cannot be mapped with GWAS using these breeds alone. Alternatively, these breeds can be conveniently compared with breeds not showing the trait (e.g. across-breeds GWAS, see **INTRODUCTION: Table 2**) or used in the detection of selective sweeps shared

across the breeds with the common phenotype; this second approach, namely selection mapping, has been successfully applied to canine traits such as wrinkled skin, dropped ears or body size (see **INTRODUCTION: Table 3**).

In Paper II of this thesis we searched for selective sweeps in the Boxer breed using high-density genome-wide SNP data. Specifically, genomic regions with extended loss of SNP heterozygosity were sought.

Results and discussion

Detection of Regions of Homozygosity (ROHs) in the Boxer

In order to identify selective sweeps, chromosomes were analyzed in 50-SNPs sliding windows, moving one SNP each time. The average observed heterozygosity of the SNPs in each window ($H_{o\ avg}$) was calculated and windows in the lowest 1% empirical distribution were selected; of these, overlapping and close enough windows were merged as single ROHs. After SNP data cleaning, with this method we searched for ROHs in two Boxer datasets, one with 25 samples genotyped at ~22,000 SNPs and the second with 273 samples genotyped at ~172,000 SNPs (denoted as sets A and B, respectively). Here we focus on the results from set B as it includes a higher number of samples genotyped for a larger number of SNPs; results from set A are used later for comparison. Based on set B, we reported a list of 27 ROHs that might be indicative of selective sweeps in the Boxer (see **PAPER II: Table 2**).

Two studies have provided an extensive catalogue of genome-wide selective sweeps for 30 dog breeds (**Akey** *et al.* **2010**; **Vaysse** *et al.* **2011**) but direct comparison with our results is difficult because the Boxer breed was not used in these works. Boyko *et al.* (**Boyko** *et al.* **2010**) also performed a genome-wide scan for recent selection in an enormous canine dataset but it was based on single-SNP F_{ST} values and few regions are available in the publication.

Table 7. Comparison of Paper II with other surveys of selective sweeps in the dog.

	(Akey *et al.* 2010)	Paper II	(Vaysse *et al.* 2011)
Methods			
Breeds	10 (no Boxer)	1 (Boxer)	30 (no Boxer)
Dogs / breed	21–44	273	10–26
Statistic	d_i	$H_{o\ avg}$	S_i, d_i[1]
SNPs	21,114	171,772	172,115
Window size (Kb)	1,000	~650[2]	150[3]
SNPs in window	9.5 (average)	50	10 (average)
Windows excluded	<4 SNPs	—	<5 SNPs
# windows	1,933	169,812	NA
Windows selection criteria	Extreme 1% tail of d_i distribution	Extreme 1% tail of $H_{o\ avg}$ distribution	Extreme 1% tail of S_i or d_i distributions[4]
Results			
Sweeps / breed	20	27	6–39
Average sweep size in Kb	1,000 —	1,520 (564–8,906[5])	475 (266–3,138)
Genome coverage (%)	NA	1.7[6]	0.12–0.75

[1]A third statistic, XP-EHH (**Sabeti** *et al.* **2007**), was used to validate results found with S_i, d_i.

[2]The average SNP density of the array was 13 Kb/SNP so that this window size was roughly equivalent to 650 Kb.

[3]Windows overlapped by 25 Kb.

[4]In addition, for each region in each breed, a marginal p-value was computed as the proportion of simulated regions that were longer, and the p-values were corrected for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR) method (**Benjamini and Hochberg 1995**). None of the regions identified by the d_i statistic remained significant after FDR correction and the results reported in this table are based on the S_i only.

[5]1,236 Kb (564–3,069 Kb) when the longest region (8,906 Kb) is not considered.

[6]1.3% when the longest region (8,906 Kb) is not considered.

Comparing our results with those obtained for the ten breeds analyzed by Akey and colleagues (**Akey** *et al.* **2010**) is inconclusive because sweeps were defined so that for each breed twenty regions were retained and all of them were of size 1 Mb (**Table 7**); in addition, the marker density was much lower than that in the set B of our study.

The numbers of putative sweeps we found in the Boxer fall in the range of what was found by Vaysse *et al.* for 30 other breeds. On the other hand, the average sweep size and the proportion of the genome covered by the Boxer regions more than doubles the value found in any breed in the compared study, even when not considering the largest sweep of >8 Mb in the Boxer (**Table 7**). It has been estimated, however, that the fraction of long-range homozygous regions is not exceptional in the Boxer compared with other breeds (based on the analysis of ~6% of the genome of 224 dogs from 34 breeds) (**Lindblad-Toh** *et al.* **2005**). There are at least two possible explanations for this difference between our study and that of Vaysse *et al*. First, Vaysse *et al.* used a shorter window size and thus obtained a finer resolution of the regions with reduced heterozygosity. It is likely that the actual length of our regions is shorter and/or that they break into smaller ones. This can be effectively seen in many of the regions whose chromosome locations overlap in both studies, provided that they indicate the same sweeps in different breeds (**Figure 14**). Second, Vaysse *et al.* did test whether the observed regions of reduced variability in each breed were more likely explained by selection or by genetic drift. To do so, they generated breed-specific simulated datasets accounting for both the respective breed creation bottlenecks suffered and variable dog population recombination rates along the genome; in addition, significance values were corrected for multiple-testing using the Benjamini-Hochberg FDR method (**Benjamini and Hochberg 1995**). In this way, they retained 524 high confidence putative sweeps that were more likely to be caused by selection in the 30 breeds studied. This procedure was not applied to our dataset and it might be that some of the regions in the Boxer are better explained by genetic drift.

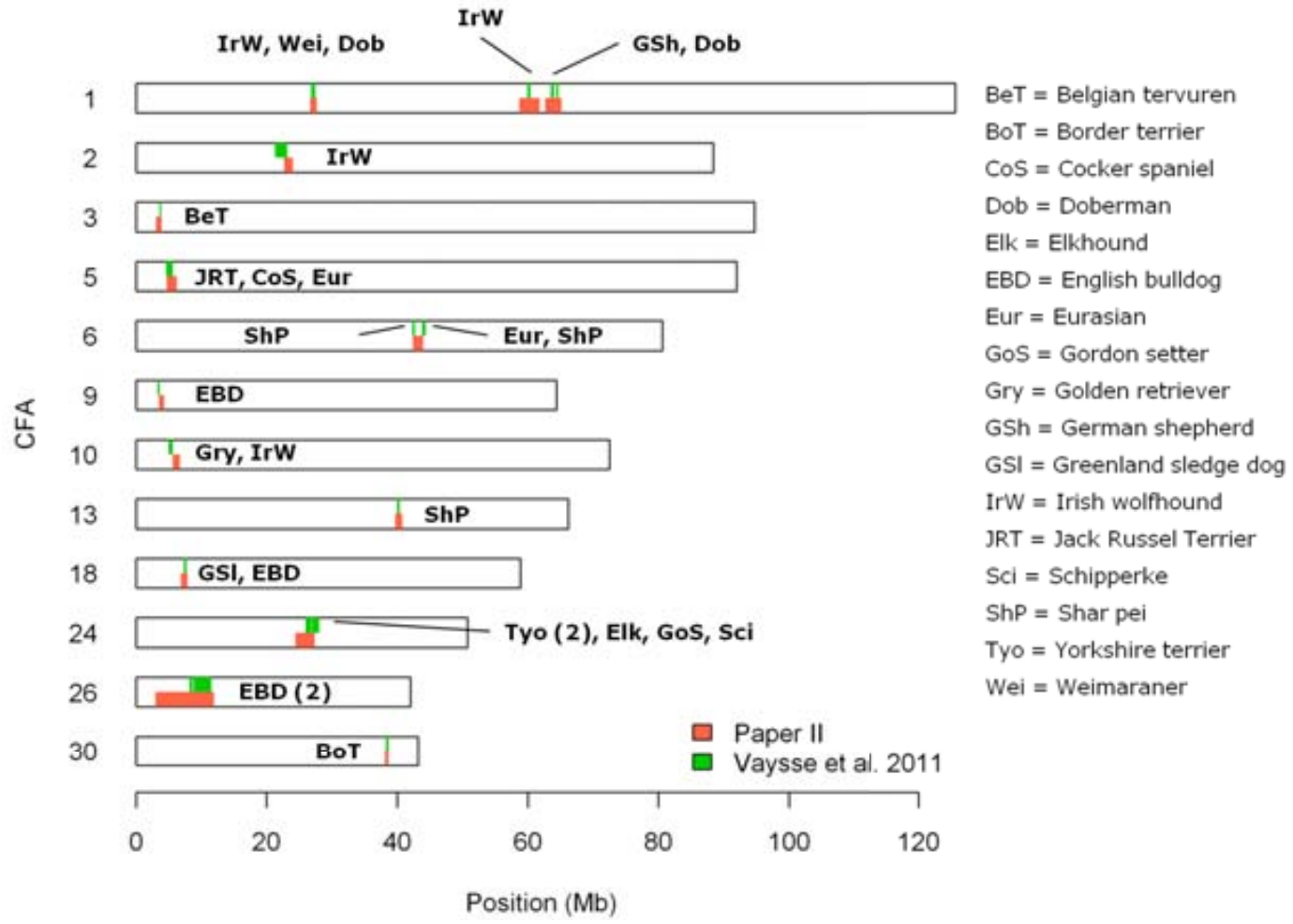


Figure 14. Map of overlapping regions between Paper II (Boxer, set B) and other breeds (**Vaysse** *et al.* **2011**)**.** Note that only chromosomes and regions where overlapping between both studies exists are displayed. The number in parenthesis indicates two regions very close in the same breed.

Nonetheless, Vaysse and colleagues concluded that extended blocks of homozygosity on the Mb scale appear to be best explained by selection (**Vaysse** *et al.* **2011**). It is therefore plausible that a good number of our Boxer regions are indeed caused by selection (>10 regions exceed 1 Mb and had an average heterozygosity <5%).

CFA 26 and brachycephaly: guilty by association?

Amongst the ROHs in the Boxer found in our study, the region on CFA 1:58,710,420–61,801,815 bp (as defined in set B, but also present in set A) matches a previously published location associated with canine brachycephaly (CFA 1:59 Mb (**Bannasch** *et al.* **2010**)). Indeed, for the region associated in Bannasch *et al.*, we observed reduced levels of genetic variation and allelic-matching with the ROH in the Boxer in other brachycephalic breeds such as English bulldog, Pug and French bulldog. Yet, in the latter the reduction of variation was not as extended as in the other two breeds and seemed to be located slightly upstream in the chromosome (see **PAPER II: Figure 3**). On the other hand, in the analysis performed by Vaysse *et al.*, segments covering this part of CFA 1 also shown extremely elevated S_i and d_i values in the English bulldog (region numbers 3,092 and 1,220, respectively; (**Vaysse** *et al.* **2011**)) but were less likely to be explained by selection than by genetic drift.

Brachycephaly is a morphological trait characterized by severe shortening of the muzzle and it is present in a substantial proportion of breeds including the Boxer (**Figure 15**). The phenotype was originally selected in dogs that were used for fighting, based on the idea that this head shape was more powerful for biting (**Alderton and Bailey 2006**; **Ellis** *et al.* **2009**), but with the illegalization of dog fights at least one hundred years ago the brachycephalic head type has been favored due to its similarity to that of human infants (**Nöller** *et al.* **2002**). In contrast, brachycephaly is associated with certain medical problems including breathing abnormalities, cleft palate and lip, increased risk of glioma in some breeds (**Hayes** *et al.* **1975**; **Foley** *et al.* **1979**; **Nelson and Couto 2003**) and a frequent need for Caesarean sections during birth.

The largest ROH in the Boxer found in our study was on CFA 26:3,008,718–11,914,284 bp. The presence of a region of >8 Mb on CFA 26 with almost total loss of genetic variation unavoidably posed this location as a candidate to be scrutinized for its association with brachycephaly. Interestingly, it results that it is on CFA 26 where it is located the second strongest genome-wide significant association for canine brachycephaly reported by Bannasch and colleagues (**Bannasch** *et al.* **2010**), who used 3–10 brachycephalic breeds including the Boxer (**Table 8**). The authors did not discard that, although the significance of the second strongest association was several orders of magnitude lower than that of CFA 1:59 Mb, the associations across many different chromosomes were contributing to the genetically complex nature of the brachycephalic head phenotype (**Bannasch** *et al.* **2010**). Unfortunately, no clue about the exact position of the association on CFA 26 and the levels of variation in the nearby region were published. We reason that there are evidences that suggest that this region represents an additional footprint of artificial selection for brachycephaly.

Presence in Boxers from different geographic locations

Because brachycephaly is a breed-defining standard of the Boxer, if the region on CFA 26 is associated with such morphology the sweep should be present at near fixation in the breed. It was nearly fixed in 273 Boxers (**Figure 16a**) sampled from different areas in Spain, with major representation of the east coast of the country, and with a small fraction of samples (n=6) coming from Italy, Greece and Portugal (data not shown in Paper II). We confirmed that it was also present in an independent cohort of Boxers (n=56) genotyped with the same SNP array and with the same geographic representation (data not shown in Paper II). Moreover, the sweep was present with virtually the same length in set A, which contained Boxer samples from the United Kingdom genotyped on arrays of lower density. It is therefore sensible to think that this region is representative at least of the European Boxer population. Nonetheless, it is worthwhile to consider that some studies have reported genome-wide sub-structure and different haplotype patterns in European and United States populations within the same breed (**Karlsson** *et al.* **2007**; **Quignon** *et al.* **2007**).

Figure 15. The brachycephalic head. Dramatic morphological differences are noticeable between the wild-type head shape present in wild canids (**b**) and several modern dog breeds (non-brachycephalic) (**c**, **d**) and the brachycephalic head present in breeds where it has been selected (**f–i**). The brachycephalic head implies a severe shortening of the bones underlying the muzzle and a more modest shortening and widening of the skull (**a**, **e**) (**Stockard** *et al.* **1941**) (Photography legend: **a** = skull from a German shepherd; **b** = wolf; **c** = German shepherd; **d** = Golden retriever; **e** = skull from a Boxer; **f** = Boxer; **g** = French bulldog; **h** = English bulldog; **i** = Pug) [Photography credits: **a**, **c**, **e**, **f** (**Bannasch** *et al.* **2010**); **b** (http://freewallpapers4desktop.com/); **d**, **h** (courtesy of Verónica Martínez); **i** (courtesy of Violeta Llorens)].

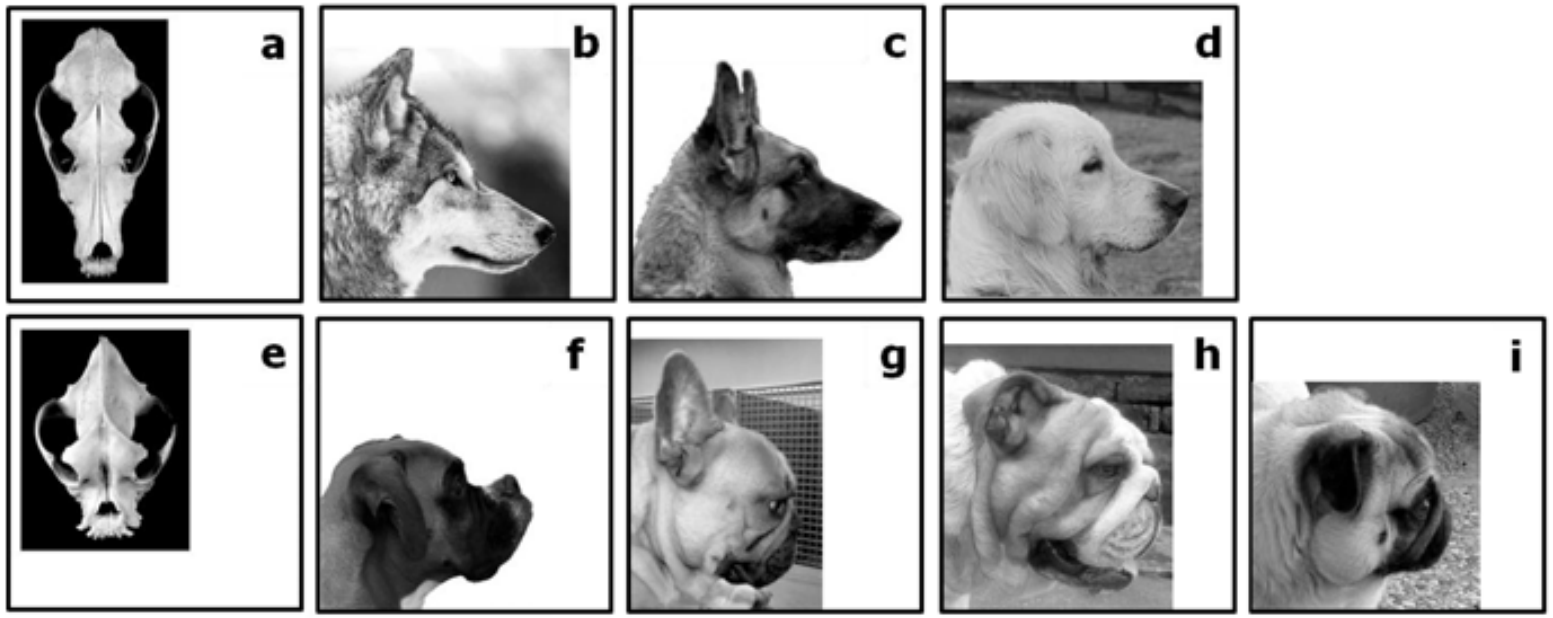


Table 8. Summary of samples and findings on CFA 26 by different authors. Bannasch *et al.* grouped dogs into brachycephalic and non-brachycephalic breeds pools for their analyses, and the representation of each breed relative to the brachycephalic pool is shown in parenthesis.

	(Bannasch *et al.* 2010)		Paper II	(Vaysse *et al.* 2011)
	Dataset 1[1]	**Dataset 2**[2]		
# samples				
Boxer	3 (0.15)	10 (0.36)	273[3]	__
English bulldog	3 (0.15)	__	4	13
French bulldog	3 (0.15)	__	6	__
Pug	3 (0.15)	10 (0.36)	10	__
Evidence on CFA 26				
	2nd strongest association with brachycephaly	NA	Sweeps in: Boxer (3.0–11.9 Mb) English B. (6.8–11.3 Mb) French B. (8.6–9.2 Mb)	Sweeps in English B.: 8.2–8.7 Mb 8.8–11.6 Mb 12.1–13.4 Mb

[1]20 dogs (10 brachycephalic breeds) and 33 dogs (14 non-brachycephalic breeds).

[2]28 dogs (3 brachycephalic breeds) and 120 dogs (13 non-brachycephalic breeds).

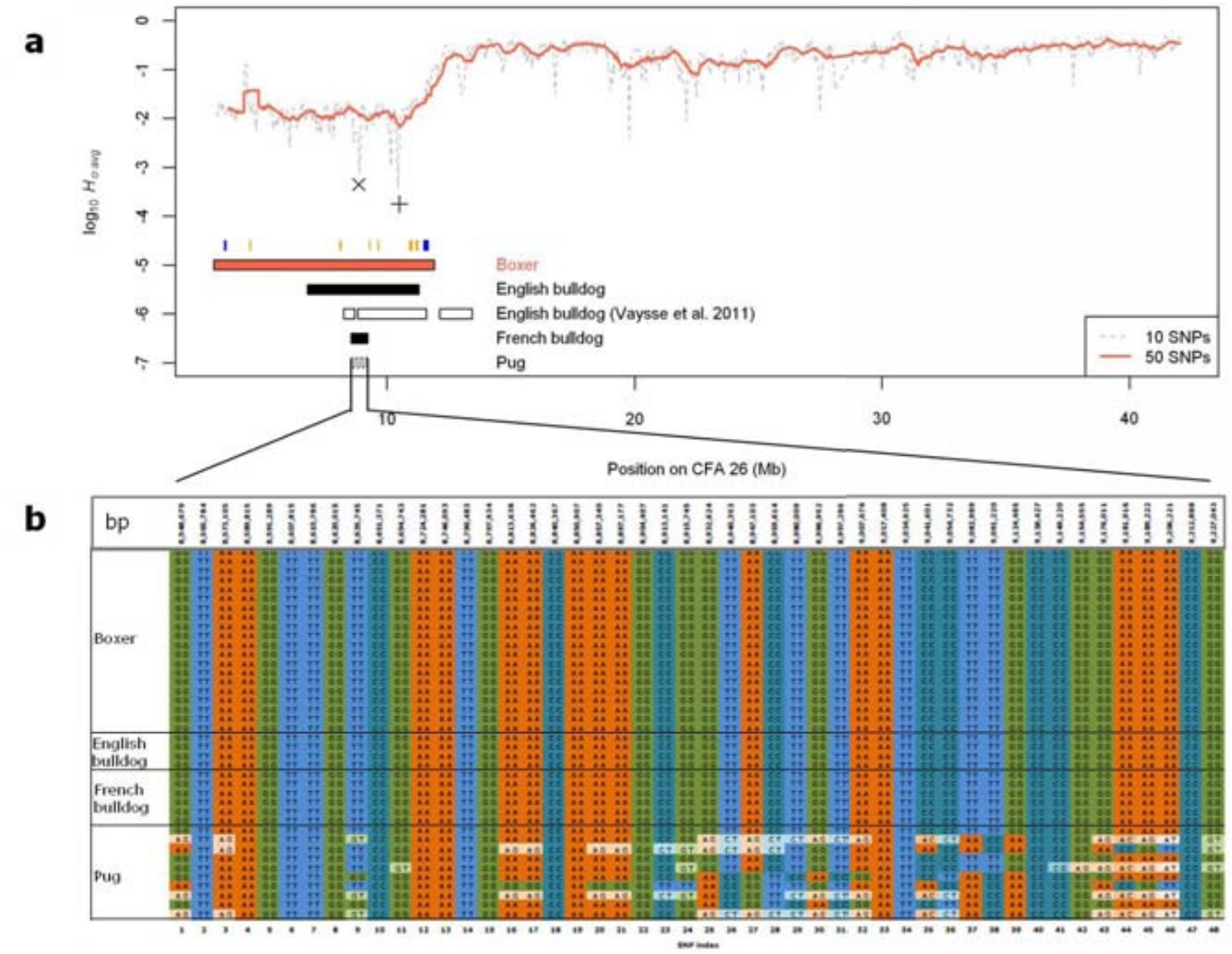
[3]25 Boxers genotyped on the array of lower density are not considered.

Figure 16. Selective sweep on CFA 26. (a) The y-axis displays the logarithm of the averaged observed SNP heterozygosities in windows of 10 (dashed gray) and 50 (solid red) SNPs for the Boxer set B. A clear loss of heterozygosity is seen for the region 3.0–11.9 Mb (red bar), with the greatest reduction (based on 10-SNP windows) on CFA 26:8.9 Mb (×) and CFA 26:10.4 Mb (+). The sweep is partially seen in our data for English bulldog (CFA 26:6.8–11.3 Mb) and French bulldog (CFA 26:8.6–9.2 Mb) (black bars). The same sweep was seen for the English bulldog by Vaysse *et al*. (white bars) (**Vaysse** *et al.* **2011**) (selective sweeps were not searched in Boxer, French bulldog and Pug breeds in that study). Genes that are associated with skeletal and muscular system development and function and tissue morphology categories (orange) as well as disease-related *POLE* and *MYLE* genes (blue) are represented **(b)** In the region of the sweep where the same haplotype is fixed in Boxer and English and French bulldogs (CFA 26:8.6–9.2 Mb), a good number of SNPs in the Pug are fixed for the same alleles. For the Boxer, a random sample of 20 dogs is displayed but fixation was observed in the whole Boxer set (figure modified from Paper II).

Presence in other breeds

Because brachycephaly is present in various breeds, it would be expected that the sweep on CFA 26 is present in brachycephalic breeds whereas variable levels of genetic diversity are seen in non-brachycephalic breeds. We explored further the presence of the sweep on CFA 26:3.0–11.9 found in the Boxer in other breeds according to their brachycephalic status. Part of the sweep was fixed and showed allelic-matching with the Boxer in the English and French bulldogs samples but not in the Pug samples (**Table 8**, **Figure 16**). The haplotype block formed by a single fixed haplotype in the first three brachycephalic breeds (**Figure 16b**) was broken into three shorter haplotype blocks in the Pug, in which the haplotypes that showed fixation in the other three brachycephalic breeds were seen at moderate frequencies (0.25–0.60) (**Figure 17**). Importantly, the sweep was also found in the English bulldog

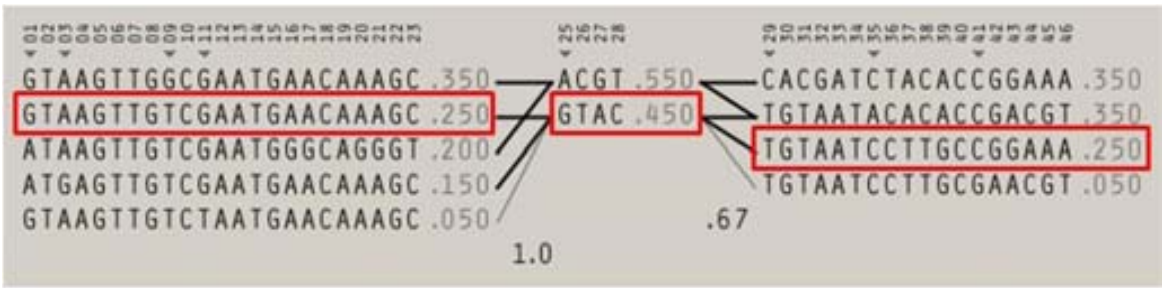
Figure 16 (*Continued*)



by Vaysse *et al.* and the chromosome locations largely overlapped in both studies (**Figure 14**, **Table 8**, **Figure 16a**). No other sweeps were reported by Vaysse *et al.* in the nearby region on CFA 26, and Boxer, French bulldog and Pug were not used in the search of selective sweeps.

The selective sweep on CFA 26 was not observed in the samples from non-brachycephalic dogs and wolf that we investigated. Normal levels of heterozygosity were observed in a first dataset containing 118 samples from 6 different dog breeds and 7 Iberian wolf samples genotyped using the same panel of SNPs as in set A (~22,000 SNPs) and on a second dataset containing 43 samples from German shepherd genotyped using the same SNP array as in set B (~172,000 SNPs) (see **PAPER II: Figure 3**).

Figure 17. Haplotype structure in the Pug for the region fixed in Boxer and English and French bulldogs (CFA 26:8.6–9.2 Mb). Haplotype blocks were inferred with Haploview (**Barrett** *et al.* **2005**) using the Four Gamete Rule requiring that the fourth gamete was observed at a population frequency >0.05 (so that it was seen in at least one of the 20 Pug chromosomes). Three haplotype blocks are shown with the different haplotypes in each block and their frequencies in the Pug samples. The numbers in the crossing areas indicate the level of recombination between two adjacent blocks. On the top, SNP indexes in accordance with those presented in **Figure 16b** and triangles indicating tag SNPs in blocks are shown. The haplotypes fixed in the other three brachycephalic breeds are highlighted in red.



Evolutionary relationships of the breeds

Next, we investigated on the relationship of the brachycephalic breeds studied in Paper II during the breed creation process and theorized about the evolutionary history of the sweep on CFA 26. It is known that the English bulldog contributed to the breed creation of both Boxer and French bulldog breeds (**Alderton and Bailey 2006**). The Boxer is believed to have originated from a long-existing and now extinct German breed, the Bullenbeisser, which was crossed with a small number of English bulldog exemplars exported from the United Kingdom. Likewise, the French bulldog originated from toy varieties of English bulldog that were more popular in France.

Given that the sweep is partially found in the three breeds it seems rational to think a possible scenario in which the local standing neutral variation present in the original English bulldog was passed to both Boxer and French bulldog during the breed creation process. Some variants would have been beneficial thereafter when selection of brachycephaly started, which is reasonable to think that happened during the breeds creation process since brachycephaly is a breed standard in these three types of dogs. Thus, strong selection of variants close to the position 8–10 Mb on CFA 26 contributing to brachycephaly might have swept nearby genetic variation. Variable selective sweep length in the three breeds would response to different breed histories as it depends on the strength of selection, the amount of recombination and the effective population size (**Maynard Smith and Haigh 1974**; **Kaplan** *et al.* **1989**; **Kim and Stephan 2002**). Therefore, if one assumes the recombination rate to be similar across breeds for a given chromosome region, different across-breeds strength of selection and population sizes might have probably caused the variable length in the sweeps on CFA 26, which in the Boxer is more than ten times larger than in the French bulldog (**Figure 16a**).

The fact that the haplotype that is fixed in these three breeds only showed intermediate frequencies in the Pug (**Figure 17**), also brachycephalic, might be explained by an independent history of this breed and by the brachycephalic phenotype being defined by multiple chromosomes. The Pug dates to the ancient China and it is suggested that interbreeding with Pekingese, Japanese chin and possibly Shih tzu contributed to the breed

creation process. Pugs were imported to Europe through Holland around 1,600s (**Alderton and Bailey 2006**). In fact, the two largest surveys of SNP and haplotype variation in dogs (**VonHoldt** *et al.* **2010**; **Vaysse** *et al.* **2011**) grouped together Boxer with English and French bulldogs, within the cluster of terriers, but separated from the Pug, which belongs to the group of toy dogs (see **Introduction: Figure 3**). The genetic nature of the brachycephalic head type phenotype is complex and many unlinked causal variants with variable effect may segregate at different frequencies in brachycephalic breeds.

Genetic content, putative targets of selection and potential negative side effects

Firstly, we tried to have a broad picture of the genetic content of the sweep on CFA 26 in its longest form, that is, in the Boxer, a genomic region that includes 135 annotated elements. Advantage was taken of the existence of human-dog synteny for the region of interest (CFA 26:3.0–11.9 Mb) and of its better annotation in the human genome in order to retrieve the functional annotation with regard to biological processes and diseases. Interestingly, the region is enriched for genes involved in skeletal and muscular system development and function as well as tissue morphology (**Figure 18**), biological processes necessarily involved in changes affecting the conformation of muzzle and skull. Besides, these genes are located nearby the local peaks of lowest genetic variation in the Boxer (**Figure 16a**) and selective sweeps are characterized by the strongest loss of variability happening in the immediate vicinity of the selected mutations (see **INTRODUCTION: Box 2**). Nonetheless, asymmetry in the valleys of reduced heterozygosity may provide imprecise information about the location of the sweep (**Kim and Stephan 2002**). On the other hand, the region contains genes that are linked to inherited diseases overrepresented in the Boxer (**Sargan 2004**) (**Figure 16a**). Lymphoblastic lymphoma is a type of non-Hodgkin lymphoma characterized by uncontrolled growth of either T- or B-cells and associated with *POLE*. The frequency of T-cell lymphoma in the Boxer is higher than in other breeds (**Lurie** *et al.* **2004**, **2008**; **Pastor** *et al.* **2009**). Dilated cardiomyopathy is a disorder characterized by cardiac enlargement (especially of the left ventricle), poor myocardial contractility, and congestive heart failure. *MYL2* is involved in the

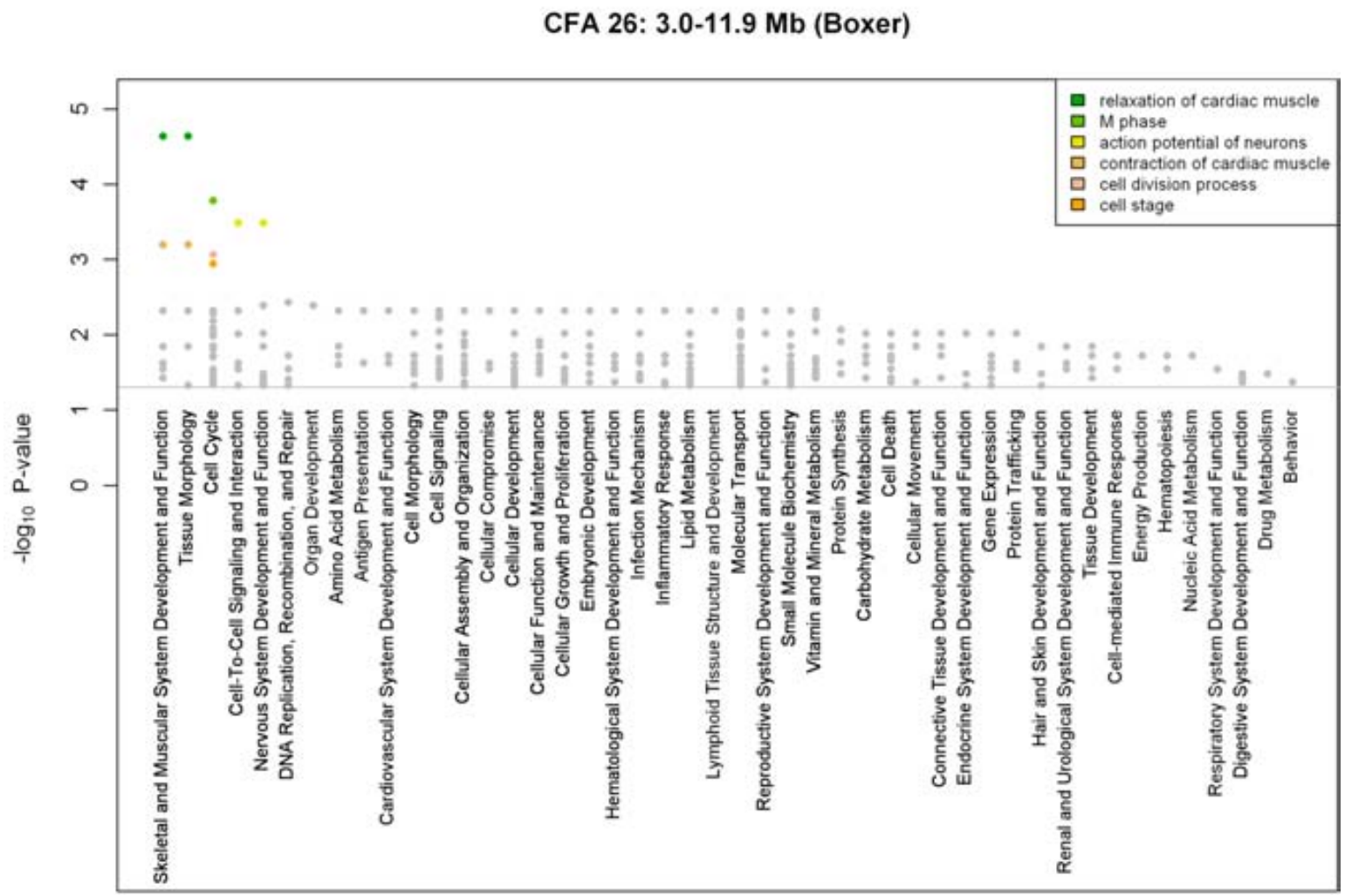
development of the sarcomere and muscle contraction and has also been associated with cardiomyopathy of the heart ventricle.

Secondly, we looked closer at the genes located in the part of the sweep shared by Boxer and English and French bulldogs and hypothetically associated with brachycephaly (CFA 26: 8.6–9.2 Mb) in order to shorten the list of putative selected sites. In general terms, genes in this region are involved in protein metabolism, cell cycle as well as in embryonic and tissue development (**Table 9**). Likewise, some genes in this region have been associated with genetic disorders in humans (**Table 9**). Loss-of-function mutations in *ATP6V0A2* cause autosomal recessive cutis laxa type II and some cases of wrinkly skin syndrome (**Kornak** *et al.* **2008**). This type of cutis laxa represents a spectrum of clinical entities with variable severity of loose skin, abnormal growth, developmental delay and associated skeletal abnormalities, and it is closely related with the wrinkly skin syndrome. *EIF2B1* encodes a subunit of a translation initiation factor and its mutation-caused malfunction can result in leukoencephalopathy with vanishing white matter (**van der Knaap** *et al.* **2002**), a neurological disorder manifesting progressive cerebellar ataxia, spasticity, inconstant optic atrophy and relatively preserved mental abilities. *SETD8* encodes a protein that regulates tumor suppressor p53 protein (**Shi** *et al.* **2007**).

Table 9. Genetic content of the part of the sweep shared in three brachycephalic breeds: Boxer and English and French bulldogs. The genes that were associated with biological processes or diseases (italic bold) enriched in the functional annotation analysis are presented. For simplicity, in the case of biological processes broader biological process categories are shown.

Human gene	Dog gene	Dog Start (bp)	Dog End (bp)	Dog Ensembl Gene ID	Biological process category or disease
ZNF664	*ZNF664*	8,718,870	8,719,655	ENSCAFG00000006994	__
CCDC92	*CCDC92*	8,759,998	8,790,267	ENSCAFG00000006996	Amino Acid Metabolism Protein Synthesis Small Molecule Biochemistry
DNAH10	*DNAH10*	8,791,705	8,871,910	ENSCAFG00000007042	__
ATP6V0A2	*ATP6V0A2*	8,929,687	8,968,952	ENSCAFG00000007234	***Cutis laxa, Wrinkly skin syndrome***
TCTN2	*TCTN2*	8,971,431	8,996,550	ENSCAFG00000023546	__
EIF2B1	*EIF2B1*	9,029,462	9,038,275	ENSCAFG00000007434	Gene Expression Protein Synthesis ***Leukoencephalopathy with vanishing white matter***
DDX55	*DDX55*	9,040,108	9,053,306	ENSCAFG00000007452	***Infection by HIV and of tumor and cervical cancer cell lines***
TMED2	*TMED2*	9,056,220	9,064,265	ENSCAFG00000007468	***Infection by HIV and of tumor and cervical cancer cell lines***
RILPL1	*RILPL1*	9,090,154	9,133,958	ENSCAFG00000007482	__
SNRNP35	*SNRNP35*	9,137,386	9,138,129	ENSCAFG00000007484	__
RILPL2	*RILPL2*	9,162,606	9,176,646	ENSCAFG00000007489	__
SETD8		9,188,951	9,202,238	ENSCAFG00000007493	Tissue development Cell cycle, death and movement Embryonic development

Figure 18. Biological process significantly enriched in the selective sweep on CFA 26 in the Boxer. Each dot represents a single biological process and they are grouped into broader biological process categories (labels in the x-axis). For the most significant associations (colored) the specific biological processes are in the legend.



Concluding remarks

In Paper II, we found a selective sweep previously associated with canine brachycephaly (**Bannasch** *et al.* **2010**) and a novel sweep of >8 Mb; both sweeps are likely to be representative of the Boxer breed as they were found in samples of variable geographic locations. We hypothesized about the possibility that the sweep we described encompasses an additional loci governing the brachycephalic phenotype in the dog: first, brachycephaly, characterized by severe shortening of the muzzle, is a breed-defining trait in the Boxer and thus it has been strongly selected in the breed; second, the sweep was partially detected in a few other breeds with brachycephaly but was absent in non-brachycephalic dog breeds and wolf. In addition, in Paper II we examined the putative loci in the sweep that would have been targeted by artificial selection for this trait and the possible undesired health consequences derived from such selection.

CONCLUSIONS

The specific aims of each of the two works presented in this thesis were different and so they are the conclusions that can be drawn from them.

With regard to the genetic control of canine leishmaniasis:

- Our results support that progression towards clinical disease from *Leishmania* infection in dogs is partly determined by host genetics.
- A substantial proportion of the genome is affecting the phenotype and its heritability could be as high as 60%, which is the first heritability estimate for this trait in any species.
- There was a significant predictive value from using the genomic information; one might anticipate that further genotyping of samples would increase accuracy to levels that have the potential for making an impact on both veterinary diagnostics and breeding.
- Three genomic locations on chromosomes 1, 4 and 20 showed the strongest association with the phenotype and the haplotype structure in these associations correlated with the affection status.
- However, genome-wide significance was not reached for any of the SNPs individually and therefore confirmation of these associations will require replication in an independent sample.

From the search of selective sweeps in the Boxer breed:

- We described a novel selective sweep of >8 Mb on chromosome 26 in the Boxer breed.
- We propose that the selective sweep on chromosome 26 encompasses an additional locus governing the complex brachycephalic head type in the dog.
- Selection for this morphological trait might have led to undesired health consequences for the breeds bearing it.

REFERENCES

Abitbol M, Thibaud JL, Olby NJ, Hitte C, Puech JP, Maurer M, Pilot-Storck F, Hedan B, Dreano S, Brahimi S *et al.*, (2010) A canine Arylsulfatase G (ARSG) mutation leading to a sulfatase deficiency is associated with neuronal ceroid lipofuscinosis. Proceedings of the National Academy of Sciences 107(33):14775-14780.

Abranches P, Silva-Pereira MC, Conceição-Silva FM, Santos-Gomes GM, Janz JG, (1991) Canine leishmaniasis: pathological and ecological factors influencing transmission of infection. J Parasitol 77(4):557-561.

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW, (2010) Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A 107(3):1160-1165.

Alderton D, Bailey G, PUBLISHING D, Kindersley D, Club AK, (2006) The Complete Dog Book. Random House Digital; 858 p.

Altet L, Francino O, Solano-Gallego L, Renier C, Sanchez A, (2002) Mapping and Sequencing of the Canine NRAMP1 Gene and Identification of Mutations in Leishmaniasis-Susceptible Dogs. Infect.Immun. 70(6):2763-2771.

Andersson L, (2009) Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. Genetica 136(2):341-349.

Andersson L, Georges M, (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet 5(3):202-212.

Awano T, Johnson GS, Wade CM, Katz ML, Johnson GC, Taylor JF, Perloski M, Biagi T, Baranowska I, Long S *et al.*, (2009) Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A 106(8):2794-2799.

Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K, (2012) Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Res 22(1):51-63.

Bannasch D, Young A, Myers J, Truvé K, Dickinson P, Gregg J, Davis R, Bongcam-Rudloff E, Webster MT, Lindblad-Toh K *et al.*, (2010) Localization of canine brachycephaly using an across breed mapping approach. PloS one 5(3):e9632.

Barber RM, Schatzberg SJ, Corneveaux JJ, Allen AN, Porter BF, Pruzin JJ, Platt SR, Kent M, Huentelman MJ, (2011) Identification of risk loci for necrotizing meningoencephalitis in Pug dogs. J Hered 102 Suppl 1:S40-S46.

Barrett JC, Fry B, Maller J, Daly MJ, (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-265.

Barros Roque J, O'Leary CA, Duffy DL, Kyaw-Tanner M, Latter M, Mason K, Vogelnest L, Shipstone M, (2011) IgE responsiveness to Dermatophagoides farinae in West Highland white terrier dogs is associated with region on CFA35. J Hered 102 Suppl 1:S74-S80.

Benjamini Y, Hochberg Y, (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) :289-289.

Bern C, Maguire JH, Alvar J, (2008) Complexities of assessing the disease burden attributable to leishmaniasis. PLoS neglected tropical diseases 2(10):e313.

Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L *et al.*, (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324(5926):522-528.

Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM *et al.*, (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324(5926):528-532.

Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R *et al.*, (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. Proc Natl Acad Sci U S A 106(33):13903-13908.

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

Browning SR, Browning BL, (2011) Population structure can inflate SNP-based heritability estimates. Am J Hum Genet 89(1):191-3; author reply 193.

Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, VonHoldt BM, Rhue A, Boyko A, Byers A *et al.*, (2009) Coat variation in the domestic dog is governed by variants in three genes. Science 326(5949):150-153.

Chinwalla MGSCT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD *et al.*, (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520-562.

Clark LA, Wahl JM, Rees CA, Murphy KE, (2006) Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc Natl Acad Sci U S A 103(5):1376-1381.

Clutton-Brock J, (1995) Origins of the dog: domestication and early history. In: Serpell J, ed. *The Domestic Dog, its Evolution, Behavior and Interactions With People.* Cambridge: Cambridge University Press;7-20.

Crossa J, Campos GDL, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J *et al.*, (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713-724.

Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD, (2007) Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Research 17(12):1743-1754.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM, (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1):375-385.

de los Campos G, Gianola D, Allison DB, (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Genes Immun 11(12):886-886.

Devlin B, Roeder K, (1999) Genomic control for association studies. Biometrics 55(4):997-1004.

Ding ZL, Oskarsson M, Ardalan A, Angleby H, Dahlgren LG, Tepeli C, Kirkness E, Savolainen P, Zhang YP, (2011) Origins of domestic dog in Southern East Asia is supported by analysis of Y-chromosome DNA. Heredity.

Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW *et al.*, (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet 7(6):e1002141.

Dodman NH, Karlsson EK, Moon-Fanelli A, Galdzicka M, Perloski M, Shuster L, Lindblad-Toh K, Ginns EI, (2010) A canine chromosome 7 locus confers compulsive disorder susceptibility. Mol Psychiatry 15(1):8-10.

Downs LM, Wallin-Håkansson B, Boursnell M, Marklund S, Hedhammar Å, Truvé K, Hübinette L, Lindblad-Toh K, Bergström T, Mellersh CS *et al.*, (2011) A frameshift mutation in golden retriever dogs with progressive retinal atrophy endorses SLC4A3 as a candidate gene for human retinal degenerations. PloS one 6(6):e21452.

Drogemuller C, Karlsson EK, Hytonen MK, Perloski M, Dolf G, Sainio K, Lohi H, Lindblad-Toh K, Leeb T, (2008) A Mutation in Hairless Dogs Implicates FOXI3 in Ectodermal Development. Science 321(5895):1462-1462.

Elferink MG, Megens HJ, Vereijken A, Hu X, Crooijmans RPMA, Groenen MAM, (2012) Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PloS one 7(2):e32720.

Ellis JL, Thomason J, Kebreab E, Zubair K, France J, (2009) Cranial dimensions and forces of biting in the domestic dog. J Anat 214(3):362-373.

Falconer DS, Mackay TFC, (1996) Introduction to quantitative genetics. Longman Pub Group; 464 p.

Fan B, Onteru SK, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF, (2011) Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. PloS one 6(2):e14726.

Finlay EK, Berry DP, Wickham B, Gormley EP, Bradley DG, (2012) A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle. PloS one 7(2):e30545.

Foley CW, Lasley JF, Osweiler GD, Digestive system. *Abnormalities of companion animals.* Iowa State Pr; (1979);119-120.

Fu WX, Liu Y, Lu X, Niu XY, Ding XD, Liu JF, Zhang Q, (2012) A Genome-Wide Association Study Identifies Two Novel Promising Candidate Genes Affecting Escherichia coli F4ab/F4ac Susceptibility in Swine. PloS one 7(3):e32127.

Galibert F, André C, Chéron A, Chuat JC, Hitte C, Jiang Z, Jouquand S, Priat C, Rénier C, Vignaux F *et al.*, (1998) [The importance of the canine model in medical genetics]. Bull Acad Natl Med 182(4):811-21; discussion 822.

Garrick DJ, (2010) The nature, scope and impact of some whole-genome analyses in beef cattle. 9th World Congress on Genetics Applied to Livestock; Leipzig.

Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC *et al.*, (2011) Genetics of venous thrombosis: insights from a new genome wide association study. PloS one 6(9):e25581.

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(6982):493-521.

Goldstein O, Mezey JG, Boyko AR, Gao C, Wang W, Bustamante CD, Anguish LJ, Jordan JA, Pearce-Kelling SE, Aguirre GD *et al.*, (2010) An ADAM9 mutation in canine cone-rod dystrophy 3 establishes homology with human cone-rod dystrophy 9. Mol Vis 16:1549-1569.

González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendaño S, (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178(4):2305-2313.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS, (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37(5):549-554.

Habier D, Fernando RL, Dekkers JCM, (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177(4):2389-2397.

Harris B, Johnson D, Spelman R, Sattler J, others, (2009) Genomic selection in New Zealand and the implications for national genetic evaluation. (13):325-325.

Hayes HM, Priester WA, Pendergrass TW, (1975) Occurrence of nervous-tissue tumors in cattle, horses, cats and dogs. Int J Cancer 15(1):39-47.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME, (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2):433-443.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME *et al.*, (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432(7018):695-716.

Hirschhorn JN, Daly MJ, (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95-108.

Holzapfel C, Grallert H, Huth C, Wahl S, Fischer B, Döring A, Rückert IM, Hinney A, Hebebrand J, Wichmann HE *et al.*, (2010) Genes and lifestyle

factors in obesity: results from 12,462 subjects from MONICA/KORA. Int J Obes (Lond) 34(10):1538-1545.

Humphray SJ, Scott CE, Clark R, Marron B, Bender C, Camm N, Davis J, Jenks A, Noon A, Patel M *et al.*, (2007) A high utility integrated map of the pig genome. Genome Biol 8(7):R139.

Ibrahim M, Lambson B, Yousif A, Deifalla N, Alnaiem D, Ismail A, Yousif H, Ghalib H, Khalil E, Kadaro A *et al.*, (1999) Kala-azar in a high transmission focus: an ethnic and geographic dimension. Am.J.Trop.Med.Hyg. 61(6):941-944.

International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748-752.

International Sheep Genomics Consortium, Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW *et al.*, (2010) The sheep genome reference sequence: a work in progress. Anim Genet 41(5):449-453.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R *et al.*, (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Genes Immun 451(7181):1003-1003.

Jeronimo SMB, Holst AKB, Jamieson SE, Francis R, Martins DRA, Bezerra FL, Ettinger NA, Nascimento ET, Monteiro GR, Lacerda HG *et al.*, (2007a) Genes at human chromosome 5q31.1 regulate delayed-type hypersensitivity responses associated with Leishmania chagasi infection. Genes Immun 8(7):539-551.

Jeronimo SMB, Duggal P, Ettinger NA, Nascimento ET, Monteiro GR, Cabral AP, Pontes NN, Lacerda HG, Queiroz PV, Gomes CEM *et al.*, (2007b) Genetic Predisposition to Self-Curing Infection with the Protozoan Leishmania chagasi: A Genomewide Scan. Journal of Infectious Diseases 196(8):1261-1269.

Kaplan N, Hudson R, Langley C, (1989) The "hitchhiking effect" revisited. Genetics 123:887-899.

Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ *et al.*, (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet 39(11):1321-1328.

Karlsson EK, Lindblad-Toh K, (2008) Leader of the pack: gene mapping in dogs and other model organisms. Nat Rev Genet 9(9):713-725.

Karplus TM, Jeronimo SMB, Chang H, Helms BK, Burns TL, Murray JC, Mitchell AA, Pugh EW, Braz RFS, Bezerra FL *et al.*, (2002) Association between the Tumor Necrosis Factor Locus and the Clinical Outcome of Leishmania chagasi Infection. Infect.Immun. 70(12):6919-6925.

Kedzierski L, (2010) Leishmaniasis Vaccine: Where are We Today? Journal of global infectious diseases 2(2):177-185.

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J *et al.*, (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21(10):1233-1237.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F *et al.*, (2008) Mapping and sequencing of structural variation from eight human genomes. Genes Immun 453(7191):64-64.

Kim Y, Stephan W, (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160(2):765-777.

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.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L *et al.*, (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318(5849):420-426.

Kornak U, Reynders E, Dimopoulou A, van Reeuwijk J, Fischer B, Rajab A, Budde B, Nürnberg P, Foulquier F, ARCL Debré-type Study Group *et al.*, (2008) Impaired glycosylation and cutis laxa caused by mutations in the vesicular H+-ATPase subunit ATP6V0A2. Nat Genet 40(1):32-34.

Lander ES, Schork NJ, (1994) Genetic dissection of complex traits. Science 265(5181):2037-2048.

Lander IHGSCES, 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(6822):860-921.

Lee SH, Wray NR, Goddard ME, Visscher PM, (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88(3):294-305.

Lee SH, Decandia TR, Ripke S, Yang J, The Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), The International Schizophrenia Consortium (ISC), The Molecular Genetics of Schizophrenia Collaboration (MGS), Sullivan PF, Goddard ME, Keller MC *et al.*, (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet 44(3):247-250.

Lewis TW, Blott SC, Woolliams JA, (2010a) Genetic evaluation of hip score in UK Labrador Retrievers. PloS one 5(10):e12797.

Lewis TW, Woolliams JA, Blott SC, (2010b) Genetic evaluation of the nine component features of hip score in UK Labrador Retrievers. PloS one 5(10):e13610.

Lewis T, Swift S, Woolliams JA, Blott S, (2011) Heritability of premature mitral valve disease in Cavalier King Charles spaniels. Vet J 188(1):73-76.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC *et al.*, (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069):803-819.

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, González-Recio O, (2010) Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics research 92(3):209-225.

Lund MS, Sahana G, de Koning DJ, Su G, Carlborg O, (2009) Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC proceedings 3 Suppl 1:S1.

Lurie DM, Lucroy MD, Griffey SM, Simonson E, Madewell BR, (2004) T-cell-derived malignant lymphoma in the boxer breed. Veterinary and comparative oncology 2(3):171-175.

Lurie DM, Milner RJ, Suter SE, Vernau W, (2008) Immunophenotypic and cytomorphologic subclassification of T-cell lymphoma in the boxer breed. Vet Immunol Immunopathol 125(1-2):102-110.

Madsen MB, Olsen LH, Haggstrom J, Hoglund K, Ljungvall I, Falk T, Wess G, Stephenson H, Dukes-McEwan J, Chetboul V *et al.*, (2011) Identification of 2 loci associated with development of myxomatous mitral valve disease in Cavalier King Charles Spaniels. J Hered 102 Suppl 1:S62-S67.

Martinez A, Simon L, Tejedor D, Artieda M, Bartolome N, Escaich J, Velasco A, Selles M, Chetrit C, Martinez D *et al.*, (2012) Markers for joint displasia, osteoarthritis and conditions secondary thereto. Patent USA 61/49739915. Appl. No.: Patent WO2012038382 (A2)

Maynard Smith J, Haigh J, (1974) The hitch-hiking effect of a favourable gene. Genetic Research 23:23-25.

Meddeb-Garnaoui A, Gritli S, Garbouj S, Ben Fadhel M, El Kares R, Mansour L, Kaabi B, Chouchane L, Ben Salah A, Dellagi K *et al.*, (2001) Association analysis of HLA-class II and class III gene polymorphisms in the susceptibility to mediterranean visceral leishmaniasis. Hum Immunol 62(5):509-517.

Meredith BK, Kearney FJ, Finlay EK, Bradley DG, Fahey AG, Berry DP, Lynn DJ, (2012) Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. BMC Genet 13(1):21.

Merveille AC, Davis EE, Becker-Heck A, Legendre M, Amirav I, Bataille G, Belmont J, Beydon N, Billen F, Clément A *et al.*, (2011) CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. Nat Genet 43(1):72-78.

Meuwissen T, Hayes BJ, Goddard ME, (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Miranda S, Roura X, Alberola J, Ferrer L, Ramis A (2005) Clinically patent canine leishmaniasis shows age, breed and sex predilection. World-leish3, Third World Congress on Leishmaniosis; 2005; Palermo,Terrassini, Sicily, Italy.

Mogensen MS, Karlskov-Mortensen P, Proschowsky HF, Lingaas F, Lappalainen A, Lohi H, Jensen VF, Fredholm M, (2011) Genome-wide association study in Dachshund: identification of a major locus affecting intervertebral disc calcification. J Hered 102 Suppl 1:S81-S86.

Nelson R, Couto C, Disorders of the Laynx and Pharynx. *Small Animal Internal Medicine. third ed..* Missouri: Mosby; (2003);:248-249.

Nielsen R, (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197-218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG, (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8(11):857-868.

Nöller C, Hueber J, Aupperle H, Seeger J, Oechtering TH, (2002) New Aspects of Brachycephalia in Dogs & Cats Basics: Insights Into Embryology. Anatomy & Pathophysiology; 2002; Granada, Spain.

Olsson M, Meadows JRS, Truvé K, Rosengren Pielberg G, Puppo F, Mauceli E, Quilez J, Tonomura N, Zanna G, Docampo MJ *et al.*, (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 7(3):e1001332.

Ostrander EA, Giniger E, (1997) Semper fidelis: what man's best friend can teach us about human biology and disease. Am J Hum Genet 61(3):475-480.

Ostrander EA, Kruglyak L, (2000) Unleashing the canine genome. Genome Res 10(9):1271-1274.

Ostrander EA, Galibert F, Patterson DF, (2000) Canine genetics comes of age. Trends Genet 16(3):117-124.

Ostrander EA, (2005) The canine genome. Genome Res 15(12):1706-1716.

Paigen K, (1995) A miracle enough: the power of mice. Nat Med 1(3):215-220.

Pang JF, Kluetsch C, Zou XJ, Zhang AB, Luo LY, Angleby H, Ardalan A, Ekstrom C, Skollermo A, Lundeberg J *et al.*, (2009) mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Mol Biol Evol 26(12):2849-2864.

Parker HG, Ostrander EA, (2005) Canine genomics and genetics: running with the pack. PLoS Genet 1(5):e58.

Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG *et al.*, (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325(5943):995-998.

Pastor M, Chalvet-Monfray K, Marchal T, Keck G, Magnol JP, Fournel-Fleury C, Ponce F, (2009) Genetic and environmental risk indicators in canine non-Hodgkin's lymphomas: breed associations and geographic distribution of 608 cases diagnosed throughout France over 1 year. J Vet Intern Med 23(2):301-310.

Patterson DF, Haskins ME, Jezyk PF, (1982) Models of human genetic disease in domestic animals. Adv Hum Genet 12:263-339.

Patterson DF, Haskins ME, Jezyk PF, Giger U, Meyers-Wallen VN, Aguirre G, Fyfe JC, Wolfe JH, (1988) Research on genetic diseases: reciprocal benefits to animals and man. J Am Vet Med Assoc 193(9):1131-1144.

Peacock CS, Collins A, Shaw MA, Silveira F, Costa J, Coste CH, Nascimento MD, Siddiqui R, Shaw JJ, Blackwell JM *et al.*, (2001) Genetic epidemiology of visceral leishmaniasis in northeastern Brazil. Genet Epidemiol 20(3):383-396.

Pielberg G, Golovko A, Sundstrm E, Curik I, Lennartsson J, Seltenhammer M, Druml T, Binns M, Fitzsimmons C, Lindgren G *et al.*, (2008) A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. Nat Genet 40(8):1004-1004.

Pong-Wong R, Hadjipavlou G, (2010) A two-step approach combining the Gompertz growth model with genomic selection for longitudinal data. BMC proceedings 4 Suppl 1:S4.

Price A, Zaitlen N, Reich D, Patterson N, (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459-463.

Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, Galibert F, André C, Ostrander EA, Hitte C *et al.*, (2007) Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. PloS one 2(12):e1324.

Quinnell RJ, Kennedy LJ, Barnes A, Courtenay O, Dye C, Garcez LM, Shaw MA, Carter SD, Thomson W, Ollier WER *et al.*, (2003) Susceptibility to visceral leishmaniasis in the domestic dog is associated with MHC class II polymorphism. Immunogenetics 55(1):23-28.

Rubin CJ, Brändström H, Wright D, Kerje S, Gunnarsson U, Schutz K, Fredriksson R, Jensen P, Andersson L, Ohlsson C *et al.*, (2007) Quantitative trait loci for BMD and bone strength in an intercross between domestic and wildtype chickens. J Bone Miner Res 22(3):375-384.

Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S *et al.*, (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464(7288):587-591.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES *et al.*, (2006) Positive natural selection in the human lineage. Science 312(5780):1614-1620.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R *et al.*, (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913-918.

Sanchez-Robert E, Altet L, Sanchez A, Francino O, (2005) Polymorphism of Slc11a1 (Nramp1) gene and canine leishmaniasis in a case-control study. J Hered 96(7):755-758.

Sanchez-Robert E, Altet L, Utzet-Sadurni M, Giger U, Sanchez A, Francino O, (2008a) Slc11a1 (formerly Nramp1) and susceptibility to canine visceral leishmaniasis. Vet Res 39(3):36.

Sanchez-Robert E, Altet L, Alberola J, Rodriguez-Cortés A, Ojeda A, López-Fuertes L, Timon M, Sanchez A, Francino O, (2008b) Longitudinal analysis of cytokine gene expression and parasite load in PBMC in Leishmania infantum experimentally infected dogs. Vet Immunol Immunopathol 125(1-2):168-175.

Sargan DR, (2004) IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. Mamm Genome 15(6):503-506.

Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T, (2002) Genetic evidence for an East Asian origin of domestic dogs. Science 298(5598):1610-1613.

Seppälä EH, Jokinen TS, Fukata M, Fukata Y, Webster MT, Karlsson EK, Kilpinen SK, Steffen F, Dietschi E, Leeb T *et al.*, (2011) LGI2 truncation causes a remitting focal epilepsy in dogs. PLoS Genet 7(7):e1002194.

Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, Bis JC, Smith AV, Carassquillo MM, Lambert JC *et al.*, (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA 303(18):1832-1840.

Shi X, Kachirskaia I, Yamaguchi H, West L, Wen H, Wang E, Dutta S, Appella E, Gozani O, (2007) Modulation of p53 function by SET8-mediated methylation at lysine 382. Mol Cell 27(4):636-646.

Simonson MA, Wills AG, Keller MC, McQueen MB, (2011) Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. BMC Med Genet 12:146.

Solano-Gallego L, Llull J, Ramos G, Riera C, Arboix M, Alberola J, Ferrer L, (2000) The Ibizian hound presents a predominantly cellular immune response against natural Leishmania infection. Vet Parasitol 90(1-2):37-45.

Solano-Gallego L, Morell P, Arboix M, Alberola J, Ferrer L, (2001) Prevalence of Leishmania infantum infection in dogs living in an area of canine leishmaniasis endemicity using PCR on several tissues and serology. J Clin Microbiol 39(2):560-563.

Stockard CR, Anderson OD, James WT, Wistar Institute of Anatomy and Biology, The genetic and endocrinic basis for differences in form and behavior. (1941);775 p.

Sturtevant AH, (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool 14(1):43-59.

Sutter NB, Ostrander EA, (2004) Dog star rising: the canine genetic system. Nat Rev Genet 5(12):900-910.

The International HapMap Consortium, (2005) A haplotype map of the human genome. Nature. 437(7063):1299-1320.

The International HapMap Consortium, (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Tsai KL, Noorai RE, Starr-Moss AN, Quignon P, Rinz CJ, Ostrander EA, Steiner JM, Murphy KE, Clark LA, (2012) Genome-wide association studies for multiple diseases of the German Shepherd Dog. Mamm Genome 23(1-2):203-211.

Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC *et al.*, (2010) Predicting phenotype from genotype: normal pigmentation. J Forensic Sci 55(2):315-322.

van der Knaap MS, Leegwater PAJ, Könst AAM, Visser A, Naidu S, Oudejans CBM, Schutgens RBH, Pronk JC, (2002) Mutations in each of the five subunits of translation initiation factor eIF2B can cause leukoencephalopathy with vanishing white matter. Ann Neurol 51(2):264-270.

van Hoek M, Dehghan A, Witteman JCM, van Duijn CM, Uitterlinden AG, Oostra BA, Hofman A, Sijbrands EJG, Janssens ACJW, (2008) Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 57(11):3122-3128.

Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M *et al.*, (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425(6960):832-836.

VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, Schenkel F, (2009) Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16-16.

Vattikuti S, Guo J, Chow CC, (2012) Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. PLoS Genet 8(3):e1002637.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MST, Lawley CT *et al.*, (2011) Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLoS Genet 7(10):e1002316.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*, (2001) The sequence of the human genome. Science 291(5507):1304-1351.

Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK, (1997) Multiple and ancient origins of the domestic dog. Science 276(5319):1687-1689.

VonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A *et al.*, (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464(7290):898-902.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR *et al.*, (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326(5954):865-867.

Wayne RK, (1993) Molecular evolution of the dog family. Trends Genet 9(6):218-224.

Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, Rosa GJM, Gianola D, (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92(10):5248-5257.

WHO Expert Committee on the Control of the Leishmaniases, Control of the Leishmaniases: Report of a Meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva, 22-26 March 2010.

Wilbe M, Jokinen P, Truvé K, Seppala EH, Karlsson EK, Biagi T, Hughes A, Bannasch D, Andersson G, Hansson-Hamlin H *et al.*, (2010) Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. Nat Genet 42(3):250-254.

Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C *et al.*, (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41(1):25-34.

Wood SH, Ke X, Nuttall T, McEwan N, Ollier WE, Carter SD, (2009) Genome-wide association analysis of canine atopic dermatitis and identification of disease related SNPs. Immunogenetics 61(11-12):765-772.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al.*, (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565-569.

Yang J, Lee SH, Goddard ME, Visscher PM, (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76-82.

Zhao J, Bradfield JP, Li M, Wang K, Zhang H, Kim CE, Annaiah K, Glessner JT, Thomas K, Garris M *et al.*, (2009) The role of obesity-associated loci identified in genome-wide association studies in the determination of pediatric BMI. Obesity (Silver Spring) 17(12):2254-2257.

Zhao X, Dittmer KE, Blair HT, Thompson KG, Rothschild MF, Garrick DJ, (2011) A novel nonsense mutation in the DMP1 gene identified by a genome-wide association study is responsible for inherited rickets in Corriedale sheep. PloS one 6(7):e21739.

PAPER I

OPEN ACCESS Freely available online

PLoS one

Genetic Control of Canine Leishmaniasis: Genome-Wide Association Study and Genomic Selection Analysis

Javier Quilez[1,2]*, Verónica Martínez[1,2], John A. Woolliams[4], Armand Sanchez[1,2,3], Ricardo Pong-Wong[4], Lorna J. Kennedy[5], Rupert J. Quinnell[6], William E. R. Ollier[5], Xavier Roura[7], Lluís Ferrer[7], Laura Altet[2,3], Olga Francino[2,3]

1 Departament de Genètica Animal, Centre de Recerca en Agrigenòmica (CRAG), Universitat Autònoma de Barcelona, Barcelona, Spain, **2** Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Barcelona, Spain, **3** Servei Veterinari de Genètica Molecular, Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Barcelona, Spain, **4** The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Scotland, United Kingdom, **5** Centre for Integrated Genomic Medical Research (CIGMR), University of Manchester, Stopford Building, Oxford Road, Manchester, United Kingdom, **6** Institute of Integrative and Comparative Biology, University of Leeds, Leeds, United Kingdom, **7** Hospital Clínic Veterinari, Universitat Autònoma de Barcelona, Barcelona, Spain

Abstract

Background: The current disease model for leishmaniasis suggests that only a proportion of infected individuals develop clinical disease, while others are asymptomatically infected due to immune control of infection. The factors that determine whether individuals progress to clinical disease following *Leishmania* infection are unclear, although previous studies suggest a role for host genetics. Our hypothesis was that canine leishmaniasis is a complex disease with multiple loci responsible for the progression of the disease from *Leishmania* infection.

Methodology/Principal Findings: Genome-wide association and genomic selection approaches were applied to a population-based case-control dataset of 219 dogs from a single breed (Boxer) genotyped for ~170,000 SNPs. Firstly, we aimed to identify individual disease loci; secondly, we quantified the genetic component of the observed phenotypic variance; and thirdly, we tested whether genome-wide SNP data could accurately predict the disease.

Conclusions/Significance: We estimated that a substantial proportion of the genome is affecting the trait and that its heritability could be as high as 60%. Using the genome-wide association approach, the strongest associations were on chromosomes 1, 4 and 20, although none of these were statistically significant at a genome-wide level and after correcting for genetic stratification and lifestyle. Amongst these associations, chromosome 4: 61.2–76.9 Mb maps to a locus that has previously been associated with host susceptibility to human and murine leishmaniasis, and genomic selection estimated markers in this region to have the greatest effect on the phenotype. We therefore propose these regions as candidates for replication studies. An important finding of this study was the significant predictive value from using the genomic information. We found that the phenotype could be predicted with an accuracy of ~0.29 in new samples and that the affection status was correctly predicted in 60% of dogs, significantly higher than expected by chance, and with satisfactory sensitivity-specificity values (AUC = 0.63).

Citation: Quilez J, Martínez V, Woolliams JA, Sanchez A, Pong-Wong R, et al. (2012) Genetic Control of Canine Leishmaniasis: Genome-Wide Association Study and Genomic Selection Analysis. PLoS ONE 7(4): e35349. doi:10.1371/journal.pone.0035349

Editor: Amanda Ewart Toland, Ohio State University Medical Center, United States of America

Received December 7, 2011; **Accepted** March 14, 2012; **Published** April 25, 2012

Funding: This work was funded by the European Commission (LUPA, GA-201370) as well as by the Consolider-Ingenio 2010 Programme CSD2007-00036 "Centre for Research in Agrigenomics" from the Spanish Ministry of Science and Innovation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: javier.quilez@cragenomica.es

Introduction

Leishmaniasis is a vector-borne disease affecting humans and animals, caused by parasitic species of the genera *Leishmania* and transmitted by the bite of phlebotomine sand flies. Around the Mediterranean basin, visceral (VL) and cutaneous (CL) human leishmaniasis as well as canine leishmaniasis (CanL) are caused by *Leishmania infantum*. The current disease model for leishmaniasis suggests that infected individuals may live without progression towards clinical disease manifestation probably due to immune control of the infection.

The factors that determine whether individuals progress to clinical disease following *Leishmania* infection are unclear, but previous studies suggest a large contribution of the host genetic background, as reviewed elsewhere [1,2]. Studies in mice [1] provided early support for a strong genetic component to susceptibility to *Leishmania* infection. In humans, most epidemiological studies [3,4,5,6], candidate gene studies [7,8,9,10,11,12] and genome-wide approaches [7,13,14] have offered further support for genetic susceptibility to leishmaniasis, however they did not specifically dissect the genetic factors that cause progression of the disease following infection. Some studies have

investigated genetic differences between healthy infected and symptomatic individuals, but most of these were either not aimed to identify candidate loci [15,16] or targeted at few candidate genes [17,10,18]. Only Jeronimo *et al.* [19] have studied progression of leishmaniasis following infection using a genome-wide linkage approach in humans based on a few hundred microsatellite markers. In dogs, genetic susceptibility to progression of disease from *Leishmania* infection is supported by the fact that the percentage of infected dogs in endemic areas is as high as 60% [20] whereas rates of clinical CanL are much lower in these areas [21,22]. Similarly to familial aggregation and ethnic differences of leishmaniasis prevalence seen in humans, dog breeds show variable susceptibility to CanL. Some breeds such as Boxer, German shepherd and Rottweiler [23,24,25] appear more predisposed to overt CanL. In contrast, the Ibizan hound, a dog breed believed to have been relatively isolated in an endemic area such as Ibiza (Balearic Islands, Spain), is reported to be resistant to CanL [26].

Understanding the genomic factors controlling progression to clinical disease in dogs is critical since the dog is the main natural reservoir of *Leishmania infantum* infection for humans, and CanL is a disease of great importance in veterinary medicine because of its severity in the dog. Despite the importance of leishmaniasis in dogs, there have been very few genetic studies of this species and these have focused on a few candidate genes [27,25,28,29,30], which have confirmed some genes previously found in mice and humans. There have been no previous genome-wide studies of genetic susceptibility to visceral leishmaniasis in the dog.

The dog has been previously proposed as a comparative animal genetic model for disease mapping. For complex diseases, a strategy with a first genome-wide scan genotyping tens of thousands of single-nucleotide polymorphisms (SNPs) for a few hundreds of dogs from one or few breeds has been suggested [31,32,33] based on calculations of statistical power. This approach has been based on simulation studies. For complex phenotypes, these simulation studies demonstrate that 100–300 cases and 100–300 controls provide adequate power to detect alleles conferring 2 to 5-fold multiplicative risk [33]. As a proof of principle, the efficacy of the proposed design has recently been demonstrated on several different studies [34,35,36,37,38,39, 40,41,42,43,44,45]. Moreover, Daetwyler and collaborators [46] showed that the predictive accuracy depends upon the genomic structure of the species, and this is favorable for canine studies because of its low effective population increases the power in genomic selection techniques [47].

The aim of this work was therefore to carry out a genome-wide study of the genetic contribution to the progression of clinical CanL from *Leishmania* infection. Our working hypothesis were: (i) that the observed phenotypic variance in the progression of leishmaniasis in infected dogs is partly explained by the genetics of the host; (ii) that CanL is a complex disease with multiple loci involved and an environmental component; and (iii) that genomic information may be used to predict the progression of the disease. We applied both genome-wide association study (GWAS) and genomic selection approaches to a population-based case-control dataset of 219 dogs from a single dog breed (Boxer) genotyped for ~170,000 single-nucleotide polymorphisms (SNPs) in order to study host genetic susceptibility to progression of clinical leishmaniasis from *Leishmania* infection. Firstly, we tried to identify loci in the canine genome associated with the disease progression phenotype. Secondly, we investigated the genetic component of the observed phenotypic variance. Thirdly, we examined whether genome-wide SNP data could be used to predict accurately the phenotype.

Results

Genome-wide scan of loci affecting disease progression

A GWAS analysis testing markers individually was performed in order to find loci associated with the progression to clinical CanL from *Leishmania* infection, using a dataset of 115 healthy infected and 104 affected Boxer dogs. All dogs had genotypes for 126,607 SNPs distributed across the genome.

Three statistical models were applied by fitting additional covariates in order to correct for the two confounding effects considered (described in **Materials and Methods**). When no covariates were included (Model 1), the strongest associations were found on *Canis familiaris* chromosomes (CFA) 1:39,058,553 bp ($P_{raw} = 1.0\times10^{-5}$, $P_{genome} = 0.21$), CFA 4: 68,238,371 bp ($P_{raw} = 1.1\times10^{-5}$, $P_{genome} = 0.22$) and CFA 20: 30,132,329 bp ($P_{raw} = 2.5\times10^{-5}$, $P_{genome} = 0.43$) (**Figure S1** and **Table S1**). Although healthy infected and affected samples generally clustered together in the MDS plot (**Figure S2**), genetic stratification was observed in our cleaned dataset based on the genomic inflation factor ($\lambda = 1.29$), with C1 capturing twice the stratification captured by C2. The associations on CFA1 and 4 remained when confounding effects were accounted for although significance did not reach the genome-wide level (**Table S1**). Genetic stratification, corrected by fitting the two first dimensions from the multidimensional scaling analysis (C1 and C2), is likely to explain part of the initial association in Model 1, as P_{raw} values for the ten strongest associated SNPs on CFA 1 and 4 were an order of magnitude higher when stratification was accounted for (Model 2). Nevertheless, associations of C1 and C2 with each of these markers were not significant (data not shown). Inclusion of dog lifestyle as a confounding effect did not affect the significance of the markers. After correction for the confounder effects (Model 3) the inflation factor was reduced to $\lambda = 1.17$ and this was not reduced by adding three additional MDS dimensions which altogether captured an extra 5% of the genetic variance in the markers (**Table S2**).

We examined candidate loci previously reported to have associations with host response to *Leishmania* infection and susceptibility to leishmaniasis in *Homo sapiens* (49 loci) and *Mus musculus* (33 loci) to test in a systematic way if any of these loci showed a stronger association in our canine dataset. When possible, these were mapped to their orthologues in the dog genome, and this was successful for 78 loci (95%; **Dataset S1**). We selected SNPs in the GWAS data contained within these candidate loci and their flanking regions (±1 Mb) and assigned them to sets of non-overlapping candidate regions. This resulted in 4,751 SNPs in 37 sets with a median of 108 SNPs (**Dataset S1**). Sets were tested one at a time for association with the phenotype controlling for within-set linkage disequilibrium (LD) and multiple testing arising from the number of SNPs in the set as described elsewhere [48] ($r^2 = 0.80$ and $p = 0.05$ were used). Three sets of SNPs on CFA 4, and one each on CFA 9 and 10 showed an empirical set-specific p-value (EMP1)<0.05 (**Dataset S1**). The same sets showed EMP1<0.01 when $r^2 = 0.10$ and $p = 0.01$ were applied (see **Materials and Methods**). Although EMP1 does not account for the fact that multiple sets are tested, the sets on CFA 4 showed EMP1 values notably lower than for other sets (**Figure S3**). The sets on CFA 4 spanned the region 61.2–76.9 Mb which had previously showed the strongest associations in the initial GWAS (**Table S1**). All the sets contain loci associated with *Leishmania* infection, and the three sets on CFA 4 included several genes (*Il7r*, *Lifr*, *C6*, *C7* and *Csf1r*) that lie within a locus involved in lesion development in murine *Leishmania major* infection [49,50,51,52].

Finally, the same dataset used in the GWAS was analysed using genomic selection with the BayesB method [47] with some modifications previously published [53]. Briefly, the BayesB method first proposed by Meuwissen *et al.* [47] is a Bayesian model in which the effect of SNPs on the total genetic values are predicted simultaneously, with an *a priori* assumption that only few SNPs are useful for predicting the trait. With the modified BayesB method we used (from now onwards referred just as BayesB for simplicity), Models 1–3 produced a similar genome-wide plot of both estimated marker effects (**Figure S4**) and the proportion of realisations a given marker was estimated to have a non-zero effect (data not shown), with a most detectable peak on CFA 4:61–77 Mb. This region overlapped with both the strongest association in GWAS and the region in which SNP sets covering candidate genes were significant (EMP1$<$0.01).

Estimating genetic variance in the phenotype

GWAS methodology is concerned with identifying individual SNPs that may be a causative variant for the phenotype or in LD with such a variant. Despite the failure to detect any such SNP, it was possible to detect genetic variation relating to the leishmaniasis phenotype. Two different approaches were adopted, the first using a modified BayesB methodology [53] and the second a Restricted Maximum Likelihood (REML) methodology implemented within the GCTA package [54]. The estimates of heritability obtained were 0.64 and 0.58 (s.e. 0.17) from BayesB and GCTA, respectively (**Table 1**). These estimates were corrected for genetic stratification (C1, C2) and lifestyle. Note that these estimates are likely to be biased upwards because of the selection of the samples contributing to the study – as would be expected in a case-control study. Given the uncertainty of the actual prevalence of the disease we decided to explore this using GCTA by varying the prevalence from 0.01 to 0.6. As show in **Table S3**, in all cases heritability was found notably greater than zero and it went down to 0.32 with prevalence equal to 0.01.

Using BayesB the fraction of markers contributing to the genetic variance was estimated as 0.015 (s.e. 0.011), however experience with such methods suggests that this fraction is sensitive to the distribution of allele effects that is assumed (results not shown). The inclusion of an additional MDS dimension (C3) did not change the results compared with Model 3.

Prediction of the phenotype

Cross-validation was used to test the predictive potential of genomic evaluation. Five cross-validation sets (denoted A–E) were produced at random from the full dataset to estimate the predictive benefit when new individuals, which have not been used to estimate the effects of markers and covariates, are genotyped in order to predict their phenotypes. Two approaches to assess the predictive value were adopted: the accuracy to predict the phenotype and the capability to diagnose individuals from genomic information.

Table 1. Summary results from the BayesB and GCTA analyses.

	Model 1	Model 2	Model 3	Model 4
BayesB				
Posterior 1–π (%)	1.65	1.57	1.54	1.57
h^2	0.61	0.63	0.64	0.65
GCTA				
h^2 (s.e.)	0.53 (0.18)	0.55 (0.18)	0.58 (0.17)	0.59 (0.17)

The estimates for the percentage of markers affecting the phenotype (1–π) and its heritability (h^2) are shown for the different statistical models: Model 1 included no covariates; Model 2 included the first two dimensions of the MDS analysis; Model 3 included the first two dimensions of the MDS analysis plus the lifestyle; Model 4 included an additional dimension of the MDS analysis to Model 3.

doi:10.1371/journal.pone.0035349.t001

Accuracy. The correlation between predicted fitted values for the new individuals and their known actual phenotype was calculated as a measure of accuracy (r) for predicting the phenotype. The Model 1 results suggest that the combined SNP effects predict the phenotype with an accuracy of 0.18 and that, by comparison with Model 2, little accuracy is added by including covariates correcting for genetic stratification (**Table 2** and **Table 3**). Including lifestyle, which was identified as a risk factor in previous analyses, improved the accuracy to 0.29 (**Table 4**). Still, the key question is whether the genomic data adds accuracy and this was assessed in different ways.

Firstly, cross-validation was performed on permuted data prior to the running of the BayesB analyses, where genotypes were randomized with respect to both phenotypes and covariates, whilst the link between phenotypes and covariates was maintained. In general, accuracy values were notably lower with permuted data than with the actual data, regardless of which of Models 1 to 3 were fitted. Within-set accuracies from permuted data were very close to zero when no covariates (Model 1) and genetic stratification (Model 2) were included. Statistical significance was observed only when lifestyle was included (Model 3), which confirms the earlier result that lifestyle has predictive value.

Secondly, to test the contribution of the genomic data, the predictions obtained from the BayesB analysis were decomposed into the component from the covariates and the component from the SNPs. The SNP component was then permuted within the cross-validation set as described in the **Materials and Methods**, but maintaining the link between the predictor from covariates and the phenotypes. For each permutation the accuracy of prediction was calculated. **Tables 2, 3, 4** show that the accuracy from the observed data with the true link between phenotypes and genotypes was in the upper tail of the distribution of accuracies (P$<$0.05). Collectively this demonstrates the Models have significant predictive value and that, within the predictor, the genomic data makes a significant contribution to the accuracy.

Finally, the magnitude of the benefit from the genomic data was assessed by predictions that excluded all genomic data. Overall accuracy obtained by C1 and C2 as explanatory factors alone was not significant (**Table 3**). Accuracy using only covariates in Model 3 was significant although the accuracy achieved was only half the value obtained from the full Model (**Table 4**). Altogether, these three ways to look at the data proved that prediction of the phenotype was more accurate when genetic markers were included.

Nevertheless, as may be expected from the relatively small data sets, there is considerable variation among the cross validation sets, and confidence intervals within individual cross-validation sets are large. Predictive accuracies were significant in sets C and D, but were not significant in sets A, B and E (**Table 2**), coinciding with a slightly higher posterior fraction of markers with a non-zero (1–π) effect for sets A, B and E than for C and D (data not shown). Overall, there was an improvement in prediction by using SNPs.

Prediction of the trait. Our second approach to assess the capability of our data to be used for prognosis of disease development required individuals to be classified as either healthy infected or affected for increasing thresholds of fitted values. Note that the phenotype was defined as one or two for

Table 2. Summary of cross-validation results after constructing five sets (labelled A–E), showing the predictive accuracy when the set is excluded from the training set for Model 1.

Model 1						
Set	**A**	**B**	**C**	**D**	**E**	**A–E**
$N_{training}$	175	175	177	176	173	
N_{cases}	21	21	20	20	22	104
Full model						
Accuracy (*r*)	0.02	0.09	0.41	0.49	0.07	0.18
(95% CI)	(−0.28, 0.32)	(−0.21, 0.38)	(0.13, 0.64)	(0.22, 0.69)	(−0.23, 0.35)	(0.05, 0.30)
Empirical significance	0.42	0.34	<0.01	<0.01	0.44	<0.01
Permuted genotypes						
Accuracy (r)	−0.11	−0.05	−0.17	−0.23	−0.13	−0.14
(95% CI)	(−0.39, 0.19)	(−0.34, 0.25)	(−0.45, 0.14)	(−0.49, 0.08)	(−0.41, 0.16)	(−0.27, −0.01)

Empirical significance was obtained from the fraction of permutations that showed a correlation higher than in the real data.
doi:10.1371/journal.pone.0035349.t002

healthy infected and affected, respectively, and therefore fitted values were approximately in this range. Receiver operating characteristic (ROC) curves were generated from sensitivity and specificity values for different thresholds and the area under the curve (AUC) was calculated as an indicative of the balance between sensitivity and specificity. AUC values were notably higher than randomness and Model 3 achieved the best performance (**Figure 1**). Regardless of the model, a threshold of 1.5 to diagnose individuals would reach the highest fraction of correct predictions (g), notably higher than the expected by chance alone (for Model 3, $g_{1.5} = 0.63$; $g_{95\%\text{limit}} = 0.55$) (**Figure 2**).

Discussion

In this study we have explored the contribution of genetic loci in the dog genome for determining clinical progression of disease following *Leishmania* infection and how such information may be used to predict disease course. Our first analysis was focused on identifying individual loci in the canine genome which contributed medium to large effects for determining disease development. Different analyses associated CFA 4: 61–77 Mb. The strongest association in the GWAS analysis was for markers in this region, even when we considered confounding factors such as lifestyle and genetic stratification, whose causes are discussed below. However, these associations were not significant when corrected for multiple testing (**Figure S1**, **Table S1**). The lack of genome-wide significance at the individual SNP level may indicate that our study was underpowered for GWAS due to the small sample size of our study. However the size of the study was at the lower end of the range of 100–300 cases and 100–300 controls that has been suggested for GWAS in dogs in complex diseases [33]. The lack of genome-wide significance may also be evidence of a complex genetic nature for leishmaniasis. This provides justification for the genomic selection approach which is more suited to prediction of complex traits (e.g. [47]).

Table 3. Summary of cross-validation results after constructing five sets (labelled A–E), showing the predictive accuracy when the set is excluded from the training set for Model 2.

Model 2						
Set	**A**	**B**	**C**	**D**	**E**	**A–E**
$N_{training}$	175	175	177	176	173	
N_{cases}	21	21	20	20	22	104
Full model						
Accuracy (*r*)	0.05	0.05	0.41	0.53	0.12	0.20
(95% CI)	(−0.26, 0.34)	(−0.25, 0.34)	(0.12, 0.64)	(0.27, 0.71)	(−0.18, 0.39)	(0.07, 0.32)
Empirical significance	0.37	0.27	0.02	0.03	0.34	0.04
Permuted genotypes						
Accuracy (r)	−0.03	−0.09	0.09	0.11	0.07	0.02
(95% CI)	(−0.32, 0.27)	(−0.38, 0.21)	(−0.22, 0.38)	(−0.20, 0.40)	(−0.23, 0.35)	(−0.11, 0.15)
Covariates alone						
Accuracy (r)	0.003	−0.06	0.23	0.43	0.17	0.11
(95% CI)	(−0.29, 0.30)	(−0.35, 0.24)	(−0.08, 0.50)	(0.15, 0.65)	(−0.13, 0.43)	(−0.02, 0.24)

Empirical significance was obtained from the fraction of permutations that showed a correlation higher than in the real data.
doi:10.1371/journal.pone.0035349.t003

Table 4. Summary of cross-validation results after constructing five sets (labelled A–E), showing the predictive accuracy when the set is excluded from the training set for Model 3.

Model 3						
Set	**A**	**B**	**C**	**D**	**E**	**A–E**
$N_{training}$	175	175	177	176	173	
N_{cases}	21	21	20	20	22	104
Full model						
Accuracy (*r*)	0.10	0.14	0.46	0.56	0.23	0.29
(95% CI)	(−0.20, 0.38)	(−0.16, 0.42)	(0.18, 0.67)	(0.32, 0.74)	(−0.06, 0.49)	(0.16, 0.41)
Empirical significance	0.48	0.24	0.02	0.01	0.46	0.03
Permuted genotypes						
Accuracy (r)	0.09	−0.01	0.28	0.29	0.26	0.15
(95% CI)	(−0.22, 0.37)	(−0.31, 0.29)	(−0.03, 0.54)	(−0.01, 0.54)	(−0.03, 0.51)	(0.02, 0.28)
Covariates alone						
Accuracy (r)	0.11	0.03	0.35	0.43	0.34	0.22
(95% CI)	(−0.19, 0.40)	(−0.27, 0.32)	(0.05, 0.59)	(0.15, 0.65)	(0.05, 0.57)	(0.09, 0.35)

Empirical significance was obtained from the fraction of permutations that showed a correlation higher than in the real data.
doi:10.1371/journal.pone.0035349.t004

Interestingly, when we tested for association focusing only on SNPs residing within candidate loci related to host response to *Leishmania* and susceptibility to leishmaniasis in humans and mice [50,55,56], loci on chromosome 4: 61–77 Mb were significant after correcting for multiple testing and linkage disequilibrium (**Dataset S1** and **Figure S3**).

In addition, from the BayesB analysis, markers in this region of CFA 4 had a larger estimated effect on the phenotype than other genome-wide markers (**Figure S4**). Chromosome 4: 61–77 Mb is syntenic to a locus that mediates host response to *Leishmania major* in mice, which includes the candidate genes *Il7r*, *Lifr*, *C6* and *C7* [50]. *Il7r* (CFA 4: 75.8 Mb) is of special interest as, although healthy infected and affected samples showed similar MAF and observed heterozygosity values along CFA 4: 61–77 Mb, in both groups three SNPs (CFA 4: 75.7–75.9 Mb) flanking *Il7r* significantly deviated from HWE (p-value$<10^{-5}$) (data not shown). Extended patterns of markers deviating from HWE may indicate copy number variants. Variation in the number of copies between affected and healthy infected cannot be detected through differences in genotype frequencies though it might affect the phenotype. In fact, structural variations have been described for the syntenic region in the human genome [57,55,58], which encompasses *SPEF2*, *CAPSL*, *UGT3A1* and *UGT3A2* in addition to

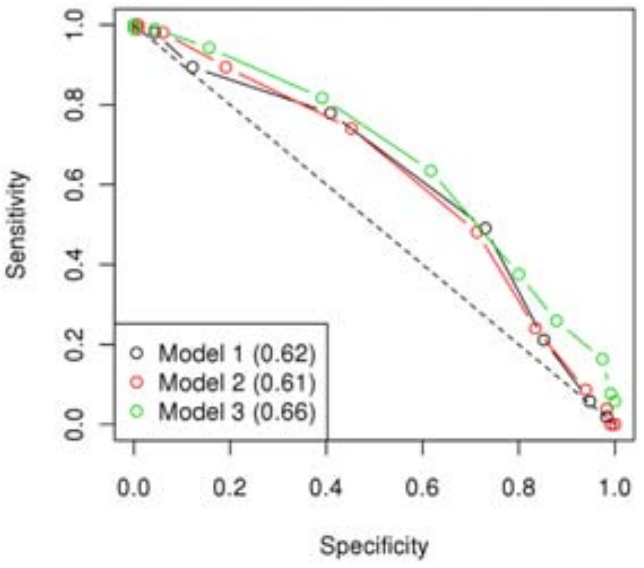


Figure 1. Receiver Operating Characteristic (ROC) curves. Sensitivity and specificity values were obtained for increasing classification thresholds to produce the ROC curves. In the legend, the values for the area under the ROC curve (AUC) are indicated in parenthesis for each model. AUC can range between 0.5 (randomness, dashed line) and 1.0 (ideally).
doi:10.1371/journal.pone.0035349.g001

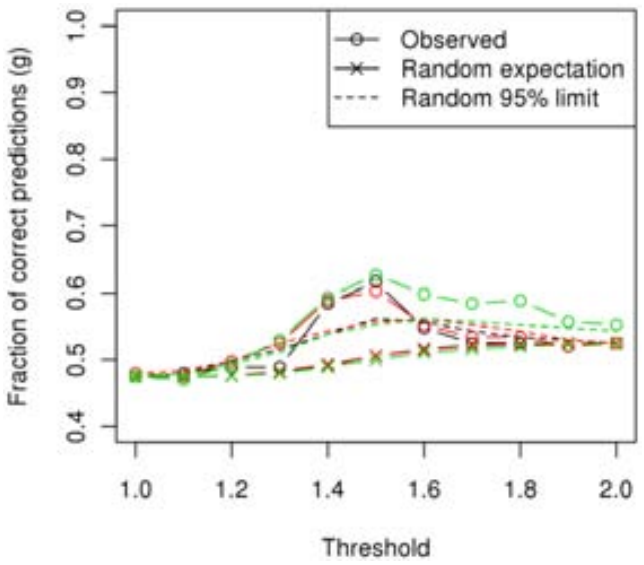


Figure 2. Fraction of correct predictions. For increasing classification thresholds percentages of correct classifications were compared to those expected by chance. Calculations for the random expectation and the random 95% limit were drawn from a hypergeometric distribution and are detailed in **Text S1**.
doi:10.1371/journal.pone.0035349.g002

IL7R. In mice, structural variation has also been reported for a shorter region overlapping *Ugt3a1* [59]. However, replication in an independent sample is needed to confirm the association on chromosome 4, as well as those on chromosomes 1 and 20, and the identification of these regions only represents a first discovery step for a better understanding of the genetic variants that control genetic susceptibility to clinical progression of leishmaniasis from *Leishmania* infection.

Next, we studied the extent to which the additive effects of loci throughout the genome determine the disease development following *Leishmania* infection. Our data suggest that the trait is complex with many different gene segments contributing to the phenotype and that the genetic variance may explain as much as 60% of the total observed phenotypic variance. Whilst this estimate was fairly consistent across the different methodologies used for its estimation (**Table 1**), the estimation is made more complex and very likely to be biased upward, by the case-control nature of the data. This is the first clear evidence that there is a significant genetic component to leishmaniasis in dogs within breeds. In addition, it is the first heritability estimate for progression of clinical leishmaniasis from *Leishmania* infection in any species, although an estimate of heritability for a marker of healed *Leishmania* infection and protection against subsequent reinfection in humans has been reported [19].

An important finding of this study was that whilst no single SNP was found to be reliably predictive, there was significant predictive value of the genomic data through using the genomic evaluation as proposed by Meuwissen *et al.* [47]. The best predictor included information on lifestyle as well as the genomic predictor, but it was clearly established that the genomics made a substantial contribution to the accuracy. The model including the covariates and the genomic data reached an accuracy of 0.29 for a dog that was outside the current dataset (e.g. a newborn dog), and thus is only weakly predictive of the phenotype. However three points should be remembered. Firstly this accuracy was achieved using 80% of the data (the other 20% were used for cross-validation), and that the total data consisted of only 219 infected animals, of which only 104 had developed the disease. Secondly, this accuracy is the prediction of a phenotype and not the underlying genetic liability, and the accuracy of predicting the genetic liability is likely to be greater. In random sample with continuous traits the accuracy would be scaled by $1/h$ (>1) where h is the square root of the heritability. The structure of the data prevents us from proposing any correction. Thirdly the value of using genomics is that the genomic data can be accumulated over time with increasing accuracy of prediction. One might anticipate that further collection of cases and controls would increase accuracy to levels that have the potential for making a clinical impact on breeding for resistance away from the development of pathology, i.e. toleration of the parasite.

Finally, we would like to comment on the possible causes of the genetic stratification seen in our dataset, which especially affected the GWAS results and could only be reduced to $\lambda = 1.17$, and to compare with other GWAS in dogs. Roughly half of the dozen published GWAS in dogs provided information with regard to stratification. Three studies [37,45,43] observed good clustering of cases and controls when plotting the first two MDS dimensions, in spite of different geographical origin of the samples in the study from Madsen et al.. Barber et al. [42] also used MDS in order to detect stratification and excluded a good number of outlier samples. Wilbe et al. [34] and Downs et al. [60] reported inflation factor values, before correction, of 1.3 and 1.4, respectively. Both studies observed clustering of either samples with similar geographic provenance [34] or known to be related [60] and performed a Cochran-Mantel-Haenszel stratified analysis within the clusters as a measure of correction. However, no value of the inflation factor after the correction was presented. Only Olsson and collegues reported an inflation factor of 1.2 after removing two outliers in the MDS plot [35]. We consider that it is unlikely that our lambda value was inflated due to population stratification because we neither observed geographical clustering of samples within Spain (the majority of the samples were collected from different areas in the country) nor differentiation of samples collected in other countries (i.e. Italy, Greece and Portugal). It is reasonable to think that geographical stratification would have been noticed if present, as it has happened with some other canine GWAS. Although population or geographical stratification is a common cause of increased inflation factor, there are other confounder effects that can produce the same results [61]. We tried to avoid differential bias by following the same procedures in the collection of samples and clustering of samples that went through different DNA extraction protocols or genotyping batches was ruled out. Although we tried to avoid family structure by not including members of the same family, cryptic relatedness might have certainly inflated the lambda value. Nonetheless, we note that lambda values >1.05 are typically considered to denote stratification in human studies [61]. Although this is a statistical rule-of-thumb and it should be the same regardless of the species, we wonder if certain relatedness owing to founder effects, inbreeding, popular sire effects and repeated mating might be inherent to GWAS in dogs in spite of a careful study design.

Materials and Methods

Ethics statement

The dogs in the study were examined during routinary veterinary procedures by the veterinary clinics participating in the study. All samples were collected for routine diagnostic and clinical purposes. The samples were obtained during veterinary procedures that would have been carried out anyway and DNA was extracted from residual surplus of samples and used in the study with verbal owner consent. This is a very special situation in veterinary medicine. As the data are from client-owned dogs that underwent normal veterinary exams, there was no ''animal experiment'' according to the legal definitions in Spain and the United Kingdom, and approval by an ethical committee was not necessary.

Study population and epidemiology

The study population consisted of a single breed of dogs (Boxer). This design was chosen as the use of a single breed for the first stage of GWAS in dogs will increase power by reducing effects of genetic differentiation between breeds and increasing the degree of linkage disequilibrium [33]. Moreover, this breed appears more predisposed to overt CanL than others [23,62,63] . An age criterion for study inclusion was applied (see below) as it has been reported that age distribution of the prevalence of infection follows a bimodal pattern, with the first peak including dogs diagnosed at 2 to 4 years of age and the second peak including dogs about 7 years old [64]. The study was carried out in collaboration with the Hospital Clínic Veterinari of the Universitat Autònoma de Barcelona (HCV-UAB) and therefore most Boxers included were from the metropolitan area of Barcelona, Spain, where the HCV-UAB is located. A number of Boxers from other areas where the disease is endemic were also included (Spain, Greece, Italy and Portugal).

Veterinary clinics recruitment and samples collection

Veterinary clinics and dog owners were encouraged to participate in the study, in the case of the latter through their veterinary centre. Two millilitres of EDTA peripheral blood and 2 ml of serum were required. In addition, other tissues (*e.g.* bone marrow) as well as conjunctiva or lesion swabs were received occasionally. With regard to CanL, no pre-screening of the samples sent to the laboratory was done by the veterinary clinics. Dogs affected by CanL included those with a documented history of disease, undergoing a relapse or newly diagnosed. For those showing a disease episode when samples were collected, additional samples were requested one month after treatment or if requested by the referring clinician in order to both confirm the diagnosis and help inform then clinician on treatment response. For healthy infected dogs (see below), additional samples and medical information were requested to confirm the absence of CanL development. Inclusion of additional samples relied on the collaboration of veterinary clinics.

Phenotype definition, clinical classification and laboratory tests

Dogs were classified by clinical signs, clinical biochemistry, direct parasite detection and anti-*Leishmania* immune reactions into the following groups: (i) healthy infected: healthy and >4 years old but with evidence of prior infection and (ii) affected: manifest clinical disease and diagnosed before the age of 4 years. Ages were recorded as age-of-onset of disease for affected and current age at sample collection for healthy infected dogs. Lifestyle (*i.e.* living indoors, outdoors, both or undetermined), gender, level of relatedness and geographic location of origin were also collected. *Leishmania* quantitative polymerase chain reaction (qPCR) and anti-*Leishmania* Enzyme-Linked ImmunoSorbent Assay (ELISA) were performed on all samples from all dogs, although additional results from direct parasite detection and anti-*Leishmania* immune tests were provided by veterinary clinics for most samples. *Leishmania* qPCR was performed at the Servei Veterinari de Genètica Molecular, UAB, as described previously [62] and ELISA was performed at UNIVET Servicio de Diagnóstico Veterinario SL, UAB, as described elsewhere [63].

DNA extraction, SNP genotyping and data quality control

DNA was extracted from peripheral blood and bone marrow samples using either QIAamp® DNA Blood Mini Kit (QIAGEN) or by conventional phenol-chloroform DNA extraction and deproteinization methods. All samples were genotyped using the Illumina CanineHD BeadChip (174,376 markers) [65] at the Centre National de Génotypage, France. Data cleaning was conducted using PLINK [48] and R version 2.13.0 [66] packages. Quality control was performed independently on two genotyping batches. In total eight samples with call rate <90% were excluded. Intensity probes were excluded together with markers on the boundary autosomal region on the CFA X and SNPs on the non-pseudoautosomal region on CFA X for which heterozygous genotypes in male samples were observed. Markers with call rate <90% were also excluded. Multidimensional scaling (MDS) analysis based on the genotypes was performed to detect samples with a very different genetic content (explained below). Three affected dogs were excluded because they appeared as outliers when the first two dimensions from the MDS analysis were plot (data not shown). After data cleaning 115 healthy infected and 104 affected dogs remained. In addition, markers were filtered to have a minor allele frequency (MAF) >1.5% and a Hardy-Weinberg Equilibrium (HWE) test p-value>0.005 (a threshold set based on the empirical distribution of our data). This left 126,607 markers for analysis. Finally, for logistic regression and BayesB analyses, one SNP of a pair was removed for those SNP pairs showing complete genotypic correlation, resulting in 99,997 SNPs left for analysis.

Statistical analysis

Covariates. Two confounding effects were considered and fitted into the statistical analyses: genetic stratification and dog lifestyle. Using PLINK [48], an identical-by-state correlation matrix for *n* individuals was calculated from which *n* dimensions were extracted using MDS analysis, resulting in a matrix of *n*-samples by *n*-dimensions eigenvalues. The fraction of genetic variance explained by each dimension was calculated as the variance for a given dimension along all samples divided by the sum of variances for all the dimensions extracted. The eigenvalues for the first two dimensions (C1 and C2) of the MDS analysis were used as continuous covariates. For simplicity, only C1 and C2 were used because the fraction of additional genetic variance explained by each of the subsequent 217 MDS dimensions extracted was minimal (**Figure S2**). Although healthy infected and affected samples generally clustered together in the MDS plot (**Figure S2**), genetic stratification was observed in our cleaned dataset based on the genomic inflation factor ($\lambda = 1.29$), with C1 capturing twice the stratification captured by C2. Therefore, C1 and C2 values for each sample were fitted as continuous covariates in the indicated models. CanL is known to be a complex disease with an environmental component and thus dogs living outdoors, more exposed to infection, are believed to more frequently develop the disease. Hence, lifestyle was also included as a factor in the analyses for the models indicated. For Model 3, the inflation factor was reduced to $\lambda = 1.17$. In the logistic regression, lifestyle was fitted as a factor (one degree of freedom) using dummy variables for indoors, indoors/outdoors, outdoors and 'undetermined'. In the genomic selection analyses, lifestyle was considered as a categorical covariate with four levels. Three genetic models varying in whether covariates were fitted were defined to explain the phenotype (y), treated as binary (i.e. either healthy infected or affected):

Model 1: y~SNPs
Model 2: y~SNPs+C1+C2
Model 3: y~SNPs+C1+C2+lifestyle

GWAS and candidate genes analysis

Markers were tested for association using the Cochran-Armitage for trend test (Model 1) and logistic regression (Models 2 and 3). Genome-wide significance (P_{genome}) was obtained after 10,000 permutations. Based on the permutations carried out in our dataset, the uncorrected p-value that would reach genome-wide significance (at the 5% level) after correction for multiple testing in our study would be $p = 2.08 \times 10^{-6}$. Formally, first, for each permutation the maximum statistic across all SNPs was recorded and, second, from this distribution of maximum statistics, the statistic in the top 5% is used to give the $p = 2.08 \times 10^{-6}$ that would be significant after permutations.

Candidate loci reported as related to host response to *Leishmania* and susceptibility to leishmaniasis in *Mus musculus* and *Homo sapiens* [56,67] were used to retrieve homologous loci in *C. familiaris* using Biomart [68] with CanFam 2.0. Sets were defined with SNPs in the Illumina's CanineHD Beadchip residing within the retrieved candidate loci and their flanking regions (±1 Mb). Loci for which at least one SNP overlapped were merged into the same set (**Dataset 1**). Set-based association tests were performed as described in PLINK [53] with two different sets of parameters:

(1) $r^2 = 0.80$, p-value = 0.05; and (2) $r^2 = 0.10$, p-value = 0.01. In both cases, a maximum set size of 10 SNPs (~10% of the median set size) was used. The Cochran-Armitage for trend test was used and 10,000 within-set permutations were conducted to obtain empirical set-based p-values (EMP1).

Modified BayesB method. Datasets were analysed with the BayesB method [47] with some modifications previously published [53] assuming Models 1–3. The phenotype was treated as continuous. A flat prior distribution for the proportion of markers with non-zero additive effect ($1–\pi$) was set to follow a beta distribution with parameters $\alpha = 1$ and $\beta = 1$ and a starting value of 0.2. An informative distribution for the variance for the additive parameter was set to follow an inverse chi-squared distribution with two degrees of freedom ($v = 2$) and a scale parameter (S) of 0.001 (weak prior). The Markov chain Monte Carlo (MCMC) of BayesB was run for 160,000 cycles and the first 10,000 cycles were discarded as burn in; 3,000 realisations of sampling were performed with 50 cycles between realisations. Absolute variation between each of the 3,000 sampled values of posterior S and their prior S (either $S = 0.001$ or $S = 0.1$) was calculated and a Welch two sample t-test was applied. The same test was applied to the sampled values of ($1–\pi$) produced by each of weak and stronger S priors. BayesB produced estimates of the genomic breeding values (GEBV) and of the effects of C1, C2 and each lifestyle category, which were used to calculate fitted values of the phenotype ($\hat{y}$) according to the different predictive models (**Text S1**).

In order to assess the effect of the weak informative prior distribution used for the variance of the additive parameter on the resulting posterior value, for Model 1 the analysis was repeated with $S = 0.1$ (strong prior), which was 100 times higher than the value used otherwise. In absolute values, the posterior S value changed a 58.7% respect to the weak prior ($S = 0.001$) whereas this variation was significantly greater, 78.9%, for $S = 0.1$ (p-value = 7×10^{-6}). Another effect of giving a stronger prior S was that the posterior proportion of markers with a non-zero effect ($1–\pi$) was 100 times lower compared to the obtained for the weak prior (0.02% and 1.65%, respectively, p-value = 2×10^{-16}) and the fraction of genetic variance was also lower (0.27 and 0.61, respectively). Moreover, the genome-wide pattern of estimated SNP effects was notably different depending on the prior given. With a weak prior most markers had non-zero but very low estimated effect (10%-quantile = 10^{-5}) whereas a small fraction of SNPs (10%-quantile = 0) had 10-fold estimated effects with a stronger prior. Altogether these results can be explained by a scenario in which fewer SNPs with greater effect contribute to the phenotype when a greater prior variance of the additive parameter is allowed.

Restricted Maximum Likelihood (REML) analysis (GCTA software). When calculating the genetic relationship matrix (GRM) with GCTA [54], no adjustment was specified to correct for imperfect LD between genotyped markers and causal loci. REML analyses were run assuming Models 1–3. As input parameters, genetic and environmental variances were not specified and default values of 0.12 for both were used. Model 3 was run with varying phenotype prevalence values from 0.01 to 0.60 in order to explore the sensitivity of estimates to prevalence.

Cross-validation. Samples were assigned randomly to one of five training sets (denoted A–E) so that (i) each training set had a size of approximately 4/5 of the full unpermuted dataset and (ii) in each training set the proportions of samples belonging to each phenotype (either affected or healthy infected) and lifestyle categories were approximately as in the full dataset. Samples not included in each training set were used as testing data. For each training set, BayesB was run to both estimate E_{C1}, E_{C2}, E_{life} from samples in the training set and produce GEBV for the testing data samples and then calculate their fitted values, according to the corresponding predictive model.

Accuracy (r) was calculated as the correlation between fitted values ($\hat{y}$) and true phenotypes (y). In each testing set, r was calculated as a measure of accuracy to predict the phenotype. The overall correlation for the full unpermuted dataset was calculated by combining the predictions across sets. The contribution of markers to the accuracy was analysed in three ways. First, GEBV generated with BayesB were permuted before the calculation of fitted values. In this way, the correspondence between phenotypes and covariates was not altered. Within each model, 100 sets of permuted GEBV, resulting in 100 sets of permuted fitted values, were generated for each set. The empirical p-value for the real data was computed as the fraction of permuted sets with a lower p-value that the real data. Second, genotypes were randomized respective to phenotypes and covariates, which were kept as in the original data. The BayesB analysis was then run and cross-validation applied as explained before. Third, fitted values were calculated using uniquely covariates, i.e. GEBV were not used.

For Models 1–3, receiver operating characteristic (ROC) curves were calculated as follows. A fitted value threshold was set so that below or above it individuals were predicted to be healthy infected or affected, respectively. Specificity, sensitivity and fraction of correct predictions (g) values were calculated (**Text S1**) for increasing thresholds of fitted values and ROC curves were generated by plotting specificity against sensitivity. The area under the ROC curve (AUC) was calculated as a measure of similarity between specificity and sensitivity.

Supporting Information

Figure S1 Single-marker genome-wide association plot for Model 1 after 10,000 permutations with the strongest associations indicated.
(TIFF)

Figure S2 Genetic stratification. (**A**) relative genetic variance explained by the 219 MDS dimensions extracted; (**B**) MDS plot for the first two MDS dimensions (C1 and C2) with healthy infected and affected samples coloured differently. The percentage of relative genetic variance explained by each dimension is indicated as well as the genomic inflation factor (lambda).
(TIFF)

Figure S3 Distribution of EMP1 across SNP sets of candidate regions. Sets comprise SNPs in the following regions: 6 (CFA 4:61.2–63.2 Mb), 7 (CFA 4: 70.5–74.5 Mb), 8 (CFA 4: 74.8–76.9 Mb), 19 (CFA 9: 40.0–46.5 Mb) and 22 (CFA 10: 29.6–31.5 Mb).
(TIFF)

Figure S4 Genome-wide plot of the absolute mean SNP effects estimated with BayesB for Model 1 (A), Model 2 (B) and Model 3 (C). The peak on CFA 4: 61–77 Mb (red segment) consistent across Models 1–3 coincided with both the strongest association in GWAS analysis and the region in which SNP sets covering candidate genes were significant (EMP1<0.01).
(TIFF)

Table S1 Strongest associations from each region identified in the GWAS analysis. BICF2P1345879 was not used in models 2 and 3 because, for logistic regression, SNPs were pruned based on LD (see **Materials and Methods**).The closest marker, <6 Kb upstream, was BICF2P813758 at 20:30,126,633 bp (Model 2: $P_{raw} = 4.4 \times 10^{-4}$, $P_{genome} > 0.50$, OR = 0.33; Model 3: $P_{raw} = 6.2 \times 10^{-4}$, $P_{genome} > 0.50$, OR = 0.34). Choice of

SNPs representing each genomic region was based on the strongest associations in Model 1. *Canis familiaris* genes (CanFam_2.0) were retrieved using Biomart and associated gene names are given, with the exception of some for which no gene name was available and the Ensembl ID is given instead. The same information is presented for the strongest associations on chromosomes 9 and 10 from the set-based analysis.
(XLS)

Table S2 Genomic inflation (λ) was not affected by fitting additional MDS dimensions as covariates of the model.
(DOC)

Table S3 Sensitivity of heritability (h^2) estimation using GCTA to prevalence of the phenotype is shown for Model 3.
(DOC)

Text S1 Fitted values, fraction of correct predictions, sensitivity and specificity calculation.
(DOC)

Dataset S1 Candidate genes analysis: (A) candidate genes and loci described in *H. sapiens* and *M. musculus* and retrieved genomic positions in *C. familiaris*; (B) sets of non-overlapping candidate regions plus their ±1 Mb-flanking regions; (C) results from the set-based association study.
(XLS)

Acknowledgments

We would like to thank the referring clinicians, dog owners who gave permission for their dogs to participate in this study.

Author Contributions

Conceived and designed the experiments: AS LJK RJQ WERO XR LF LA OF. Performed the experiments: JQ VM LA. Analyzed the data: JQ JAW RP. Contributed reagents/materials/analysis tools: JAW AS RP. Wrote the paper: JQ. Revised the manuscript: VM JAW RJQ WERO XR LF LA OF.

References

1. Blackwell JM (1996) Genetic susceptibility to leishmanial infections: studies in mice and man. Parasitology 112(Supplement S1): S67.
2. Blackwell JM, Fakiola M, Ibrahim ME, Jamieson SE, Jeronimo SB, et al. (2009) Genetics and visceral leishmaniasis: of mice and man. Parasite Immunol 31(5): 254–266.
3. Ibrahim M, Lambson B, Yousif A, Deifalla N, Alnaiem D, et al. (1999) Kala-azar in a high transmission focus: an ethnic and geographic dimension. Am J Trop Med Hyg 61(6): 941–944.
4. Cabello PH, Lima AMVMD, Azevedo ES, Krieger H (1995) Familial Aggregation of Leishmania chagasi Infection in Northeastern Brazil. Am J Trop Med Hyg 52(4): 364–365.
5. Zijlstra EE, El-Hassan AM, Ismael A, Ghalib HW (1994) Endemic Kala-Azar in Eastern Sudan: A Longitudinal Study on the Incidence of Clinical and Subclinical Infection and Post-Kala-Azar Dermal Leishmaniasis,. Am J Trop Med Hyg 51,(6,): 826–836.
6. Peacock CS, Collins A, Shaw MA, Silveira F, Costa J, et al. (2001) Genetic epidemiology of visceral leishmaniasis in northeastern Brazil. Genet Epidemiol 20(3): 383–396.
7. Bucheton B, Abel L, El-Safi S, Kheir MM, Pavek S, et al. (2003) A major susceptibility locus on chromosome 22q12 plays a critical role in the control of kala-azar. Genes Immun 73(5): 1052–1060.
8. Mohamed HS, Ibrahim ME, Miller EN, White JK, Cordell HJ, et al. (2004) SLC11A1 (formerly NRAMP1) and susceptibility to visceral leishmaniasis in The Sudan. Eur J Hum Genet 12(1): 66–74.
9. Faghiri Z TS (1995) Study of the Association of HLA Class I Antigens with Kala-Azar. Hum Hered 45: 258–261.
10. Meddeb-Garnaoui A, Gritli S, Garbouj S, Ben Fadhel M, El Kares R, et al. (2001) Association analysis of HLA-class II and class III gene polymorphisms in the susceptibility to mediterranean visceral leishmaniasis. Hum Immunol 62(5): 509–517.
11. Salih MA, Ibrahim ME, Blackwell JM, Miller EN, Khalil EAG, et al. (2007) IFNG and IFNGR1 gene polymorphisms and susceptibility to post-kala-azar dermal leishmaniasis in Sudan. Genes Immun 8(1): 75–78.
12. Peacock CS, Sanjeevi CB, Shaw MA, Collins A, Campbell RD, et al. (2002) Genetic analysis of multicase families of visceral leishmaniasis in northeastern Brazil: no major role for class II or class III regions of HLA. Genes Immun 3(6): 350–358.
13. Miller E, Fadl M, Mohamed H, Elzein A, Jamieson S, et al. (2007) Y Chromosome Lineage- and Village-Specific Genes on Chromosomes 1p22 and 6q27 Control Visceral Leishmaniasis in Sudan. PLoS Genet 3(5.
14. Jamieson SE, Miller EN, Peacock CS, Fakiola M, Wilson ME, et al. (2006) Genome-wide scan for visceral leishmaniasis susceptibility genes in Brazil. Genes Immun 8(1): 90–90.
15. Jeronimo SM, Duggal P, Braz RF, Cheng C, Monteiro GR, et al. (2004) An emerging peri-urban pattern of infection with Leishmania chagasi, the protozoan causing visceral leishmaniasis in northeast Brazil. Scand J Infect Dis 36(6–7): 443–449.
16. Bucheton B, Kheir MM, El-Safi SH, Hammad A, Mergani A, et al. (2002) The interplay between environmental and host factors during an outbreak of visceral leishmaniasis in eastern Sudan. Microb Infect 4(14): 1449–1457.
17. Karplus TM, Jeronimo SMB, Chang H, Helms BK, Burns TL, et al. (2002) Association between the Tumor Necrosis Factor Locus and the Clinical Outcome of Leishmania chagasi Infection. Infect Immun 70(12): 6919–6925.
18. Jeronimo SMB, Holst AKB, Jamieson SE, Francis R, Martins DRA, et al. (2007) Genes at human chromosome 5q31.1 regulate delayed-type hypersensitivity responses associated with Leishmania chagasi infection. Genes Immun 8(7): 551–551.
19. Jeronimo SMB, Duggal P, Ettinger NA, Nascimento ET, Monteiro GR, et al. (2007) Genetic Predisposition to Self-Curing Infection with the Protozoan Leishmania chagasi: A Genomewide Scan. Journal of Infectious Diseases 196(8): 1261–1269.
20. Solano-Gallego L, Morell P, Arboix M, Alberola J, Ferrer L (2001) Prevalence of Leishmania infantum infection in dogs living in an area of canine leishmaniasis endemicity using PCR on several tissues and serology. J Clin Microbiol 39(2): 560–563.
21. Baneth G, Aroch I (2008) Canine leishmaniasis: a diagnostic and clinical challenge. Vet J 175(1): 14–15.
22. Martínez V, Quilez J, Sanchez A, Roura X, Francino O, et al. (2011) Canine leishmaniasis: the key points for qPCR result interpretation. Parasites & vectors 4: 57.
23. (2005) Clinically patent canine leishmaniasis shows age, breed and sex predilection. World-leish3, Third World Congress on Leishmaniosis; Palermo,-Terrassini, Sicily, Italy. 2005 p.
24. Abranches P, Silva-Pereira M, Conceição-Silva F, Santos-Gomes G, Janz J (1991) Canine Leishmaniasis: Pathological and Ecological Factors Influencing Transmission of Infection. J Parasitol 77(4): 557–561.
25. Sanchez-Robert E, Altet L, Sanchez A, Francino O (2005) Polymorphism of Slc11a1 (Nramp1) gene and canine leishmaniasis in a case-control study. J Hered 96(7): 755–758.
26. Solano-Gallego L, Llull J, Ramos G, Riera C, Arboix M, et al. (2000) The Ibizian hound presents a predominantly cellular immune response against natural Leishmania infection. Vet Parasitol 90(1–2): 37–45.
27. Altet L, Francino O, Solano-Gallego L, Renier C, Sanchez A (2002) Mapping and Sequencing of the Canine NRAMP1 Gene and Identification of Mutations in Leishmaniasis-Susceptible Dogs. Infect Immun 70(6): 2763–2771.
28. Sanchez-Robert E, Altet L, Utzet-Sadurni M, Giger U, Sanchez A, et al. (2008) Slc11a1 (formerly Nramp1) and susceptibility to canine visceral leishmaniasis. Vet Res 39(3): 36.
29. Sanchez-Robert E, Altet L, Alberola J, Rodriguez-Cortés A, Ojeda A, et al. (2008) Longitudinal analysis of cytokine gene expression and parasite load in PBMC in Leishmania infantum experimentally infected dogs. Vet Immunol Immunopathol 125(1–2): 168–175.
30. Quinnell RJ, Kennedy LJ, Barnes A, Courtenay O, Dye C, et al. (2003) Susceptibility to visceral leishmaniasis in the domestic dog is associated with MHC class II polymorphism. Immunogenetics 55(1): 23–28.
31. Ostrander EA (2005) The canine genome. Genome Res 15(12): 1706–1716.
32. Neff MW, Rine J (2006) A fetching model organism. Cell 124(2): 229–231.
33. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069): 803–819.
34. Wilbe M, Jokinen P, Truvé K, Seppala EH, Karlsson EK, et al. (2010) Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. Nat Genet 42(3): 250–254.
35. Olsson M, Meadows JRS, Truvé K, Rosengren Pielberg G, Puppo F, et al. (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. PLoS Genet 7(3): e1001332.

36. Mausberg TB, Wess G, Simak J, Keller L, Drögemüller M, et al. (2011) A locus on chromosome 5 is associated with dilated cardiomyopathy in Doberman Pinschers. PloS one 6(5): e20042.
37. Tsai KL, Noorai RE, Starr-Moss AN, Quignon P, Rinz CJ, et al. (2011) Genome-wide association studies for multiple diseases of the German Shepherd Dog. Mamm Genome.
38. Goldstein O, Mezey JG, Boyko AR, Gao C, Wang W, et al. (2010) An ADAM9 mutation in canine cone-rod dystrophy 3 establishes homology with human cone-rod dystrophy 9. Mol Vis 16: 1549–1569.
39. Wood SH, Ke X, Nuttall T, McEwan N, Ollier WE, et al. (2009) Genome-wide association analysis of canine atopic dermatitis and identification of disease related SNPs. Immunogenetics 61(11–12): 765–772.
40. Dodman NH, Karlsson EK, Moon-Fanelli A, Galdzicka M, Perloski M, et al. (2010) A canine chromosome 7 locus confers compulsive disorder susceptibility. Mol Psychiatry 15(1): 8–10.
41. Meurs K, Mauceli E, Lahmers S, Acland G, White S, et al. (2010) Genome-wide association identifies a deletion in the 3′ untranslated region of striatin in a canine model of arrhythmogenic right ventricular cardiomyopathy. Hum Genet 128(3): 315–324.
42. Barber RM, Schatzberg SJ, Corneveaux JJ, Allen AN, Porter BF, et al. (2011) Identification of risk loci for necrotizing meningoencephalitis in Pug dogs. J Hered 102 Suppl 1: S40–S46.
43. Madsen MB, Olsen LH, Haggstrom J, Hoglund K, Ljungvall I, et al. (2011) Identification of 2 loci associated with development of myxomatous mitral valve disease in Cavalier King Charles Spaniels. J Hered 102 Suppl 1: S62–S67.
44. Barros Roque J, O'Leary CA, Duffy DL, Kyaw-Tanner M, Latter M, et al. (2011) IgE responsiveness to Dermatophagoides farinae in West Highland white terrier dogs is associated with region on CFA35. J Hered 102 Suppl 1: S74–S80.
45. Mogensen MS, Karlskov-Mortensen P, Proschowsky HF, Lingaas F, Lappalainen A, et al. (2011) Genome-wide association study in Dachshund: identification of a major locus affecting intervertebral disc calcification. J Hered 102 Suppl 1: S81–S86.
46. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3): 1021–1031.
47. Meuwissen T, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–575.
49. Badalova J, Svobodova M, Havelkova H, Vladimirov V, Vojtikova J, et al. (2002) Separation and mapping of multiple genes that control IgE level in Leishmania major infected mice. Genes Immun 3(4): 195–195.
50. Beebe AM, Mauze S, Schork NJ, Coffman RL (1997) Serial Backcross Mapping of Multiple Loci Associated with Resistance to Leishmania major in Mice. Genes Immun 6(5): 557–557.
51. Havelkova H, Badalova J, Svobodova M, Vojtiskova J, Kurey I, et al. (2006) Genetics of susceptibility to leishmaniasis in mice: four novel loci and functional heterogeneity of gene effects. Genes Immun 7(3): 233–233.
52. Vladimirov V, Badalová J, Svobodová M, Havelková H, Hart AAM, et al. (2003) Different genetic control of cutaneous and visceral disease after Leishmania major infection in mice. Infect Immun 71(4): 2041–2046.
53. Pong-Wong R, Hadjipavlou G (2010) A two-step approach combining the Gompertz growth model with genomic selection for longitudinal data. BMC proceedings 4 Suppl 1: S4.
54. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1): 76–82.
55. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318(5849): 420–426.
56. Sakthianandeswaren A, Foote SJ, Handman E (2009) The role of host genetics in leishmaniasis. Trends Parasitol 25(8): 383–391.
57. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Genes Immun 453(7191): 64–64.
58. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Genes Immun 451(7181): 1003–1003.
59. Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD (2007) Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Research 17(12): 1743–1754.
60. Downs LM, Wallin-Håkansson B, Boursnell M, Marklund S, Hedhammar Å, et al. (2011) A frameshift mutation in golden retriever dogs with progressive retinal atrophy endorses SLC4A3 as a candidate gene for human retinal degenerations. PloS one 6(6): e21452.
61. Price A, Zaitlen N, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7): 459–463.
62. Francino O, Altet L, Sánchez-Robert E, Rodriguez A, Solano-Gallego L, et al. (2006) Advantages of real-time PCR assay for diagnosis and monitoring of canine leishmaniosis. Genes Immun 137(3–4): 221–221.
63. Cortés E, Sanz A, Vela C, Ranz AI (2005) Leishmaniosis canina. Diagnóstico serológico de la leshmaniosis: análisis comparativo de ensayos inmunoenzimáticos e IFI. Información Veterinaria. pp 28–33.
64. Miranda S, Roura X, Picado A, Ferrer L, Ramis A (2008) Characterization of sex, age, and breed for a population of canine leishmaniosis diseased dogs. Res Vet Sci 85(1): 35–38.
65. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, et al. (2011) Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLoS Genet 7(10): e1002316.
66. R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (R version 2.13.0 (2011-04-13).
67. Lipoldova M, Demant P (2006) Genetic susceptibility to infectious disease: lessons from mouse models of leishmaniasis. Nat Rev Genet 7(4): 294–305.
68. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal—unified access to biological data. Nucleic Acids Research 37(suppl 2): W23–W27.

Figure S1

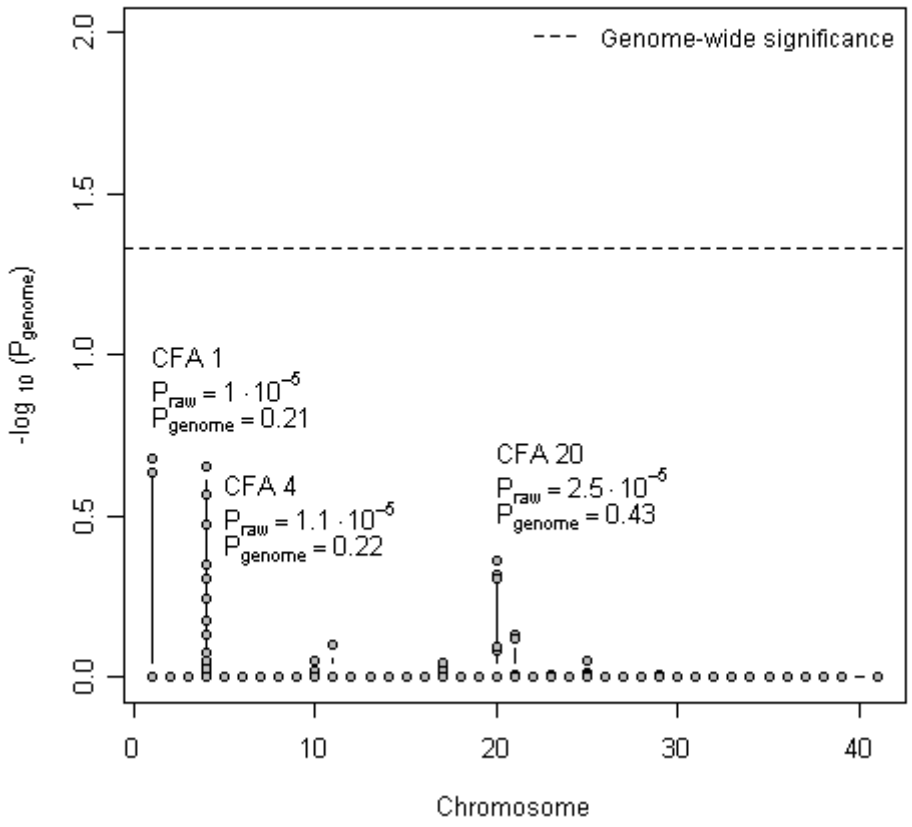

Genome-wide significance
CFA 1
$P_{raw} = 1 \cdot 10^{-5}$
$P_{genome} = 0.21$
CFA 4
$P_{raw} = 1.1 \cdot 10^{-5}$
$P_{genome} = 0.22$
CFA 20
$P_{raw} = 2.5 \cdot 10^{-5}$
$P_{genome} = 0.43$
$-\log_{10}(P_{genome})$
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Chromosome

Figure S2

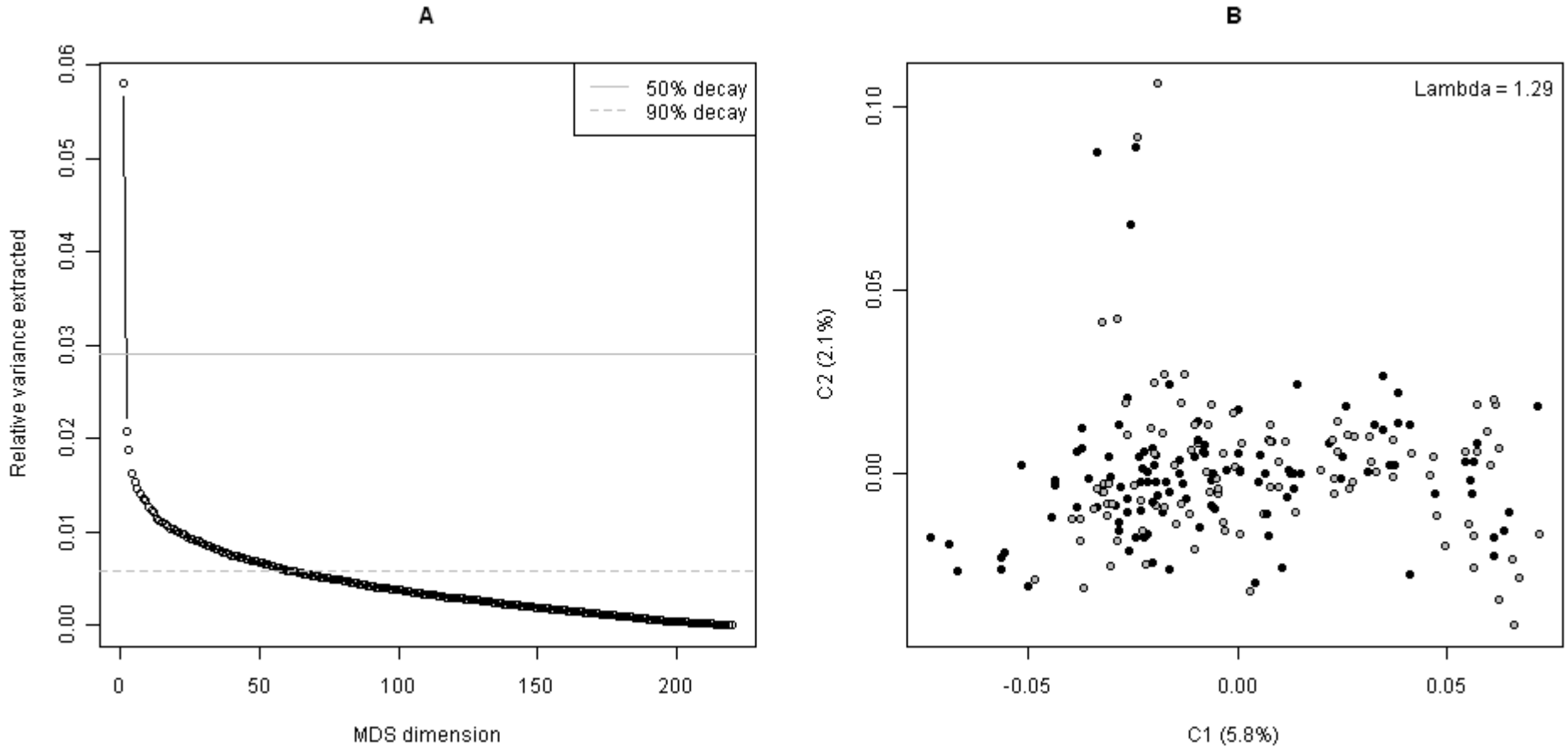

A
50% decay
90% decay
Relative variance extracted
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
50
100
150
200
MDS dimension
B
Lambda = 1.29
C2 (2.1%)
0.00
0.05
0.10
-0.05
0.00
0.05
C1 (5.8%)

Figure S3

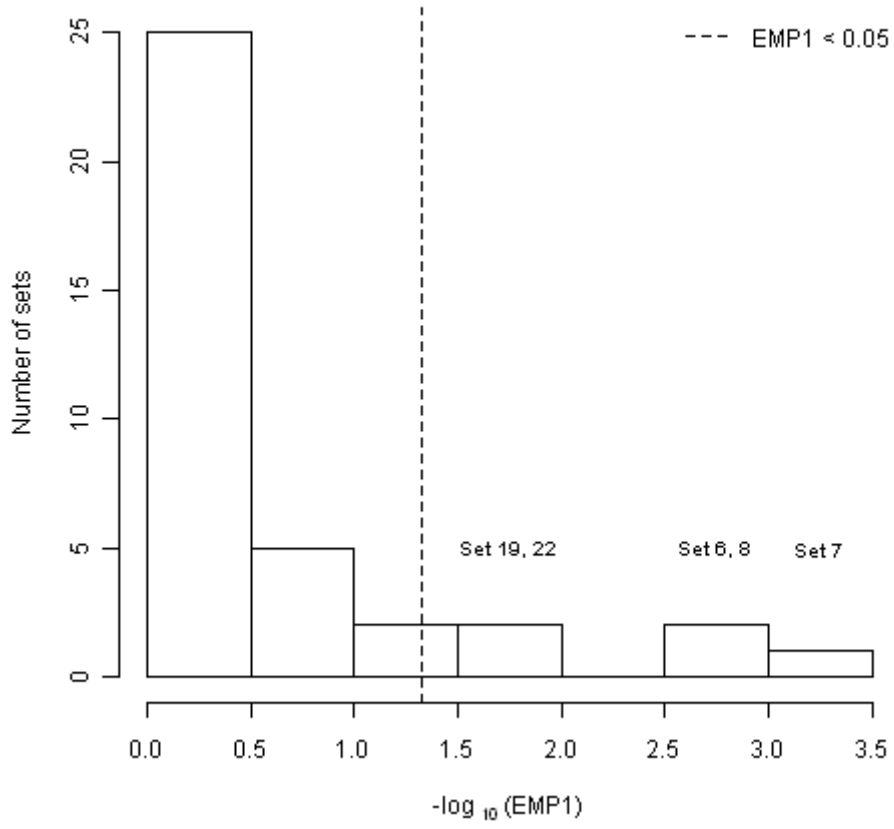

EMP1 < 0.05
Number of sets
0
5
10
15
20
25
Set 19, 22
Set 6, 8
Set 7
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
-log 10 (EMP1)

Figure S4

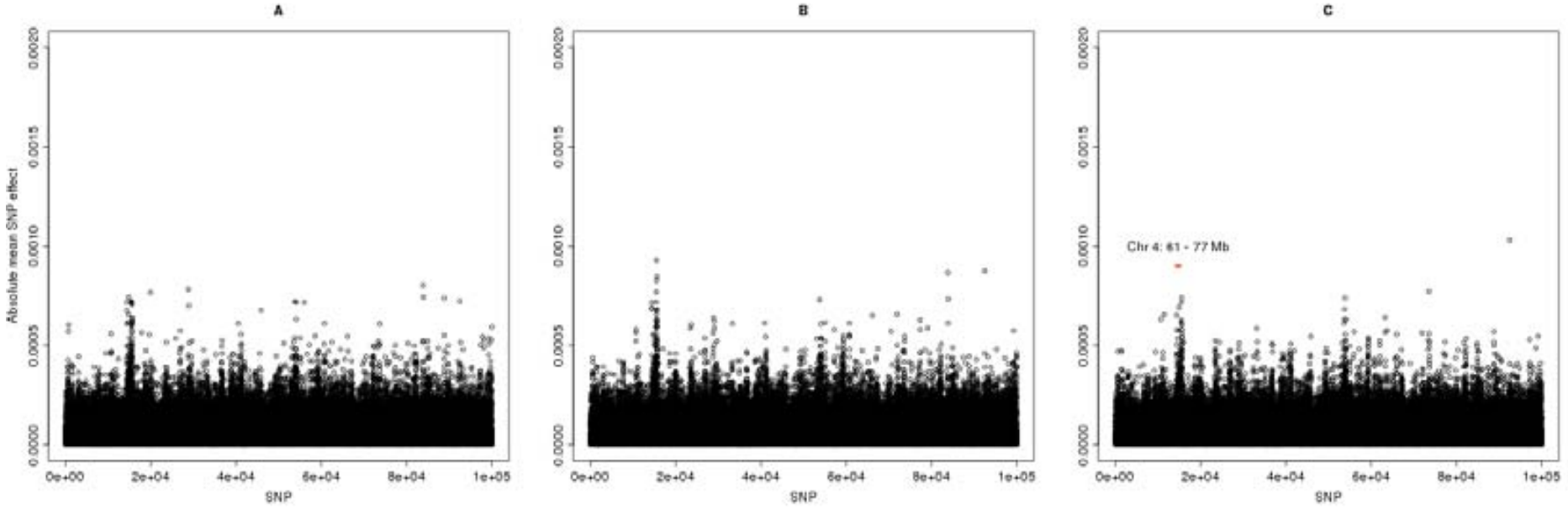

A
B
C
Absolute mean SNP effect
0.0000
0.0005
0.0010
0.0015
0.0020
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
SNP
Chr 4: 61 - 77 Mb

Table S1

			MAF		Model 1			Model 2			Model 3			
CFA	**Base pair**	**SNP**	**Affected**	**Healthy infected**	P_{raw}	P_{genome}	OR	P_{raw}	P_{genome}	OR	P_{raw}	P_{genome}	OR	**Genes**
1	**39,058,553**	**BICF2P1116416**	0.03	0.15	1.0×10^{-5}	>0.10	0.17	2.1×10^{-4}	>0.50	0.17	2.7×10^{-4}	>0.50	0.18	*STX11, NP_001012395.1*
4	**68,238,371**	**BICF2S2369174**	0.48	0.27	1.1×10^{-5}	>0.10	2.51	2.5×10^{-4}	>0.50	2.25	3.4×10^{-4}	>0.50	0.22	*ENSCAFG00000023014, XM_850162.1, ENSCAFG00000018512, MRPS30*
9	**45,276,074**	**BICF2P1187317**	0.005	0.09	2.4×10^{-4}	>0.05	0.05	0.012	>0.50	0.08	0.012	>0.50	0.08	*EVI2B, NOS2_CANFA, OMG, XM_848439.1, WSB1, KSR1, Q6QZP2_CANFA, NF1, NLK, ENSCAFG00000026192*
10	**31,276,530**	**BICF2G630481843**	0.2	0.08	1.5×10^{-4}	>0.50	2.89	2.4×10^{-4}	>0.50	3.37	1.8×10^{-4}	>0.50	3.45	*TXN2, MYH9_CANFA, FOXRED2 ENSCAFG00000001689, ENSCAFG00000023115, RBFOX2, CACNG2, EIF3D, APOL5, APOL6, MYG_CANFA, RASD2*
20	**30,132,329**	**BICF2P1345879**	0.09	0.24	2.5×10^{-5}	>0.10	0.31	NA	NA	NA	NA	NA	NA	*ADAMTS9, PRICKLE2, PSMD6, ATXN7, THOC7, C3orf49, SNTN, SYNPR*

Table S2

MDS dimensions[a]	**Relative genetic variance (%)**[b]	**λ**	**s.e.**
C3	1.83	1.19	0.00018
C3, C4	1.64	1.17	0.00017
C3, C4, C5	1.52	1.19	0.00017

[a]In addition to MDS dimensions C1, C2 and lifestyle.

[b]It refers to the relative genetic variance explained by the last MDS dimension included.

Table S3

Prevalence	h^2	**s.e.**
0.01	0.32	0.10
0.05	0.49	0.15
0.10	0.61	0.18
0.20	0.76	0.23
0.30	0.85	0.25
0.40	0.90	0.27
0.50	0.91	0.27
0.60	0.90	0.27

Text S1

Calculation of Fitted values (ŷ)

BayesB produced estimates of the genomic breeding values (GEBV) and of the effects of C1, C2 and each lifestyle category, which were used to calculate fitted values of the phenotype (ŷ) according to the different predictive models:

$$\text{Model}1: \quad y = \mu \cdot 1_n + \text{GEBV}$$

$$\text{Model}2: \quad y = \mu \cdot 1_n + E_{C1} \cdot (C1 - \overline{C1}) + E_{C2} \cdot (C2 - \overline{C2}) + \text{GEBV}$$

$$\text{Model}3: \quad y = \mu \cdot 1_n + E_{C1} \cdot (C1 - \overline{C1}) + E_{C2} \cdot (C2 - \overline{C2}) + E_{life} \cdot X_{life} + \text{GEBV}$$

Where, ŷ is the vector of fitted values, μ is the overall mean, 1_n is a vector of *n*-samples ones, E_{C1} and E_{C2} are the estimated effects for C1 and C2, respectively, C1 and C2 are vectors with the eigenvalues for C1 and C2 MDS dimensions, respectively, $\overline{C1}$ and $\overline{C2}$ are the mean values for C1 and C2, respectively, E_{life} is a vector with the estimated effect for each lifestyle, X_{life} is a matrix with as many rows as samples and as many columns as lifestyle levels with 1 degree of freedom, taking value 1 for the recorded lifestyle and zero otherwise. Throughout the analyses, correlations between actual phenotypes and fitted values were calculated as the Pearson product-moment correlation coefficient (*r*).

Fraction of correct predictions (*g*)

X was defined as a random variable following a hypergeometric distribution describing the number of healthy infected declared correctly, with parameters: a, number of true affected, c, number of true healthy infected,

n = a + c, and r, sample of individuals declared at random to be healthy infected (r ≤ n). Thus:

$$E(X) = \frac{r \cdot c}{n}$$

$$Var(X) = \frac{r \cdot c(n-r) \cdot (n-c)}{n^2 \cdot (n-1)} = \frac{r \cdot a \cdot c \cdot (n-r)}{n^2 \cdot (n-1)}$$

Y was defined as the total correct predictions:

$$Y = X + a - (r - x) = X + a - r + X = 2 \cdot X + (a - r)$$

$$E(Y) = E(2 \cdot X + (a-r)) = E(2 \cdot X) + E(a-r) = 2 \cdot E(X) + (a-r) = \frac{2 \cdot r \cdot c}{n} + (a-r)$$

$$Var(Y) = Var(2 \cdot X + (a-r)) = Var(2 \cdot X) = 4 \cdot Var(X) = \frac{4 \cdot r \cdot a \cdot c \cdot (n-r)}{n^2 \cdot (n-1)}$$

Z was defined as the fraction of correct calls:

$$Z = \frac{Y}{n} = \frac{2 \cdot X + (a-r)}{n}$$

$$E(Z) = E\left(\frac{2 \cdot X}{n} + \frac{a-r}{n}\right) = \frac{2}{n} \cdot E(X) + \frac{a-r}{n} = \frac{2 \cdot r \cdot c}{n^2} + \frac{a-r}{n}$$

$$Var(Z) = Var\left(\frac{2}{n} \cdot X + \frac{a-r}{n}\right) = \frac{4}{n^2} \cdot Var(X) = \frac{4 \cdot r \cdot a \cdot c \cdot (n-r)}{n^4 \cdot (n-1)}$$

The random 95% limit of Z was calculated from X | F(X) ≥ 0.95, where F is the distribution function.

Sensitivity and specificity

The number of individuals correctly predicted as healthy infected ($X_{observed}$) and correctly predicted as affected ($n - r_{observed}$) was used to calculate:

$$\text{Specificity} = \frac{X_{observed}}{c}$$

$$\text{Sensitivity} = \frac{\left(n - r_{observed}\right)}{a}$$

$$g = \frac{2 \cdot X_{observed} + \left(a - r_{observed}\right)}{n}$$

PAPER II

Quilez et al. BMC Genomics 2011, 12:339
http://www.biomedcentral.com/1471-2164/12/339

BMC Genomics

RESEARCH ARTICLE Open Access

A selective sweep of >8 Mb on chromosome 26 in the Boxer genome

Javier Quilez[1*], Andrea D Short[2], Verónica Martínez[1], Lorna J Kennedy[2], William Ollier[2], Armand Sanchez[1], Laura Altet[1] and Olga Francino[1]

Abstract

Background: Modern dog breeds display traits that are either breed-specific or shared by a few breeds as a result of genetic bottlenecks during the breed creation process and artificial selection for breed standards. Selective sweeps in the genome result from strong selection and can be detected as a reduction or elimination of polymorphism in a given region of the genome.

Results: Extended regions of homozygosity, indicative of selective sweeps, were identified in a genome-wide scan dataset of 25 Boxers from the United Kingdom genotyped at ~20,000 single-nucleotide polymorphisms (SNPs). These regions were further examined in a second dataset of Boxers collected from a different geographical location and genotyped using higher density SNP arrays (~170,000 SNPs). A selective sweep previously associated with canine brachycephaly was detected on chromosome 1. A novel selective sweep of over 8 Mb was observed on chromosome 26 in Boxer and for a shorter region in English and French bulldogs. It was absent in 171 samples from eight other dog breeds and 7 Iberian wolf samples. A region of extended increased heterozygosity on chromosome 9 overlapped with a previously reported copy number variant (CNV) which was polymorphic in multiple dog breeds.

Conclusion: A selective sweep of more than 8 Mb on chromosome 26 was identified in the Boxer genome. This sweep is likely caused by strong artificial selection for a trait of interest and could have inadvertently led to undesired health implications for this breed. Furthermore, we provide supporting evidence for two previously described regions: a selective sweep on chromosome 1 associated with canine brachycephaly and a CNV on chromosome 9 polymorphic in multiple dog breeds.

Background

It has been proposed that the majority of modern dog breeds recognised today have resulted from two population bottlenecks in dog evolution [1,2]. During the first genetic bottleneck, pre-domestic breeds diverged from wolves some 15,000 years ago, probably through multiple domestication events. The second bottleneck for most breeds occurred within the last few hundred years, when the breed creation process resulted in the loss of genetic variation due to strong bottleneck events which occurred in parallel with strong artificial selection for behavioural and physical characteristics favoured by humans.

The same bottlenecks and artificial selection forces that generated these breed-specific features have, in some instances, provoked undesired health effects. Random fixation of detrimental variants can occur during bottlenecks. Similarly, risk alleles may be in linkage disequilibrium with selected phenotypic variants or these may have pleiotropic effects [3,4].

Several studies have previously aimed to identify genomic regions involved in defined traits and their relationship with disease using association mapping (reviewed in Karlsson and Linblad-Toh [2]). However, phenotypic traits that have been driven to fixation by genetic drift or artificial selection within a dog breed cannot be mapped within that breed with this approach. An alternative in these cases is selection mapping, in which selective

* Correspondence: Javier.quilez@uab.cat
[1]Molecular Genetics Veterinary Service (SVGM), Department of Animal and Food Science, Veterinary School, Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Barcelona, Spain
Full list of author information is available at the end of the article

sweeps (a reduction or elimination of genetic polymorphism in a region owing to strong selection) are searched [2,5-8]. The aim of this work was to identify selective sweeps in the Boxer genome resulting from the breed creation process using high density genome-wide SNP data. These regions are likely to govern phenotypic traits of interest and may be linked to overrepresentation of certain genetic disorders in this breed.

Results

Detection and replication in Boxer

Regions of homozygosity (ROHs) in the *Canis familiaris* chromosomes (CFA) were identified in 25 Boxers from the United Kingdom (UK) which had been genotyped on microarrays for ~20,000 SNPs (set A, Table 1). Eight ROHs meeting the criteria detailed in Methods were detected (Table 2), representing 22 Mb (~0.9%) of the dog genome. Three of these ROHs, two on CFA 1 and one on CFA 26, showed a remarkable extended low heterozygosity (Figure 1a, Table 2). To confirm these ROHs, a second dataset (set B, Table 1) was generated using a higher density SNP array (~170,000 SNPs). This related to Boxers collected from different geographical locations to set A. In set B, 27 ROHs were found, which spanned 40.8 Mb (~1.7%) of the dog genome. Three regions on CFA X (Figure 1b) were discarded as these were not present when only female samples were analyzed (data not shown). Five ROHs were shared in both sets (Table 2). In general, these were notably shorter and/or split into two separated shorter regions when a higher number of samples and SNPs were genotyped (Additional file 1: Figure S1a-d, f-i). Conversely, the first 8 Mb of CFA 26 showed almost total loss of heterozygosity (average observed marker heterozygosity < 0.02) in both sets (Additional file 1: Figure S1e, j). There was a single SNP (BICF2G630807104, CFA 26:4,222,068 bp) with MAF = 0.5 within the region of extended homozygosity on CFA 26 (Additional file 1: Figure S1j), closer examination of which showed that heterozygous genotypes had been called for all Boxer samples. Possible explanations might be wrong genotype call from intensity data or a structural variation affecting that single SNP. To avoid the concern about SNPs significantly deviating from Hardy-Weinberg Equilibrium (HWE) affecting the identification of ROHs the analysis was repeated in set B after the removal of SNPs with HWE test p-value < 0.005, which resulted in similar results (Additional file 2 and Additional file 3). For the subsequent analyses we focused on the ROHs on CFA 1:58,710,420-61,801,815 bp and CFA 26:3,008,718-11,914,284 bp because these had markedly larger size and lower levels of variation than other regions common in both sets (Figure 1, Table 2). Finally, we found a region of increased heterozygosity on CFA 9:19,826,590-21,137,140 bp (Figure 1b), closer examination of which revealed a region of approximately 1.5 Mb showing a pattern of alternate heterozygous and homozygous genotypes indicating a CNV (Figure 2).

Presence in other breeds

Since reduction of genetic polymorphism in a region can result from strong selection and brachycephaly is a breed-defining trait in the Boxer, we evaluated the presence of the ROHs on CFA 1 and 26 in non-brachycephalic and brachycephalic breeds. Brachycephaly is characterized by severe shortening of the muzzle, and therefore the underlying bones, and a more modest shortening and widening of the skull [9]. For both selective sweeps on CFAs 1 and 26, normal levels of heterozygosity were observed in non-brachycephalic dog breeds and the Iberian wolf (Figure 3), based on a first dataset containing 118 samples from 6 different dog breeds and 7 Iberian wolf samples genotyped

Table 1 Samples genotyped with call rate > 90%

Group	# Samples	BeadChip
Boxer (denoted as set A)	25	Illumina's CanineSNP20 (~20,000 SNPs)
Iberian wolf	7	
Shar pei	37	
Cirneco dell'Etna	12	
Canarian warren hound	13	
Ibizan hound	39	
Pharaoh hound	5	
Labrador retriever[1]	12	Illumina's CanineHD (~170,000 SNPs)
Boxer (denoted as set B)	273	
German shepherd dog (GSD)	43	
English bulldog	4	
French bulldog	6	
Pug	10	

[1]a subset of ~17,000 SNPs shared with the Illumina's CanineSNP20 was used for this breed.

Table 2 Regions of homozygosity (ROHs) detected in sets A and B.

Set A								
Region ID	**CFA**	**BP1**	**BP2**	**KB**	**NSNP**	**KB/SNP**	**Avg. Het**	**Group**
SetA_01	1	57,121,838	67,829,167	10,707	84	127.5	0.02	*1
SetA_02	1	86,903,752	95,179,544	8,276	90	92.0	0.03	*2
SetA_03	1	112,558,604	120,632,074	8,073	71	113.7	0.05	*3
SetA_04	3	62,679,327	67,735,724	5,056	53	95.4	0.05	
SetA_05	10	3,081,933	12,082,601	9,001	68	132.4	0.04	*4
SetA_06	12	48,950,221	55,687,381	6,737	63	106.9	0.04	
SetA_07	26	3,116,745	12,410,004	9,293	107	86.9	0.01	*5
SetA_08	29	11,263,518	18,670,911	7,407	69	107.4	0.05	
Set B								
Region ID	**CFA**	**BP1**	**BP2**	**KB**	**NSNP**	**KB/SNP**	**Avg. Het**	**Group**
SetB_01	1	26,672,978	27,730,188	1,057	78	13.6	0.03	
SetB_02	1	45,218,029	46,286,798	1,069	66	16.2	0.04	
SetB_03	1	58,732,954	61,801,815	3,069	221	13.9	0.03	*1
SetB_04	1	62,722,220	65,190,321	2,468	129	19.1	0.03	*1
SetB_05	1	89,187,131	90,230,941	1,044	86	12.1	0.04	*2
SetB_06	1	102,454,189	103,320,473	866	51	17.0	0.04	
SetB_07	1	116,688,554	118,107,497	1,419	97	14.6	0.03	*3
SetB_08	1	117,979,514	118,963,939	984	51	19.3	0.04	*3
SetB_09	2	22,715,411	24,024,566	1,309	78	16.8	0.04	
SetB_10	3	3,030,299	3,903,071	873	59	14.8	0.04	
SetB_11	5	4,697,408	6,247,457	1,550	105	14.8	0.04	
SetB_12	6	25,815,666	26,601,998	786	52	15.1	0.04	
SetB_13	6	42,437,711	43,955,158	1,517	60	25.3	0.04	
SetB_14	6	58,580,737	59,356,441	776	62	12.5	0.04	
SetB_15	9	3,529,583	4,304,182	775	58	13.4	0.04	
SetB_16	10	5,626,769	6,702,961	1,076	51	21.1	0.04	*4
SetB_17	10	59,180,359	60,456,532	1,276	90	14.2	0.04	
SetB_18	10	65,209,901	66,606,742	1,397	86	16.2	0.04	
SetB_19	10	68,323,860	69,026,432	703	61	11.5	0.04	
SetB_20	13	39,705,171	40,797,838	1,093	71	15.4	0.04	
SetB_21	14	19,806,989	20,470,904	664	52	12.8	0.04	
SetB_22	18	6,868,787	7,908,159	1,039	54	19.2	0.04	
SetB_23	20	7,816,139	8,635,017	819	58	14.1	0.03	
SetB_24	24	24,444,170	27,387,620	2,943	217	13.6	0.03	
SetB_25	24	28,845,341	29,848,829	1,003	93	10.8	0.03	
SetB_26	26	3,008,718	11,914,284	8,906	707	12.6	0.01	*5
SetB_27	30	38,126,268	38,689,821	564	52	10.8	0.04	

CFA, *Canis familiaris* chromosome; BP1 and BP2, base pair positions of the first and last SNP involved in the ROH, respectively; NSNP, number of SNPs involved in the ROH; KB/SNP, SNP density in the ROH; Avg. Het, average observed heterozygosity. ROHs present in both sets are grouped numerically. Note that the end of SetB_07 overlaps with start of SetB_08.

using the same panel of SNPs as in set A and on a second dataset containing 43 samples from German shepherd dog genotyped using the same panel of SNPs as in set B (Table 1).

The selective sweep on CFA 1 was present and showed allelic match with the Boxer (data not shown) in other brachycephalic breeds such as English bulldog, Pug and French bulldog although in the latter the reduction in heterozygosity was not as extended as in the other two breeds and seemed to be located slightly upstream in the chromosome (Figure 3). The selective sweep on CFA 26 was detected with allele-matching in English bulldog (CFA 26:6,785,609-11,282,297 bp; only 3 out of 368 SNPs with non-zero MAF) and French bulldog (CFA 26:8,548,679-9,227,043; 48 SNPs involved with MAF equal to zero) (Figure 3, Additional file 4). Although the ROH was not apparent in the Pug in Figure 3, in the segment of the ROH shared between

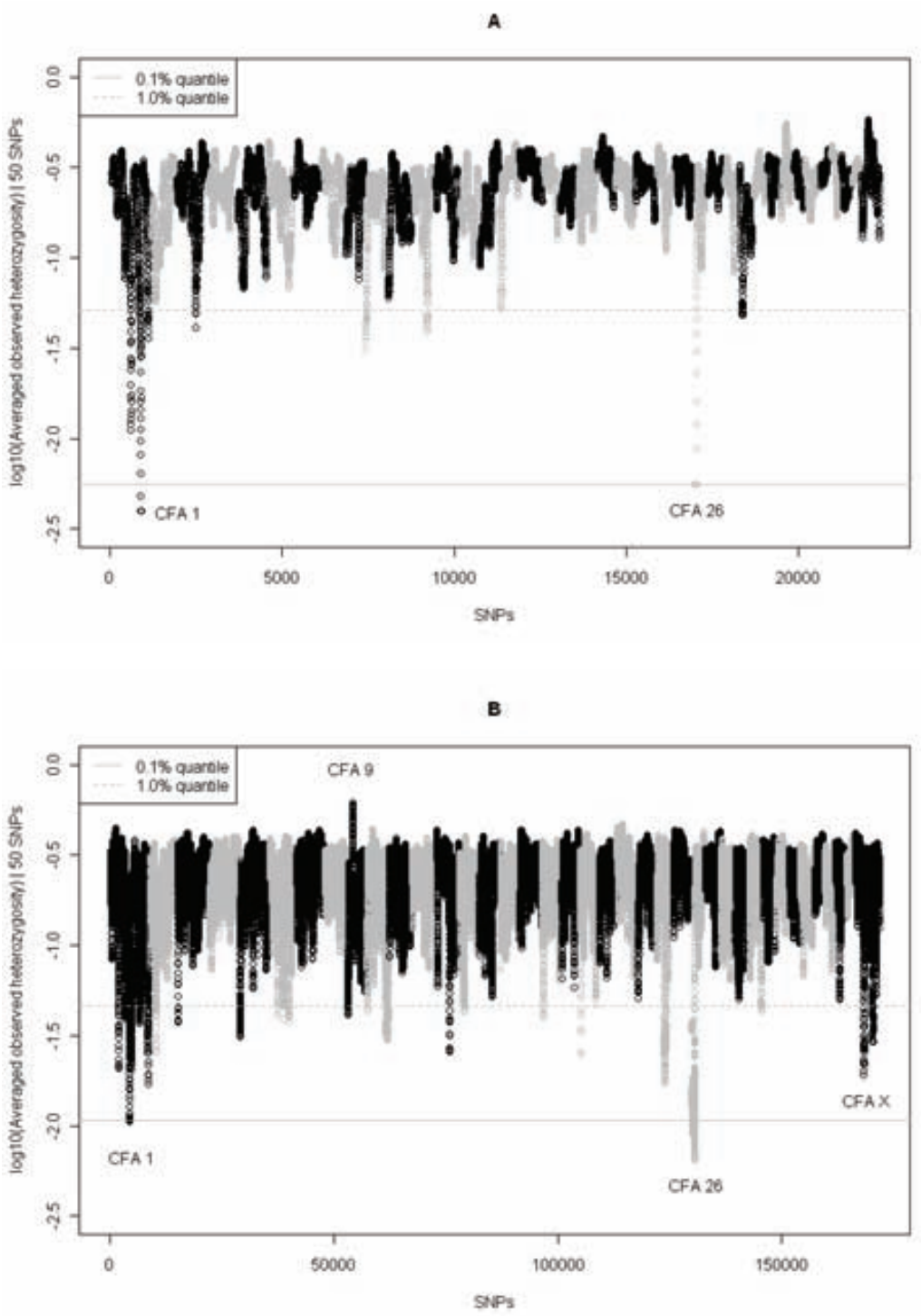


Figure 1 Genome-wide plot of the averaged observed heterozygosity with the 50-SNP sliding window for set A (a) and set B (b). The x-axis corresponds to SNPs sort by chromosomes (differently coloured) and position and the y-axis represents the 10-logarithm of the averaged observed heterozygosity calculated with sliding windows of 50 SNPs, for which 0.1 and 1.0% quantiles of the empirical distribution are displayed.

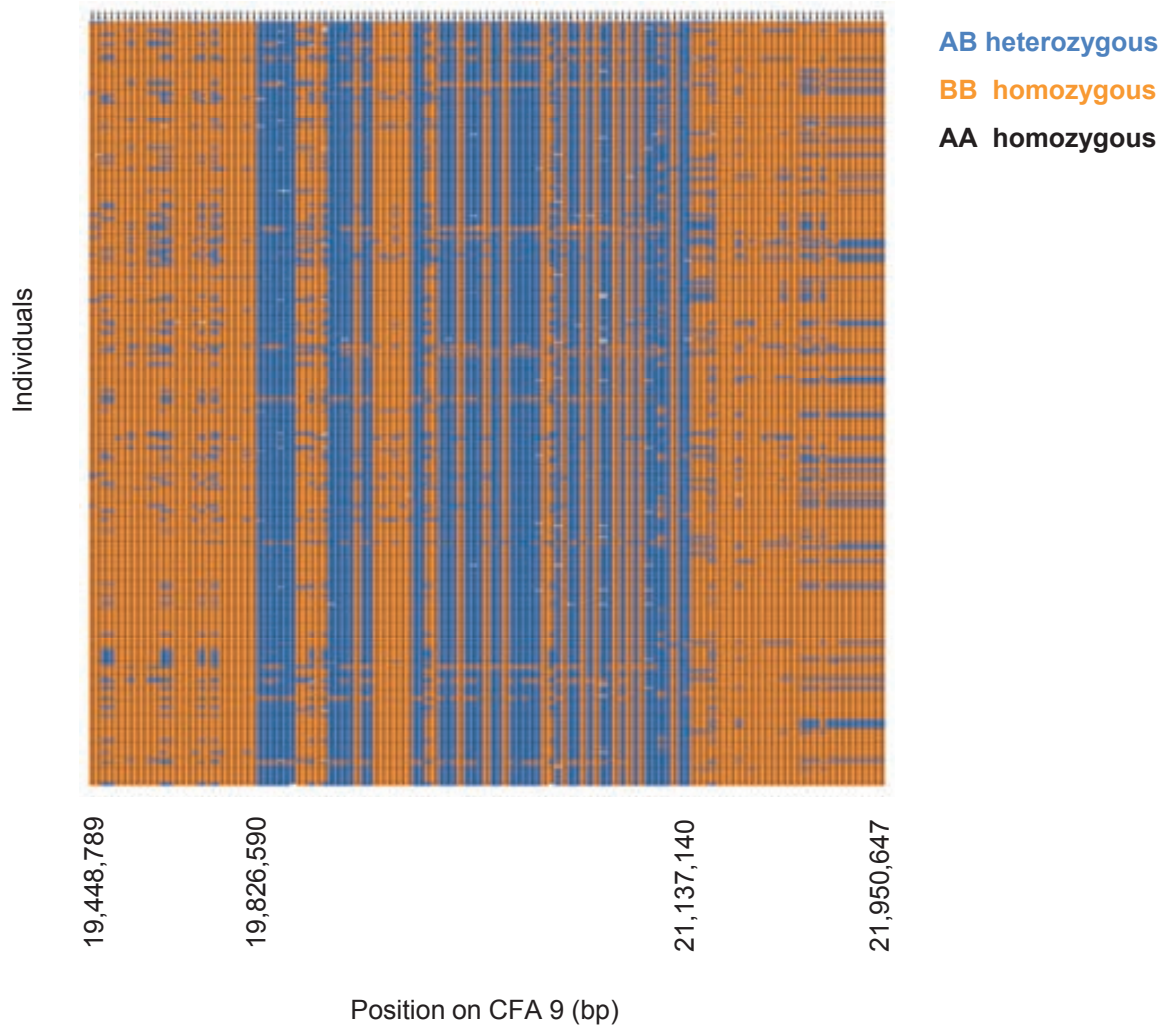


Figure 2 Pattern of alternate heterozygous/homozygous genotypes in a 1.5-Mb region corresponding to a described CNV on CFA 9. In the x-axis, the start and end positions of both the chromosome region shown and the CNV are indicated.

Boxer and English and French bulldogs a good number of SNPs in Pug samples showed the same genotypes as the other three brachycephalic breeds. In addition, the SNPs comprised between 8,565,784 and 8,620,015 bp (~55 Kb) were nearly fixed in the Pug (Additional file 4). In contrast to the Boxer, a distribution of genotypes in HWE was seen for BICF2G630807104 (CFA 26:4,222,068 bp) in the bulldog breeds studied whereas it was fixed in the Pug (data not shown).

Within the SNPs making up the Illumina's CanineSNP20 only two covered the CNV on CFA 9 (Additional file 5), both of them highly monomorphic with the exception of the SNP at position 20,274,406 bp for the Shar pei samples for which an excess of heterozygous genotypes (HWE test p-value < 0.001) were observed. A pattern of excessive heterozygous genotypes was observed for the region corresponding to the CNV on CFA 9 in the breeds genotyped with the Illumina's CanineHD Beadchip (Additional file 6).

Genetic content and functional annotation analysis

The region of decreased heterozygosity which was observed on CFA 1 in our study overlapped with a region previously associated with canine brachycephaly [7]. This was detected using dogs from brachycephalic breeds and non-brachycephalic breeds to perform across-breed association and selection mapping. Both strategies identified a region on CFA 1 at 59 Mb. The decrease in the averaged observed heterozygosity of brachycephalic dogs relative to non-brachycephalic dogs is indicative of a selective sweep at this position. Genes which have been associated with brachycephaly on CFA 1 include: *THBS2* [Ensembl:ENSCAFG00000000874], which is expressed in bone and cartilage during development and in the

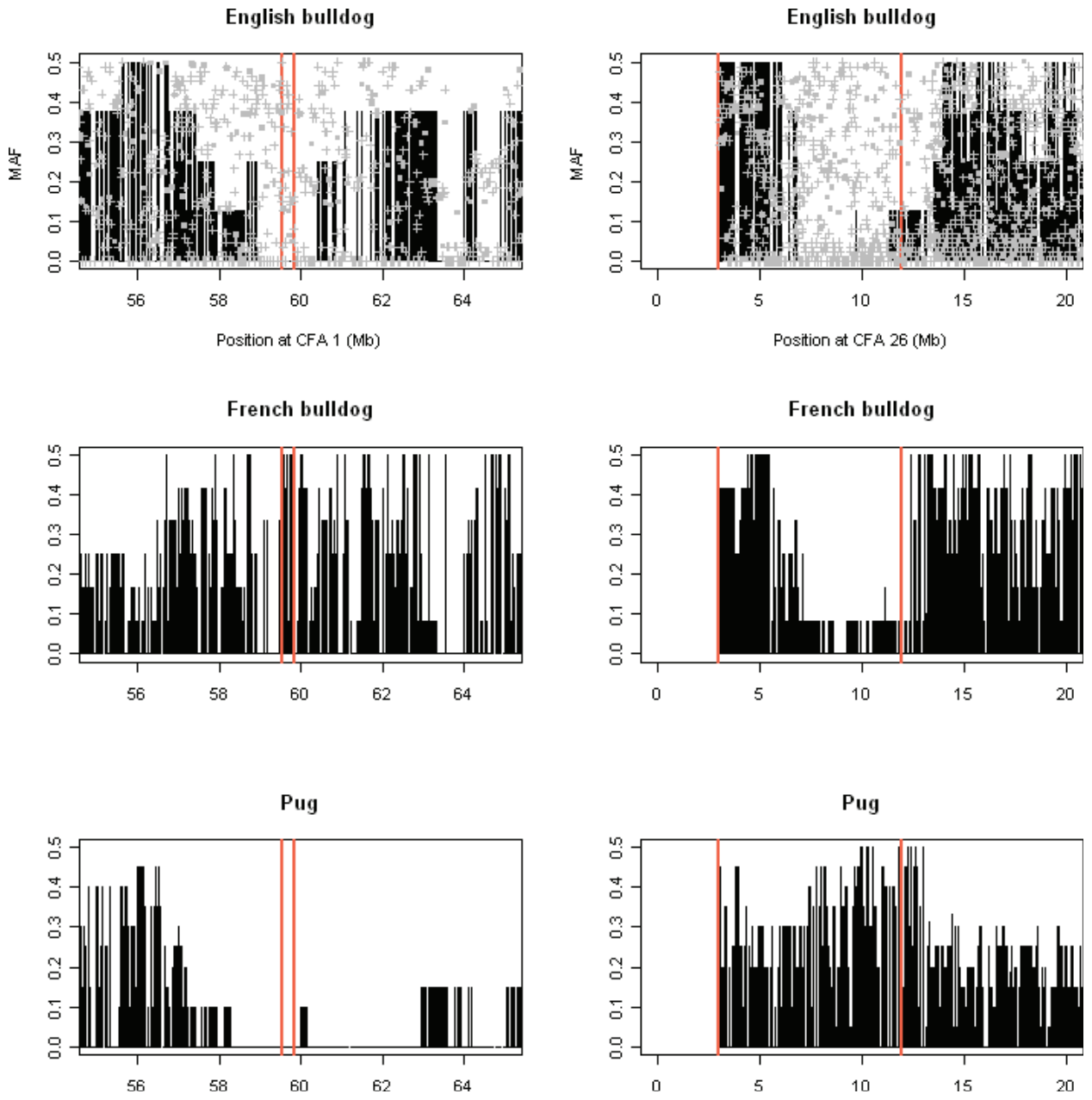


Figure 3 ROHs on CFA 1 and CFA 26 in non-brachycephalic and brachycephalic breeds other than Boxer. In the top panels, MAF values for the pool of non-brachycephalic breeds (grey squares) and German shepherd dog (grey crosses) are shown. For CFA 1, the region associated with brachycephaly by Bannasch *et al.* (CFA 1:59,536,208-59,832,965 bp) [7] is indicated (red lines). For CFA 26, the red lines show the extent of the ROH in Boxer for comparison with the other breeds.

adult skeleton [10], and *SMOC2* [Ensembl:ENSCAFG00000000868], similar in sequence to *BM-40* [Ensembl:ENSCAFG00000017855] which is expressed primarily during embryogenesis and in adult bone tissue [11].

The ROH in Boxer on CFA 26:3,008,718-11,914,284 bp contained 135 annotated elements of which 95 (71.9%) were genes with an associated name; this level was similar to that observed in the whole dog genome (69.7%). The ROH from CFA 26 mapped to two adjacent syntenic regions on *Homo sapiens* chromosome (Hs) 12 (Figure 4a). Advantage was taken of the fact that the region in the human genome syntenic to the region of interest on CFA 26 was better annotated and could be used to perform functional annotation analysis through the use of Ingenuity Pathways Analysis software [12]. One hundred and six dog to human orthologs annotated elements were

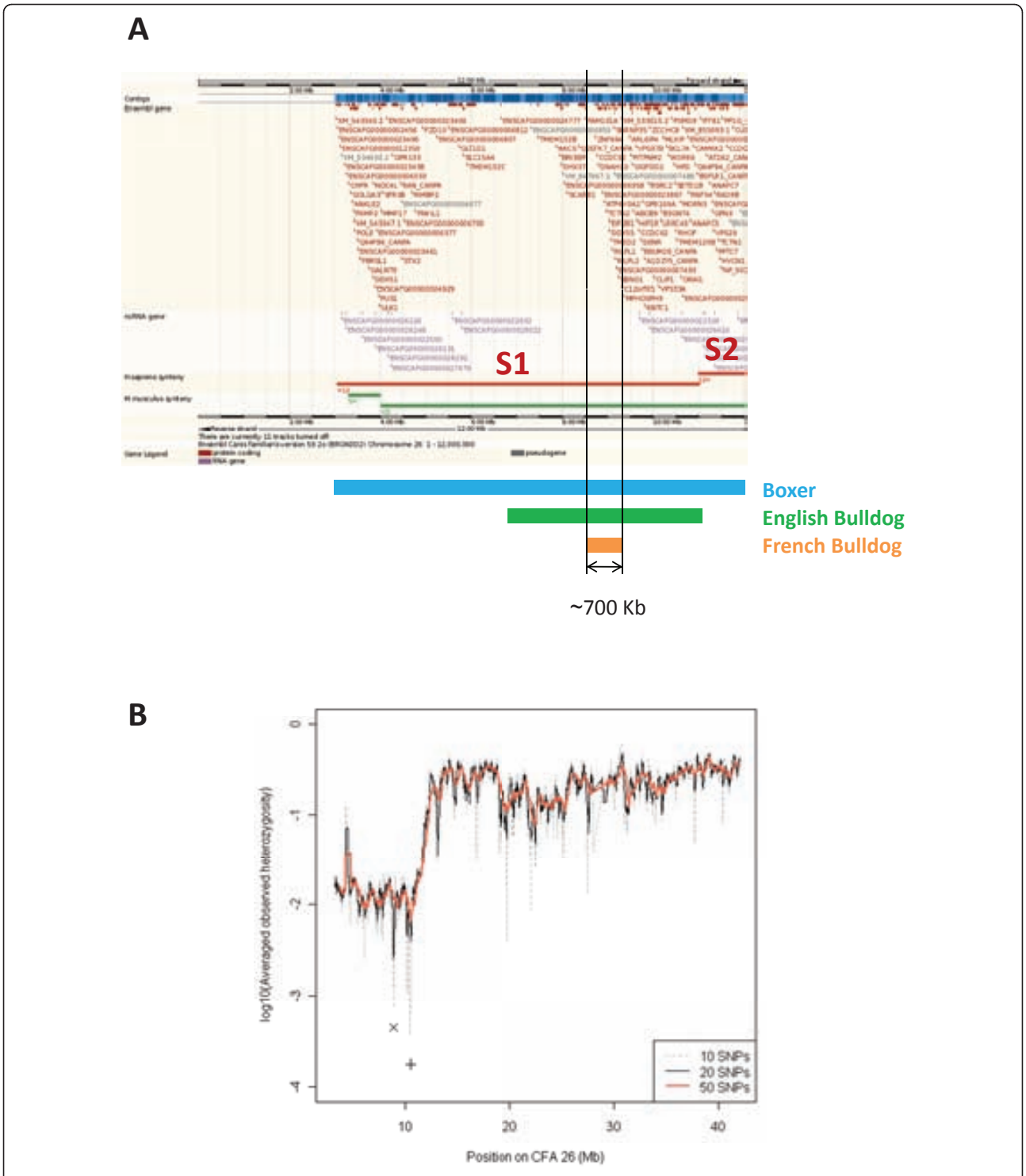


Figure 4 Genetic content of the ROH on CFA 26. **(a)** Annotated elements in the dog genome (CanFam2.0) as well as regions of synteny with *H. sapiens* and *M. musculus* are shown. CFA 26:3,008,718-11,914,284 bp maps to two syntenic regions in *H. sapiens*: Hs 12:121,547,433-133,784,108 bp (S1) and Hs 12:110,448,771-121,498,418 (S2). The end position for S2 is defined relative to the end of the ROH in dog though indeed synteny extends longer. Note that the order of the syntenic regions relative to the dog sequence indicates rearrangements events in the species; also, the different orientation of the sequences, indicated with arrows preceding the chromosome number in the compared species, designates an inversion in the orientation between *H. sapiens* and *C. familiaris* for S1. Regions of extended homozygosity for Boxer, English bulldog and French bulldog and the region of ~700 Kb shared in all three are indicated. **(b)** Averaged observed heterozygosity along CFA 26 in Boxer. Minimum values are shown for three window sizes: 10- and 50-SNPs, CFA 26:8.86 Mb (x); 20-SNPs, CFA 26:10.44 Mb (+).

used as input, resulting in a list of biological processes and disease categories that would be enriched given the genes in the region of interest. Amongst functional categories related to biological processes, skeletal and muscular system development and function as well as tissue morphology were the two most significantly associated categories (Additional file 7). Also, the selective sweep on CFA 26 contained genes that are linked to inherited diseases overrepresented in the Boxer [4] (Table 3). Lymphoblastic lymphoma is a type of non-Hodgkin lymphoma characterised by uncontrolled growth of either T- or B-cells and associated with *POLE* [Ensembl: ENSCAFG00000006215]. The frequency of T-cell lymphoma in the Boxer is higher than in other breeds [13-15]. Dilated cardiomyopathy (CMD) [OMIA: 327] is a disorder characterised by cardiac enlargement (especially of the left ventricle), poor myocardial contractility, and congestive heart failure. *MYL2* [Ensembl:ENSCAFG00000008524] is involved in the development of the sarcomere and muscle contraction and has also been associated with cardiomyopathy of the heart ventricle.

CFA 26:8,548,679-9,227,043 bp defined the region where extended homozygosity was common in Boxer, English bulldog and French bulldog (Figure 4a), which comprised six genes significant in the functional annotation analysis (Table 3). Of these genes, *ATP6V0A2* [Ensembl:ENSCAFG00000007234] and *EIF2B1* [Ensembl:ENSCAFG00000007434] are of special interest because they are involved in genetic disorders in humans. Various loss-of-function mutations of *ATP6V0A2* [Ensembl:ENSCAFG00000007234], which encodes the alpha-2 subunit of the V-type H+ ATPase, resulting in impaired glycosilation of proteins during synthesis cause autosomal recessive cutis laxa (ARCL) type II [OMIM:219200] and some cases of wrinkly skin syndrome [OMIM:278250] [16]. Mutations in each of the five subunits of the translation initiation factor eIF2B, including eIF2a encoded by *EIF2B1* [Ensembl: ENSCAFG00000007434], can cause leukoencephalopathy with vanishing white matter (VWM, [OMIM: 603896]) [17]. VWM is a neurological disorder manifesting progressive cerebellar ataxia, spasticity,

Table 3 Dog genes and human orthologs within the selective sweep on CFA 26

Boxer

Human gene	Dog gene	Start (bp)	End (bp)	Ensembl Gene ID	Annotation
HIP1R	*HIP1R*	9,641,249	9,654,192	ENSCAFG00000007764	Development of vertebral column
P2RX4	*Q64F94_CANFA*	10,917,230	10,933,173	ENSCAFG00000008363	Contraction (and relaxation) of cardiac muscle
P2RX7	*B0FLR1_CANFA*	10,959,759	11,002,514	ENSCAFG00000008401	Bone mineral density of skeleton Formation of perichondral bone Resorption of trabecular bone
ATP2A2	*AT2A2_CANFA*	11,167,290	11,232,956	ENSCAFG00000008434	Contraction (and relaxation) of cardiac muscle Contraction and relaxation of papillary muscle Elongation of cardiomyocytes
POLE	*POLE*	3,436,073	3,481,729	ENSCAFG00000006215	***Lymphoblastic lymphoma***
MYL2	*NP_001003069.1*	11,482,695	11,650,673	ENSCAFG00000008524	Contraction of cardiac muscle Development of sarcomere ***Cardiomyopathy of heart ventricle***

Boxer, English and French bulldogs

Human gene	Dog gene	Start (bp)	End (bp)	Ensembl Gene ID	Annotation
CCDC92	*CCDC92*	8,759,998	8,790,267	ENSCAFG00000006996	Amino Acid Metabolism Protein Synthesis Small Molecule Biochemistry
ATP6V0A2	*ATP6V0A2*	8,929,687	8,968,952	ENSCAFG00000007234	***Cutis laxa*** ***Wrinkly skin syndrome***
EIF2B1	*EIF2B1*	9,029,462	9,038,275	ENSCAFG00000007434	Gene Expression ***Leukoencephalopathy with vanishing white matter*** Protein Synthesis
DDX55	*DDX55*	9,040,108	9,053,306	ENSCAFG00000007452	***Cancer*** ***Infection Mechanism (by HIV)***
TMED2	*TMED2*	9,056,220	9,064,265	ENSCAFG00000007468	***Cancer*** ***Infection Mechanism (by HIV)***
SETD8		9,188,951	9,202,238	ENSCAFG00000007493	Tissue development and cell cycle

The upper part of the table (Boxer) refers to the selective sweep present in the Boxer only (CFA 26:3,008,718-11,914,284 bp). The lower part of the table (Boxer, English and French bulldogs) refers to the part of the selective sweep shared in the three breeds CFA 26:8,548,679-9,227,043 bp. Genomic positions in the dog genome and their annotated involvement in biological process and diseases (italic bold) are listed.

inconstant optic atrophy and relatively preserved mental abilities. *SETD8* [Ensembl:ENSG00000183955] encodes a lysine methyltransferase that regulates tumor suppressor p53 protein [18]. To note, the region of ~55 Kb of nearly complete homozygosity in the Pug is both upstream of these six genes and within the region of extended homozygosity shared amongst the three breeds. Interestingly, (i) the genes involved in skeletal and muscular system development and tissue morphology that were significant in the functional annotation analysis in the Boxer only (with the exception of *POLE* [Ensembl:ENSCAFG00000006215]) and (ii) the region shared in Boxer and English and French bulldogs, are both located within the region on CFA 26 that showed the greatest decay in the averaged observed heterozygosity (Figure 4b).

The region of alternate heterozygous/homozygous genotypes patterns observed for CFA 9 in the Boxers overlapped perfectly to a CNV previously described as being polymorphic in a number of dog breeds [19,20]. This 1.5-Mb CNV region contained three protein coding genes, one of which had two reported transcript variants, and four non coding RNA genes (Additional file 8). Analysis of Gene Ontology Biological Process (GO BP) terms revealed that the two transcript variants of the protein coding gene *VPS13D* [Ensembl:ENSCAFG00000016397] (CFA 9:21,079,541-21,164,823 bp) were associated with processes of protein localization and viral envelope fusion with host membrane (GO:0008104 and GO:0019064, respectively). This gene was also mapped to the Hs 1:12,290,124-12,572,099 bp but only the protein localization term was associated (GO:0008104). The remaining annotated elements had neither GO BP terms associated nor homology in *H. sapiens*.

Discussion

The substitution of a strongly selected mutation produces a selective sweep on the frequency of neutral alleles at linked loci characterised by a reduction of the local genetic variation [21-23]. Two selective sweeps detected on CFAs 1 and 26 in the Boxer genome were replicated in a larger sample size of the same breed obtained from a different geographical location and genotyped for a panel of SNPs of higher density. Assessing both the presence of these regions in other breeds and their genetic content can provide information on how they affect the phenotype and relate to the ancestral origin of breeds.

In our study, the selective sweep previously associated with brachycephaly on CFA 1 [7] was replicated in a larger sample of Boxers and in samples from other brachycephalic breeds. Moreover, samples in this study were from a different geographic area compared to the previous work [7] (Europe and US, respectively), suggesting the selective sweep is shared in the two populations within each breed.

The selective sweep on CFA 26 indicates strong artificial selection of a trait of interest in the Boxer, although the phenotypic trait resulting from this particular selective sweep is unknown. The sweep was not present in the Iberian wolf, one ancient breed (Shar pei), Labrador retrievers, German shepherd dogs or four hound breeds. On the other hand, it was present, although in shorter length, in English and French bulldogs, breeds that share with the Boxer the brachycephalic trait and a related breed creation process. Altogether, these results suggest that the selection of the sweep predated the formation of Boxer and both bulldog breeds. It is known that the English bulldog contributed to the breed creation of both Boxer and French bulldog breeds [24]. The Boxer is believed to have originated from a long-existing and now extinct German breed, the Bullenbeisser, which was crossed with a small number of English bulldog exemplars exported from the UK. Likewise, the French bulldog originated from toy varieties of English bulldog that were more popular in France. Moreover, it is interesting that the region of the selective sweep common in the three breeds coincides with the lowest reduction in the heterozygosity along the sequence (Figure 4). In selective sweeps, the reduction of genetic variation is lowest at the site of directional selection and not as great at distant sites due to recombination, although asymmetry in the valleys of reduced heterozygosity may provide imprecise information about the location of the sweep [25]. Based on our data, it is hard to assess whether the selective sweep on CFA 26 shared in Boxer and English and French bulldogs was also present in the Pug, also a brachycephalic breed, because only a short segment of reduced polymorphism within the sweep was observed in this breed (~55 Kb). Nonetheless, the history of the Pug differs from that of these three breeds mentioned before. The Pug dates to the ancient China and it is suggested that interbreeding with Pekingese, Japanese chin and possibly Shih tzu contributed to the breed creation process. Pugs were imported to Europe through Holland around 1,600s [24].

A possible scenario is that the standing neutral variation on CFA 26 present in the original English bulldog was passed to both Boxer and French bulldog during the breed creation process. Some variants would have been beneficial thereafter when selection of brachycephaly started, which is reasonable to think that happened during the breeds creation process since brachycephaly is a breed standard in these three types of dogs. Thus, strong selection of variants close to the position 8-10 Mb on CFA 26 contributing to brachycephaly might have swept nearby genetic variation. Variable selective sweep length

in the three breeds would response to different breed histories as it depends on the strength of selection, the amount of recombination and the population size [21,22,25]. Therefore, if one assumes the recombination rate to be similar across breeds for a given chromosome region, different across-breeds strength of selection and population sizes might have probably caused the variable length in the sweeps on CFA 26, which in the Boxer is more than ten times larger than in the French bulldog.

Altogether we suggest that CFA 26 may contain a footprint of selection for brachycephaly, especially in the Boxer. A brachycephalic head with a distinctive broad and blunt muzzle is a unique phenotype of the Boxer and particular attention is given to this trait by the Boxer breeding community. Although brachycephaly has been mapped to CFA 1 and the greatest association was greater than 100 times more significant than the second highest, Bannasch *et al* [7] suggested that the complex nature of the brachycephalic head phenotype may be the result of associations across multiple chromosomes. Verification as to whether the genome-wide significant markers on CFA 26, the second highest association, previously reported [7] are within the ROH on CFA 26 in our data would provide some support for a link between this region and selection for brachycephaly. Genes on CFA 1 which have been associated with brachycephaly are involved in skeletal development [7,10,11]. Similarly, genes significant in functional annotation analysis of our data were associated with skeletal and muscular system development and function as well as tissue morphology biological process (Table 3).

It is possible that the selection for certain breed-specific loci or locus might be in linkage disequilibrium with detrimental variants at other genes. Interestingly, some of the genes in the selective sweep region on CFA 26 could be related to diseases that are reported to be more common in the Boxer breed, particularly cancer (lymphoblastic lymphoma) and cardiovascular disorders (CMD) [4].

In addition, we could observe in our data a previously reported CNV on CFA 9 polymorphic in multiple dog breeds [19,20], providing evidence for within-breed variation in the number of segment copies. Our data suggest that the CNV on CFA 9 is present and variable in the 273 Boxers used in this study (Figure 1b, 2), as well as in other breeds such as German shepherd dog, Pug and English and French bulldogs (Additional file 6). We suggest this CNV may be also possibly present in the Shar pei (Additional file 5) although in this breed it should be confirmed with a panel of higher SNP density. It might be that variable numbers of copies of a gene contained within the CNV such as *VPS13D* [Ensembl:ENSCAFG00000016397], which is involved in entrance of virus into the host cell, might be functional in the susceptibility to viral infection. Likewise, the non coding RNAs (ncRNAs) and small nuclear RNAs (snRNAs) which precede a region relatively rich in genes (data not shown) could be functional in regulatory processes.

Conclusion

We have identified a selective sweep in excess of 8 Mb on CFA 26 in the Boxer which is not present in Iberian wolves or non-brachycephalic dog breeds. This region is a candidate for strong artificial selection in the Boxer for a trait of interest, possibly brachycephaly, and the inadvertent selection of genes during the enrichment for a certain phenotype may have given rise to an increased incidence of certain related afflictions in the breed. The fact that the selective sweep is also present in English and French bulldogs provides genetic evidence of a shared history of the three breeds.

Furthermore, we provide supporting evidence for two previously described regions: a selective sweep on CFA 1 associated with canine brachycephaly and a CNV on CFA 9 which is polymorphic in multiple dog breeds and contains genetic elements with potential biological implications.

Methods

Sample collection

A set of 27 Boxer samples from the UK (denoted as set A) were collected as residual samples from dogs taken for clinical investigation. They were selected from a large archive of DNA samples (UK Companion Animal DNA Archive, University of Manchester) and all samples had informed owner consent. A second set of 274 Boxer samples were collected from Spain, Greece, Italy and Portugal (denoted as set B); samples from Spain represented > 90% of set B. Dogs in this set and the remaining breeds in Table 1 came from the Hospital Clínic Veterinari of the Universitat Autònoma de Barcelona, veterinary clinics or dog owners.

DNA extraction and genotyping

DNA was extracted from peripheral blood or bone marrow samples using either QIAamp® DNA Blood Mini Kit (QIAGEN) or PureLink™ Genomic DNA (Invitrogen). Set A was genotyped at 22,362 SNPs with Illumina's CanineSNP20 BeadChip at The Genome Centre, Queen Mary University of London, UK. Set B and German shepherd dog samples were genotyped at 174,376 markers using Illumina's CanineHD BeadChip at The Centre National de Génotypage, France. The remaining samples were genotyped as indicated in Table 1 at the Universitat Autònoma de Barcelona, Spain.

Data cleaning

Data cleaning was conducted using PLINK and R packages [26,27]. Set A was filtered to have individual

and marker call rates > 90%, resulting in 25 Boxers and 22,300 SNPs left for analysis. The same filters were applied to set B and, moreover, in this set we also excluded intensity probes, markers on the boundary autosomal region on chromosome CFA X as well as those SNPs on the non-pseudoautosomal region on CFA X for which heterozygous genotypes in male samples were observed. All samples in set B had an individual call rate > 90% but one sample was excluded as it appeared as an outlier when the first two dimensions of the multidimensional scaling analysis were plotted (Additional file 9). This resulted in 273 individuals with 171,772 SNPs each left for analysis.

Statistical analysis

Averaged observed heterozygosity was calculated as the moving average of the observed heterozygosity using 50-SNPs windows both for set A (20,451 windows) and set B (169,812 windows). In each set the 1% of windows with the lowest averaged observed heterozygosity was selected (Additional file 2); windows spaced less than fifty times the mean SNP density (bp/SNP) of the beadchip used were considered as single regions of homozygosity. ROHs common in both sets were defined as those overlapping in at least one SNP. The analysis was also performed on the dataset with the SNPs with Hardy-Weinberg Equilibrium test p-value > 0.005 and the identified ROHs presented in Table 2 correspond to this second analysis.

Genetic content and functional annotation analysis

The position of the CNV detected on CFA 9 was defined as the union resulting from our data and the positions annotated in the Ensembl database [28] in two previous works describing this CNV [19,20]. This resulted in a region at CFA 9:19,778,695-21,332,928 bp that was searched for Gene Ontology biological process (GO BP) terms using Biomart [29] and regions of synteny with *H. sapiens*. For the ROH on CFA 26 Ensembl IDs of the annotated elements in the syntenic region at Hs 12:108,311,620-133,784,108 bp were retrieved using Biomart [29] and used as input for functional annotation analysis. The annotated genes in Hs 12:108,311,620-133,784,108 bp were tested for enrichment of certain biological functions or diseases by comparison with the annotations from the Ingenuity database for mouse, rat and human genomes [12]. Right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that data set was due to chance alone. The categories of diseases associated with the region of interest were compared with the reported inherited diseases in the Boxer breed [4].

Additional material

Additional file 1: Comparison of the ROHs on CFA 1, 10 and 26 common in sets A (Figure S1a-e) and B (Figure S1f-j). Red lines indicate ROH as defined in each set.

Additional file 2: Summary statistics of the 50-SNP sliding windows.

Additional file 3: Comparison of ROHs identified in Set B and Set B pruned for SNPs significantly deviating from HWE. Note that for the region setB_06 two different regions were detected and that the end of setB_07 overlaps with start of setB_08.

Additional file 4: Genotypes of samples from brachycephalic breeds for the segment of the ROH on CFA 26 shared between Boxer and English and French bulldogs. For the Boxer, a random sample of 20 dogs is displayed.

Additional file 5: Genotypes for the CNV on CFA 9 in samples from other breeds than Boxer and Iberian wolf genotyped with the Illumina's CanineSNP20. Chromosome positions corresponding to the described CNV are highlighted in green and genotypes grouped by breed. Homozygous genotypes are coded as '11' (grey) and '22' (orange), respectively, and heterozygous genotypes as '12' (blue); '00' represents missing values (white).

Additional file 6: Genotypes for the CNV on CFA 9 in samples from other breeds than Boxer and Iberian wolf genotyped with the Illumina's CanineHD. Chromosome positions corresponding to the described CNV are highlighted in green and genotypes grouped by breed. Homozygous genotypes are coded as '11' (grey) and '22' (orange), respectively, and heterozygous genotypes as '12' (blue); '00' represents missing values (white).

Additional file 7: FAA_long: functional annotation categories are classified either as biological processes (BP) or diseases (DIS). Function annotations that either involve genes laying in the region of homozygosity common in Boxer, English and French bulldogs (orange) or relate to inherited diseases overrepresented in the Boxer (blue) are highlighted. FAA_GenomicPosition: annotated elements in the selective sweep on chromosome 26 and in the syntenic region in the human genome. Whether annotated elements lay in the region of homozygosity in each breed (Boxer and English and French bulldogs) is indicated. Annotated elements significant in the functional annotation analysis are highlighted (orange). InheritedDiseases: inherited diseases reported in Boxer, English and French bulldogs [4]. Of these, those significant in the functional annotation analysis are highlighted (blue).

Additional file 8: File containing genetic content, Gene Ontology Biological Process (GO BP) annotation and synteny information for the CNV on CFA 9.

Additional file 9: Multidimensional scaling (MDS) plot of the two first dimensions C1 and C2. The excluded outlier samples are indicated by the arrows.

Acknowledgements

This work was funded by the European Commission (LUPA, GA-201370) and the Grant Number RR016466 from the National Center for Research Resources (NCRR), a component of the NIH and the American Kennel Club grant 0876-A. We would like to thank the referring clinicians, dog owners who gave permission for their dogs to participate in this study and the UK Companion Animal DNA Archive for providing the DNA from the UK samples.

Author details

[1]Molecular Genetics Veterinary Service (SVGM), Department of Animal and Food Science, Veterinary School, Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Barcelona, Spain. [2]Centre for Integrated Genomic Medical Research (CIGMR), Stopford Building, Oxford Road, Manchester, M13 9PT, UK.

Authors' contributions

AS, ASB, LA, LK, OF and WO designed the experiment. ASB, LA, OF and WO supervised the project and gave conceptual advice. AS, JQO, LK, and VMD

collected the samples. JQO and VMD performed DNA extraction and genotyping. AS provided technical support for the data analysis. JQO performed, data cleaning, statistical and genetic content analysis and wrote the manuscript. AS, LA, OF, VMD and WO edited the manuscript. All authors read and approved the final manuscript.

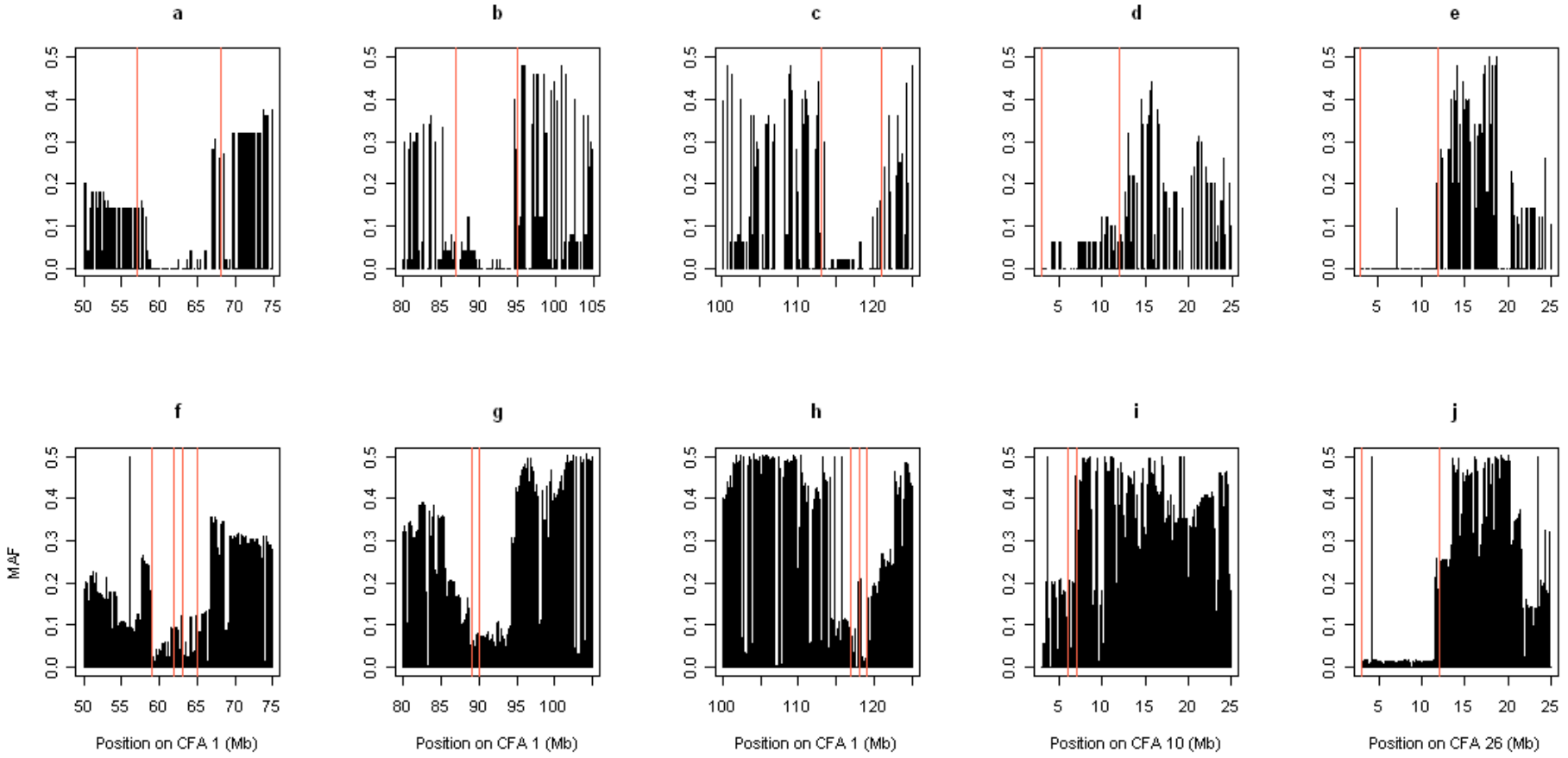
Received: 10 December 2010 Accepted: 1 July 2011
Published: 1 July 2011

References

1. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, *et al*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438**:803-819.
2. Karlsson EK, Lindblad-Toh K: **Leader of the pack: gene mapping in dogs and other model organisms.** *Nat Rev Genet* 2008, **9**:713-725.
3. Patterson DF, Haskins ME, Jezyk PF, Giger U, Meyers-Wallen VN, Aguirre G, Fyfe JC, Wolfe JH: **Research on genetic diseases: reciprocal benefits to animals and man.** *J Am Vet Med Assoc* 1988, **193**:1131-1144.
4. Sargan DR: **IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics.** *Mamm Genome* 2004, **15**:503-506.
5. 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, **316**:112-115.
6. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW: **Tracking footprints of artificial selection in the dog genome.** *Proc Natl Acad Sci USA* 2010, **107**:1160-1165.
7. Bannasch D, Young A, Myers J, Truve K, Dickinson P, Gregg J, Davis R, Bongcam-Rudloff E, Webster MT, Lindblad-Toh K, Pedersen N: **Localization of Canine Brachycephaly Using an Across Breed Mapping Approach.** *PLoS ONE* 2010, **5**:e9632.
8. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, Cargill M, Auton A, Reynolds A, Elkahloun AG, Castelhano M, Mosher DS, Sutter NB, Johnson GS, Novembre J, Hubisz MJ, Siepel A, Wayne RK, Bustamante CD, Ostrander EA: **A simple genetic architecture underlies morphological.** *variation in dogs PLoS Biol* 2010, **8**:e1000451.
9. Stockard CR: *The genetic and endocrine basis for differences in form and behavior* Philadelphia: The Wistar Institute of Anatomy and Biology; 1941.
10. Alford AIHK: **Matricellular proteins: Extracellular modulators of bone development, remodelling, and regeneration.** *Bone* 2006, **38**:749-757.
11. Porter PL SE, Lane TF, Funk SE, Gown AM: **Distribution of SPARC in normal and neoplastic human tissue.** *J Histochem Cytochem* 1995, **43**:791-800.
12. **Ingenuity® Systems.** [http://www.ingenuity.com].
13. Lurie DM, Lucroy MD, Griffey SM, Simonson E, Madewell BR: **T-cell-derived malignant lymphoma in the boxer breed.** *Vet Comp Oncol* 2004, **2(3)**:171-175.
14. Lurie DM, Milner RJ, Suter SE, Vernau W: **Immunophenotypic and cytomorphologic subclassification of T-cell lymphoma in the boxer breed.** *Vet Immunol Immunopathol* 2008, **125**:102-110.
15. Pastor M, Chalvet-Monfray K, Marchal T, Keck G, Magnol JP, Fournel-Fleury C, Ponce F: **Genetic and environmental risk indicators in canine non-Hodgkin's lymphomas: breed associations and geographic distribution of 608 cases diagnosed throughout France over 1 year.** *J Vet Intern Med* 2009, **23**:301-310.
16. Kornak U, Reynders E, Dimopoulou A, van Reeuwijk J, Fischer B, Rajab A, Budde B, Nürnberg P, Foulquier F, ARCL Debré-type Study Group, Lefeber D, Urban Z, Gruenewald S, Annaert W, Brunner HG, van Bokhoven H, Wevers R, Morava E, Matthijs G, Van Maldergem L, Mundlos S: **Impaired glycosylation and cutis laxa caused by mutations in the vesicular H+-ATPase subunit ATP6V0A2.** *Nat Genet* 2008, **40**:32-34.
17. van der Knaap MS, Leegwater PA, Könst AA, Visser A, Naidu S, Oudejans CB, Schutgens RB, Pronk JC: **Mutations in each of the five subunits of translation initiation factor eIF2B can cause leukoencephalopathy with vanishing white matter.** *Ann Neurol* 2002, **51**:264-270.
18. Shi X, Kachirskaia I, Yamaguchi H, West LE, Wen H, Wang EW, Dutta S, Appella E, Gozani O: **Modulation of p53 function by SET8-mediated methylation at lysine 382.** *Mol Cell* 2007, **27**:636-646.
19. Chen WK, Swartz JD, Rush LJ, Alvarez CE: **Mapping DNA structural variation in dogs.** *Genome Res* 2009, **19**:500-509.
20. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM: **The genomic architecture of segmental duplications and associated copy number variants in dogs.** *Genome Res* 2009, **19**:491-499.
21. Maynard Smith J, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genet Res* 1974, **23**:23-35.
22. Kaplan NL, Hudson RR, Langley CH: **The "hitchhiking effect" revisited.** *Genetics* 1989, **123**:887-899.
23. Stephan W, Wiehe THE, Lenz MW: **The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory.** *Theor Popul Bio* 1992, **41**:237-254.
24. American Kennel Club: *The Complete Dog Book* New York: G. H. Watt; 1929.
25. Kim Y, Stephan W: **Detecting a local signature of genetic hitchhiking along a recombining chromosome.** *Genetics* 2002, **160**:765-777.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a toolset for whole-genome association and population-based linkage analysis.** *American Journal of Human Genetics* 2007, **81**:559-575.
27. **The R Project for Statistical Computing.** [http://www.R-project.org].
28. Hubbard TJP, 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, *et al*: **Ensembl 2007.** *Nucleic Acids Res* 2007, , **35 Database:** D610-D617.
29. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal - unified access to biological data.** *Nucleic acids research* 2009, , **37 Web Server:** W23-27.

doi:10.1186/1471-2164-12-339
Cite this article as: Quilez *et al.*: A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. *BMC Genomics* 2011 **12**:339.

Additional file 1



Additional file 2

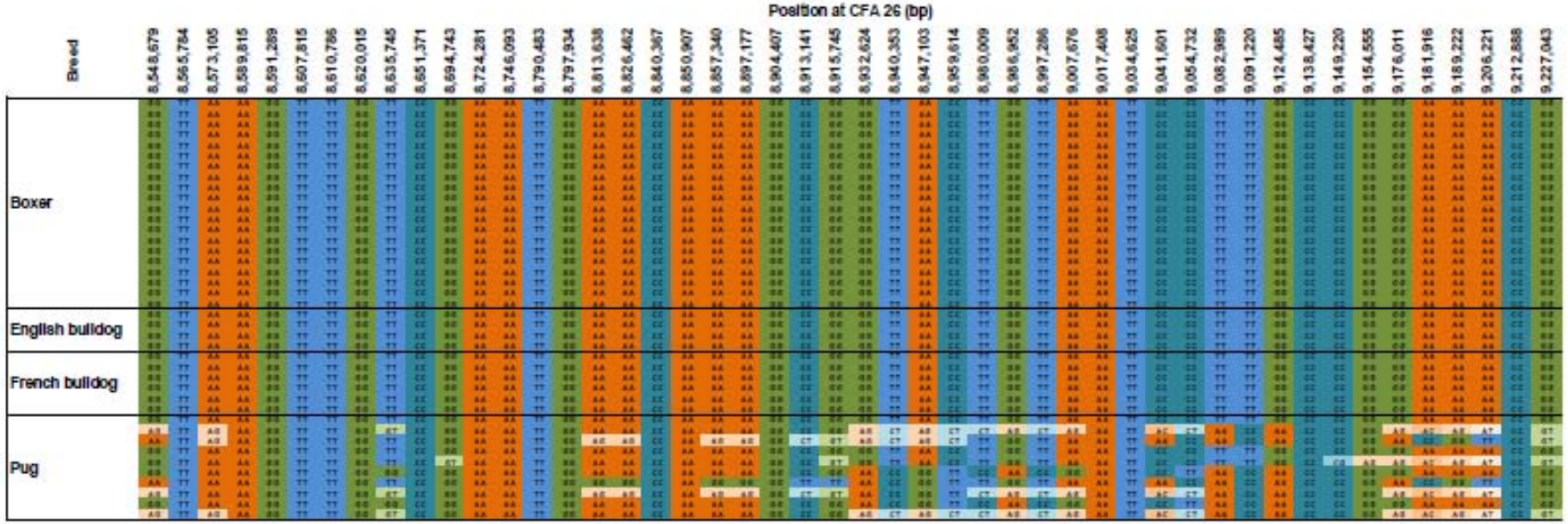
	Set A	**Set B**	**Set B (HWE test p-value > 0.005)**
Number of SNPs	22,362	171,772	168,239
Number of windows	20,451	169,812	166,279
Normalized heterozygosity			
1.0% quantile	0.051	0.046	0.045
0.1% quantile	0.006	0.011	0.011
Mean SNP density (Kb/SNP)	102	14	14
Maximum spacing between SNPs (Kb)	5,552	3,382	3,382

Additional file 3

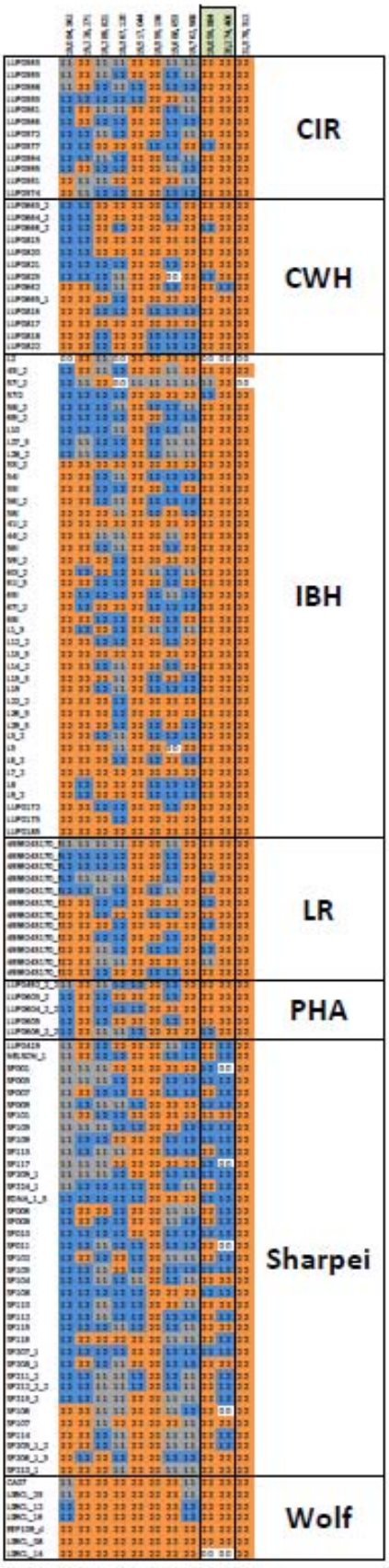
		Set B		***Set B*** ***(HWE test p-value > 0.005)***			
Region ID	**CFA**	**BP1**	**BP2**	**BP1**	**BP2**	**Difference in BP1**	**Difference in BP2**
SetB_01	1	26,672,978	27,730,188	26,672,978	27,730,188	0	0
SetB_02	1	45,207,395	46,286,798	45,218,029	46,286,798	10,634	0
SetB_03	1	58,710,420	61,801,815	58,732,954	61,801,815	22,534	0
SetB_04	1	62,722,220	65,201,836	62,722,220	65,190,321	0	11,515
SetB_05	1	89,051,381	90,230,941	89,187,131	90,230,941	135,750	0
SetB_06	1	91,184,730	92,038,243	102,454,189	103,320,473	-	-
SetB_07	1	116,683,627	118,107,497	116,688,554	118,107,497	4,927	0
SetB_08	1	117,979,514	118,963,939	117,979,514	118,963,939	0	0
SetB_09	2	22,715,411	23,895,662	22,715,411	24,024,566	0	128,904
SetB_10	3	3,030,299	3,903,071	3,030,299	3,903,071	0	0
SetB_11	5	4,697,408	6,247,457	4,697,408	6,247,457	0	0

SetB_12	6	25,815,666	26,601,998	25,815,666	26,601,998	0	0
SetB_13	6	42,437,711	43,447,406	42,437,711	43,955,158	0	507,752
SetB_14	6	58,478,112	59,356,441	58,580,737	59,356,441	102,625	0
SetB_15	9	3,529,583	4,363,516	3,529,583	4,304,182	0	59,334
SetB_16	10	5,626,769	6,840,140	5,626,769	6,702,961	0	137,179
SetB_17	10	59,180,359	60,456,532	59,180,359	60,456,532	0	0
SetB_18	10	65,161,142	66,656,760	65,209,901	66,606,742	48,759	50,018
SetB_19	10	68,323,860	69,026,432	68,323,860	69,026,432	0	0
SetB_20	13	39,894,733	40,797,838	39,705,171	40,797,838	189,562	0
SetB_21	14	19,789,648	20,470,904	19,806,989	20,470,904	17,341	0
SetB_22	18	6,868,787	7,908,159	6,868,787	7,908,159	0	0
SetB_23	20	7,816,139	8,638,174	7,816,139	8,635,017	0	3,157
SetB_24	24	24,444,170	27,387,620	24,444,170	27,387,620	0	0
SetB_25	24	28,845,341	29,848,829	28,845,341	29,848,829	0	0
SetB_26	26	3,008,718	11,914,284	3,008,718	11,914,284	0	0
SetB_27	30	38,126,268	38,689,821	38,126,268	38,689,821	0	0

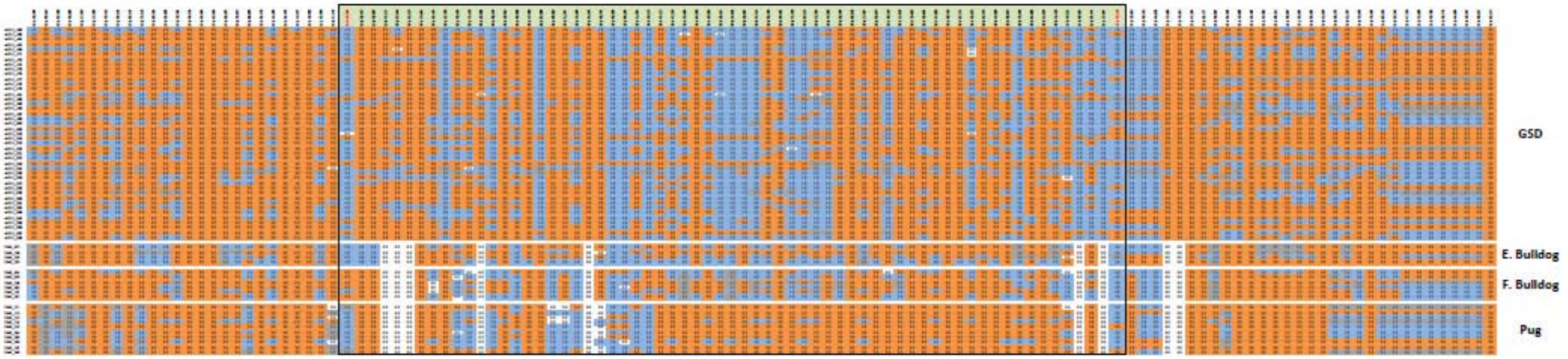
Additional file 4



Additional file 5



Additional file 6



Additional file 9