ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi:10.1111/j.1365-2052.2010.02132.x

Assessment of the functionality of genome-wide canine SNP arrays and implications for canine disease association studies

X. Ke*,†, L. J. Kennedy*, A. D. Short*, E. H. Seppälä‡, A. Barnes§, D. N. Clements¶, S. H. Wood*,§, S. D. Carter§, G. M. Happ, H. Lohi‡ and W. E. R. Ollier***

*Centre for Integrated Genomic Medical Research, School of Medicine, University of Manchester, Oxford Road, Manchester, UK. †Institute of Child Health, University College London, 30 Guilford Street, London, UK. ‡Department of Veterinary Biosciences, Department of Medical Genetics and Program in Molecular Medicine, Folkhälsan Research Center, Haartmaninkatu 8, 00014 University of Helsinki, Finland. §Department of Veterinary Pathology, Faculty of Veterinary Science, University of Liverpool, Liverpool, UK. ¶Royal (Dick) School of Veterinary Studies, Division of Veterinary Clinical Sciences, The University of Edinburgh, Hospital for Small Animals, Easter Bush Veterinary Centre, Roslin, Midlothian, Scotland, EH25 9RG, UK. **Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK, USA

Summary

Domestic dogs share a wide range of important disease conditions with humans, including cancers, diabetes and epilepsy. Many of these conditions have similar or identical underlying pathologies to their human counterparts and thus dogs represent physiologically relevant natural models of human disorders. Comparative genomic approaches whereby disease genes can be identified in dog diseases and then mapped onto the human genome are now recognized as a valid method and are increasing in popularity. The majority of dog breeds have been created over the past few hundred years and, as a consequence, the dog genome is characterized by extensive linkage disequilibrium (LD), extending usually from hundreds of kilobases to several megabases within a breed, rather than tens of kilobases observed in the human genome. Genome-wide canine SNP arrays have been developed, and increasing success of using these arrays to map disease loci in dogs is emerging. No equivalent of the human HapMap currently exists for different canine breeds, and the LD structure for such breeds is far less understood than for humans. This study is a dedicated large-scale assessment of the functionalities (LD and SNP tagging performance) of canine genome-wide SNP arrays in multiple domestic dog breeds. We have used genotype data from 18 breeds as well as wolves and coyotes genotyped by the Illumina 22K canine SNP array and Affymetrix 50K canine SNP array. As expected, high tagging performance was observed with most of the breeds using both Illumina and Affymetrix arrays when multi-marker tagging was applied. In contrast, however, large differences in population structure, LD coverage and pairwise tagging performance were found between breeds, suggesting that study designs should be carefully assessed for individual breeds before undertaking genome-wide association studies (GWAS).

Keywords canine, genome-wide SNP array, linkage disequilibrium, population structure, tagging.

Introduction

Canine genomic structure and genetic diversity have largely been shaped by at least two previous population bottlenecks. It is estimated that the first one occurred approximately 15 000 years ago when Grey Wolves were domesticated, while the more recent one is related to the creation of modern dog breeds over the past few hundred years. As a consequence, the dog genome is characterized by long stretches of regions exhibiting high linkage disequilibrium (LD) and low haplotype diversity. Within a breed, LD usually extends from hundreds of kilobases to several megabases, rather than the tens of kilobases observed in humans (Lindblad-Toh *et al.* 2005). Domestic dogs are also known to develop a similar spectrum of diseases to humans, including cancers, diabetes and epilepsy (Loscher *et al.* 1985; Khanna *et al.* 2006; Gershwin 2007). Many canine diseases have

Address for correspondence

X. Ke, MRC Centre of Epidemiology for Child Health, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK.
E-mail: xke@ich.ucl.ac.uk

Accepted for publication 21 June 2010

similar or identical underlying pathologies to their human counterparts and thus represent physiologically relevant natural models of human disorders (Chase *et al.* 2005). Furthermore, because individual dog breeds have been established over a relatively short period of time, they may be associated with a high prevalence of particular diseases. For example, approximately 15% of Golden Retrievers in the United States are affected with haemangiosarcomas (Glickman *et al.* 1999). Comparative genomic studies whereby susceptibility genes can be identified in dog diseases and then mapped onto the human genome are becoming increasingly popular and informative (Sutter & Ostrander 2004; Lindblad-Toh *et al.* 2005; Karlsson & Lindblad-Toh 2008; Andersson 2009).

Genome-wide association studies have been successfully used to identify disease susceptibility loci in humans (The Wellcome Trust Case Control Consortium. 2007; Thomson *et al.* 2007; Zeggini *et al.* 2007; Altshuler *et al.* 2008; Amos *et al.* 2008). As a high level of genetic homogeneity exists within dog breeds, the number of SNPs required to cover the whole genome for dogs is much lower than for humans (e.g. Affymetrix 50K canine SNP array vs. Genechip 500K of Affymetrix for human studies). At the same time, LD across breeds is much shorter, making approaches using multiple dog breeds ideal for fine-mapping purposes (Lindblad-Toh *et al.* 2005). The high risk of specific diseases in particular dog breeds means that considerably fewer samples are needed to map canine disease loci than for human studies. High penetrance Mendelian traits have already been mapped in dogs with approximately 10 cases vs. 10 controls (Karlsson *et al.* 2007), and it was estimated that mapping of a fivefold risk allele for a polygenic canine trait needs only 100 cases and 100 controls (Lindblad-Toh *et al.* 2005; Karlsson & Lindblad-Toh 2008). These advantages have made canine genome-wide association studies attractive tools to reveal disease risk loci, not only for dog breeds, as these studies also potentially highly relevant comparative models for analogous human conditions. Indeed, there are an increasing number of reports of successful mapping of important traits in dogs using this strategy (Karlsson *et al.* 2007; Salmon Hillbertz *et al.* 2007; Drögmüller *et al.* 2008). Large international collaborations have now been established to use this strategy to identify disease risk loci for many complex canine diseases, e.g. the LUPA project (http://www.eurolupa.org/).

Although it is known that there is generally a very high level of LD within individual dog breeds, genome-wide LD differences between individual breeds are less clear. The fact that there is no canine equivalent of the human HapMap for different dog breeds makes it difficult to accurately conduct such an assessment. The current canine map with about 2.5 million SNPs is based on a high-quality genome sequence assembly for the Boxer dog, a 1.5 × survey sequence of a Standard Poodle, 100 000 sequence reads for nine dogs from nine breeds, and 20 000 sequence reads from each of four Grey Wolves and one Coyote (Lindblad-Toh *et al.* 2005). It is therefore not surprising that the canine genome-wide SNP arrays designed so far have been selected to cover the genome evenly based on physical distance. Understanding the LD block coverage and tagging capability of these canine SNP arrays in individual breeds has important implications for the study design of genome-wide association studies. Previous studies on canine LD structure have largely focused on specific regions of the genome with a small number of markers (Sutter *et al.* 2004; Lindblad-Toh *et al.* 2005; Gray *et al.* 2009). Although Karlsson *et al.* (2007) have described the LD patterns in domestic dog breeds using the Affymetrix 27K SNP array, the emphasis was on the LD difference between within-breed and across-breed comparisons. In this report, a total of 18 dog breeds and other canids (wolves and coyotes) were genotyped by the Illumina 22K and Affymetrix 50K arrays to investigate the LD difference between breeds and the tagging performance of these arrays in different breeds, as well as to investigate the population structure, as revealed by the genome-wide SNP data.

Materials and methods

Samples and genotyping

A total of 775 DNA samples, from 18 domestic dog breeds, from Finland, the United Kingdom and the United States (see Table 1) were genotyped. A further 34 DNA samples, from other canid species, including 10 Coyotes, 12 Ethiopian Wolves and 12 Grey Wolves were also genotyped. All the samples were genotyped by either the Illumina 22K (22 362 SNPs) SNP array or Affymetrix 50K (49 663 SNPs) SNP array (Table 1).

The Coyote samples came from New Mexico, USA, while the Grey Wolves were taken from several isolated populations in Alaska, USA, and Northern Canada (Kennedy *et al.* 2007). EDTA blood samples from Ethiopian Wolves were collected as part of a rabies vaccination programme in Bale Mountain National Park, Ethiopia (Knobel *et al.* 2008).

Quality control of sample data

Two tiers of data quality control were employed. At the study-wise level, for genotype data generated by both the Illumina and Affymetrix arrays this consisted of 75% or higher genotyping success rate across all samples for a SNP and 85% or higher genotyping success rate across SNPs for each sample. For Illumina data, 20 947 SNPs in 600 of all 616 samples remained, whereas for the Affymetrix data 36 284 SNPs in 183 of the initial 193 samples remained (Table 1). At the individual analysis level, conducted within individual breeds, genotyping success rate of 85% (at sample and SNP level) and MAF threshold at 1% were used as the criterion for the Illumina and Affymetrix data, unless otherwise specified.

Table 1 Samples used in this study.

Samples	Before QC	After QC	Genotyping	Origin	Cases (Before/After QC)
Coyote	10	9	Illumina	USA	
Ethiopian wolf	12	12	Illumina	Africa	
Grey wolf	12	12	Illumina	Alaska	
Bedlington terrier	21	17	Illumina	Finland	
Bichon havanese	10	9	Illumina	Finland	
Boxer	36	35	Illumina	UK	9/8
Dobermann	62	60	Illumina	USA	25/25
German shepherd dog	60	59	Illumina	UK	25/25
Golden retriever	42	40	Illumina	UK	23/21
Labrador retriever	53	52	Illumina	UK	25/24
Nova scotia duck tolling retriever (NSDTR)	94	94	Illumina	Finland	
Rhodesian ridgeback	60	59	Illumina	USA	25/25
Samoyed	28	28	Illumina	Finland	
Cocker spaniel	45	38	Illumina	UK	20/17
Springer spaniel	39	38	Illumina	UK	15/15
Swedish vallhund	31	31	Illumina	Finland	
Australian shepherd dog	64	62	Affymetrix	Finland	
Border terrier	36	33	Affymetrix	Finland	
Brazilian terrier	10	8	Affymetrix	Finland	
Dobermann	34	34	Affymetrix	Finland	
Finnish hound	16	16	Affymetrix	Finland	
Schipperke	33	30	Affymetrix	Finland	

LD block analysis

Four gamete tests were used to define LD block structure across the genome of each breed (Wang *et al.* 2002; Lindblad-Toh *et al.* 2005). Briefly, a region was regarded as a haplotype block if 1–3 distinct haplotypes were observed for any pairwise combination of all the SNPs in the region. Blocks were searched from the start of a region by sequential addition of the next SNP to test the above condition. When the condition was no longer met, the block was defined to be terminated at the previous SNP (Wang *et al.* 2002). Haplotype phase was established using the snphap program (http://www-gene.cimr.cam.ac.uk/clayton/software/), and haplotypes with frequency <1% were disregarded. Neighbouring blocks may merge if $D' > 0.90$ was between them, as described by Lindblad-Toh *et al.* (2005). LD block coverage was defined as the proportion of the genome, in terms of physical distance, covered by LD blocks. To make LD structure comparable between different breeds, equal sample sizes were used. This was carried out by random sampling of the original sample sets without replacement. Estimates of LD block coverage were the averages of five repeated experiments.

Population structure analysis

Genetic distance (D) between individuals within breeds was calculated using the distance measure introduced by Bowcock *et al.* (1994) and modified by Kijas *et al.* (2009). First, Plink (Purcell *et al.* 2007) was used to calculate the average proportion of alleles shared (Dst) for each pairwise combination. From the Dst value, the corresponding D value was derived as 1-Dst. The D value for a breed was the average of the D values of pairwise combinations of individuals within the breed (Purcell *et al.* 2007). Genomic inflation factor (based on median chi-squared results) and multidimensional scaling (MDS) analyses were carried out using Plink (Purcell *et al.* 2007). For D value estimation and MDS analysis, SNPs were pruned with Plink to obtain a list of independent SNPs (-indep 50 5 2).

SNP tagging analysis

We used a SNP tagging strategy to assess the coverage of Illumina and Affymetrix arrays on canine genomes using the same sets of samples as generated for the LD analysis. Pairwise Pearson's correlation coefficients (r^2) were calculated for all SNPs in the arrays, using Plink (Purcell *et al.* 2007) with a tagging window size set at 100 SNPs and 10 Mb either side of each SNP and without r^2 threshold. A 'leave one out' tagging strategy was then employed (Ke *et al.* 2004). Briefly, each of the SNPs in the array was assumed to be 'nongenotyped' while other SNPs in the array were used to detect this 'nongenotyped' SNP. A 'nongenotyped' SNP was regarded as captured if the maximum r^2 between it and all other SNPs was no less than a predefined threshold (e.g. 0.80). The percentage of SNPs captured this way were regarded as the pairwise tagging capability of the

entire array. For haplotype tagging assessment, a sliding window of three SNPs was used for all 38 canine autosomes. The tagging capability of the SNP array for a particular SNP was calculated in the same way as for pairwise tagging, except that pairwise r^2 was replaced by haplotype r^2 between that SNP and the 3-marker haplotypes in individual sliding windows. Tagging was performed only on SNPs located in the same chromosomes, and no cross-chromosome tagging was carried out.

Results

LD difference between breeds

Linkage disequilibrium block coverage was investigated primarily using the multi-breed samples genotyped by the canine Illumina 22K array. To assess the LD block coverage of canine breeds, the effect of sample size was investigated first for dog breeds that had over 50 individuals. Random samples of 9, 20, 30, 40 and 50 were produced for each breed. LD block coverage was unstable and generally inflated at low sample sizes (Fig. 1a). The genome-wide LD became more conservative with increased sample size and tended to stabilize when 20 or more samples were used. This same sample size effect was also observed when the tagging capability of the genome-wide canine array was examined (data not shown). These observations suggest that to avoid unstable LD structure in the study population, a minimum sample size of 20 for both cases and controls is needed in association studies.

Large variability existed between dog breeds and other canid species in terms of LD block coverage, and this variation was consistent at different sample sizes (Fig. 1b). When all canids were examined, the Boxers, Bedlington Terriers and Swedish Vallhunds were found to have the highest LD block coverage, while the Grey Wolves and Coyotes were associated with the lowest block coverage (Fig. 1b). Among dog breeds, the Springer Spaniels and Labrador Retrievers exhibited the lowest LD block coverage (Fig. 1b).

The low LD block coverage of the Grey Wolves and Coyotes was consistent with the low number of polymorphic markers of the SNP array in their samples – 66.5% for Grey Wolves ($n = 12$) and 44.2% for Coyotes ($n = 9$), when compared to 83.5% for Bichon Havaneses ($n = 9$). The lower proportion of polymorphic SNPs in Coyotes was expected, as coyotes are genetically more distant from domestic dogs than wolves. The low LD of the Wolves and Coyotes was overall consistent with their large population sizes in general and, therefore, they were expected to harbour more ancestral chromosomes. It was surprising that, although the Ethiopian Wolves had a similar number of polymorphic markers (64.6%, $n = 12$) to the Grey Wolves, a much higher level of LD was observed. We believe that this observation could be study-specific, as the Ethiopian

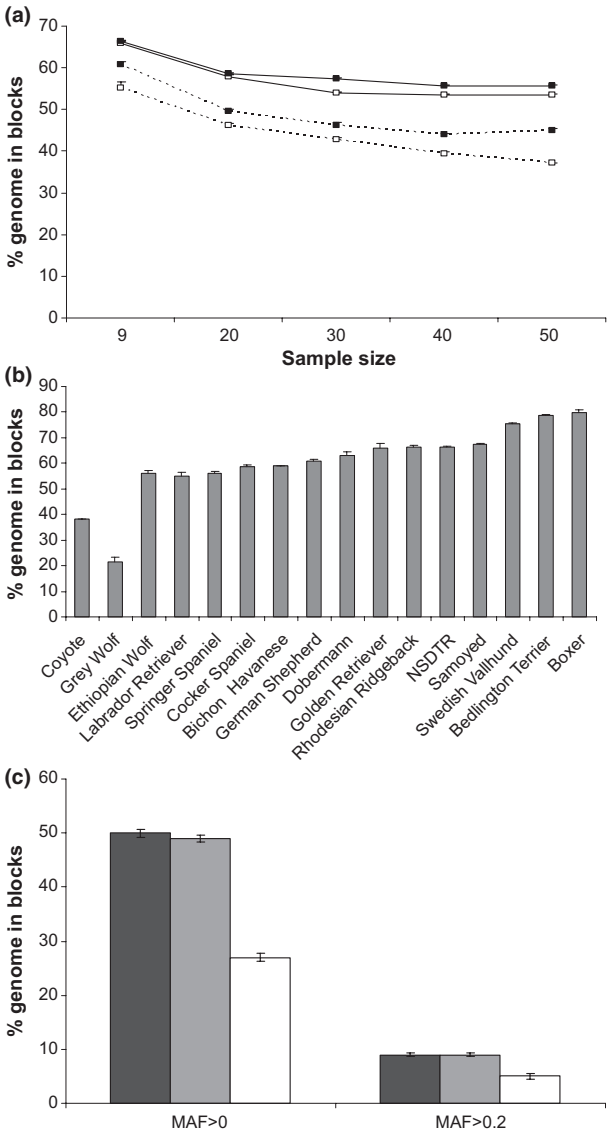

Figure 1 Linkage disequilibrium (LD) structure in canids. (a) Effect of sample size on LD structure. Nova Scotia Duck Tolling Retrievers (NSDTR, denoted by solid squares on solid line), Rhodesian Ridgebacks (white squares on solid line), German Shepherd Dogs (solid squares on dashed line) and Labrador Retrievers (white squares on dashed line); random sample sizes of 9, 20, 30, 40 and 50 were produced for each breed. These sample sets were then subjected to LD block analysis. (b) LD difference between breeds. For breeds that had more than 9 individual dogs, random sample sizes of nine were produced for each breed. These sample sets were then subjected to LD block analysis. (c) Effect of SNP selection and density on LD structure. For the Boxers, Springer Spaniels and Swedish Vallhunds, respectively, random samples of 20 were produced for each breed. A shared marker set was generated, between a random dataset from each of the three breeds, containing either polymorphic SNPs or SNPs with MAF ≥ 0.2 in all three only. The shared marker sets were then subjected to LD block analysis. For five repeat experiments, the average number of SNPs in the shared marker set was 10 985 with MAF ≥ 0 (5.1 SNPs/Mb) and 2633 with MAF ≥ 0.2 (1.2 SNPs/Mb). Block coverage (% genome in blocks) was defined according to physical sizes, and results were the averages of five repeated experiments, with standard errors shown on the error bars.

Wolf samples in this study were collected from a single isolated population. The higher level of LD coverage observed in Coyotes than in the Grey Wolves, despite a lower proportion of polymorphic SNPs, could be because of sampling bias, and the result should be viewed with caution.

The discrepancy of polymorphic SNP coverage in the different canids and dog breed collections raised a more important question, i.e. whether the LD difference observed between breeds (Fig. 1b) was largely because of SNP density and the related SNP selection scheme. To address this question, a maximum set of SNP markers from the Illumina array data were selected that were polymorphic in datasets (sample size at 20) of each of the following three breeds: Boxer, Swedish Vallhund (both had a very high LD block coverage), and Springer Spaniel (which had a very low LD block coverage). In the full marker sets, Boxers, Swedish Vallhunds and Springer Spaniels had a SNP density at 7.4, 8.5 and 6.8 SNPs/Mb, respectively, whereas in the shared marker set, the density was reduced to 5.0 SNPs/Mb. As a result of reduced SNP density, the LD block coverage (sample size at 20) was also reduced in all the three breeds with the shared marker set than with the full marker sets: Boxers from 70% to 50%; Swedish Vallhunds from 67% to 49%; and Springer Spaniels from 45% to 27%. What is more interesting is the observation that, at the same marker density, LD block coverage of the Springer Spaniels was significantly lower than the Boxers and Swedish Vallhunds, a pattern observed at the full marker sets. Rare variants are known to be able to inflate LD, but the LD difference between the three breeds remained even after the shared marker set was restricted to SNPs with MAF ≥ 0.2 (Fig. 1c). The results, therefore, demonstrated that LD difference between breeds was genuine rather than because of confounding effects from SNP density and SNP selection.

Similar between-breed variation of LD block coverage was also confirmed with samples genotyped by the Affymetrix array (data not shown). For two Dobermann sample collections, genotyped by the Illumina and Affymetrix array, using a sample size of 20, the block coverage was 52% and 61%, respectively. When only the 15 032 SNPs shared between the Illumina and Affymetrix arrays were used for the analysis, the block coverage was only slightly reduced (50% and 56%, respectively).

Population structure within dog breeds

It is known that some canine breeds are more homogeneous than others because of their breeding histories. It is also known that population structure is present in some dog breeds (Quignon *et al.* 2007; Björnerfeldt *et al.* 2008; Calboli *et al.* 2008; Chang *et al.* 2009). Genetic distance (Bowcock *et al.* 1994) between all pairwise combinations of individuals (D) was used as a measure of homogeneity of samples within each breed. Almost all of the dog breeds had an average D value between 0.19 and 0.28, with Bichon Havaneses having the highest D values. Coyotes and Grey Wolves were associated with a relatively high D value at 0.292 and 0.294, respectively, while Ethiopian Wolves were found to have a very low D value at 0.216, confirming their status as an extremely isolated population (Fig. 2).

There were two Dobermann collections, of which one was genotyped by the Illumina 22K array, and for the other the Affymetrix 50K array was used. Using a genotyping success rate of 85%, 15 032 SNPs were found to be shared between the Illumina 22K array and the Affymetrix 50K array.

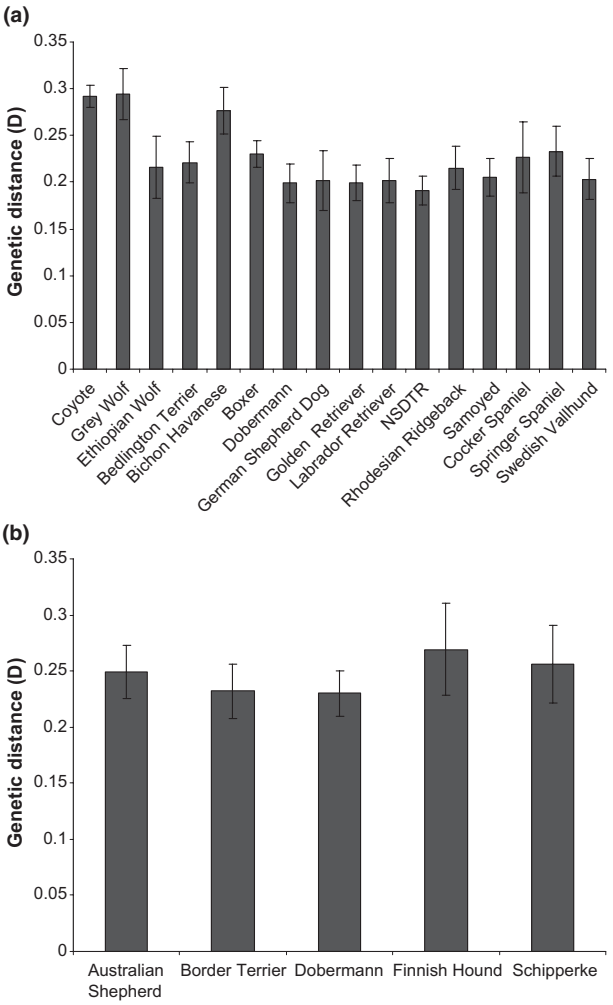

Figure 2 Genetic distance (D) between individuals within breeds. Plink was used to calculate the average proportion of alleles shared for each pairwise combination as well as the related Dst measure. From the Dst value, the corresponding D value was derived as 1-Dst. The D value for a breed was the average of the D values of all pairwise combinations of individuals with the breed. Only polymorphic SNPs were used and they were further pruned with Plink (-indep 50 5 2). A threshold of 90% genotyping success rate across all samples and all SNPs was imposed. (a) Samples genotyped with the Illumina 22K SNP array. (b) Samples genotyped with the Affymetrix 50K array. Data for Brazilian Terrier were excluded, as there were only four individuals left after QC. Error bars show the standard deviations of the estimates.

Again, this set of SNPs was pruned using the same criteria (plink–indep 50 5 2) as for the full dataset. The estimate of D was 0.202 and 0.236, respectively, almost exactly the same as using the full array data (0.199 vs. 0.230). This demonstrated that the effect of SNP array on estimated genetic distance was minimal and that the pruned set of SNPs in the shared SNP set was sufficient to distinguish the genetic distances.

The average within-breed genetic distance (D) in domestic dogs found in this study was 0.211, at a similar level as in cattle (0.210) (Kijas *et al.* 2009), and lower than the within-breed D values of sheep (0.254) (Kijas *et al.* 2009). As noted earlier, however, this figure is expected to vary across different breed collections.

One important practical issue relating to population structure is its impact on case–control disease association studies. Many of the samples used in this study were collected for such purposes, and it was found that the case–control samples of most of the breeds were associated with high values of genomic inflation factor (based on median chi-squared results) (Fig. 3a), indicating potential problems of population structure and stratification. MDS plots of the first two dimensions confirmed the presence of population substructure within Cocker Spaniel samples (Fig. 3b), which was associated with the highest inflation factor among the samples (Fig. 3a). More specifically, it was observed that between the three sub-clusters in the Cocker Spaniel samples (Fig. 3b), sub-cluster A members were predominantly controls (9/10, or 90%), sub-cluster B members were predominantly cases (7/10, or 70%), whereas sub-cluster C members were mixed at perfect balance (9/18, or 50% for both cases and controls). Such a clustering pattern of samples can clearly lead to population stratification problems and false positives in a case–control association study. Similar population stratification problems were also observed in the case–control samples of other breed collections to a lesser degree (data not shown). To correct for population stratification problems, Cochran-Mantel-Haenszel (CMH) meta-analysis was applied to combine evidence between clusters formed using the genome-wide identity by state (IBS) information, as implemented in Plink (Purcell *et al.* 2007).

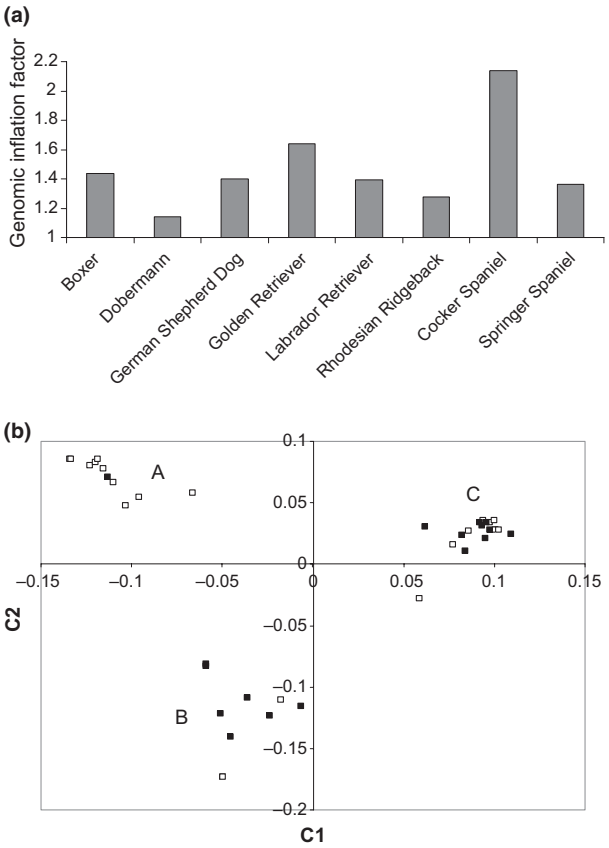

Figure 3 Population structure in case–control association study samples. (a) Genomic inflation factor estimation (based on median chi-squared results of the association test) in real disease case–control data. Boxers 8 cases vs. 27 controls; Dobermanns 25 cases vs. 35 controls; German Shepherd Dogs 25 cases vs. 34 controls; Golden Retrievers 21 cases vs. 19 controls; Labrador Retrievers 24 cases vs. 28 controls; Rhodesian Ridgebacks 25 cases vs. 34 controls; Cocker Spaniels 17 cases vs. 21 controls; and Springer Spaniels 15 cases vs. 23 controls. (b) Multidimensional scaling (MDS) plot of case (solid squares) and control (white squares) samples of the Cocker Spaniels. *X*-axis is the value of the 1st MDS principal component and *y*-axis is the 2nd component.

Power and coverage of genome-wide arrays

The success of genetic association studies relies on the presence of LD between causal variants and SNP markers that are directly genotyped. In other words, the genome-wide coverage of LD by a genotyped marker set is an important indicator of successful outcome in identifying causal variants in a case–control study (The International HapMap Consortium 2005; Barrett & Cardon 2006). Here, we adopted a 'leave one out' tagging strategy to examine the capability of canine genome-wide SNP arrays to capture nongenotyped SNPs in the genome (see Materials and Methods). Tagging was only performed for SNPs on the same chromosomes.

When pairwise tagging was used, a very large difference was observed across breeds for the tagging capability of both the Illumina (Fig. 4a) and Affymetrix (Fig. 4b) arrays. With the Illumina array, reasonable tagging power (proportion of SNPs captured) was obtained for the Boxers, but coverage for the Labrador Retrievers, Springer Spaniels and Cocker Spaniels was particularly low. The Affymetrix array has more than twice the number of SNPs when compared to the Illumina array. As a consequence, the tagging power of the Affymetrix array was much higher than the Illumina array, as illustrated by the Dobermann data (Fig. 4a,b). For the set of 15 032 SNPs shared by the two arrays, the within-set tagging power, (i.e. when the 'nongenotyped SNPs' were drawn from within the set) was 50% and 57% for the Illumina and Affymetrix samples, respectively. The power

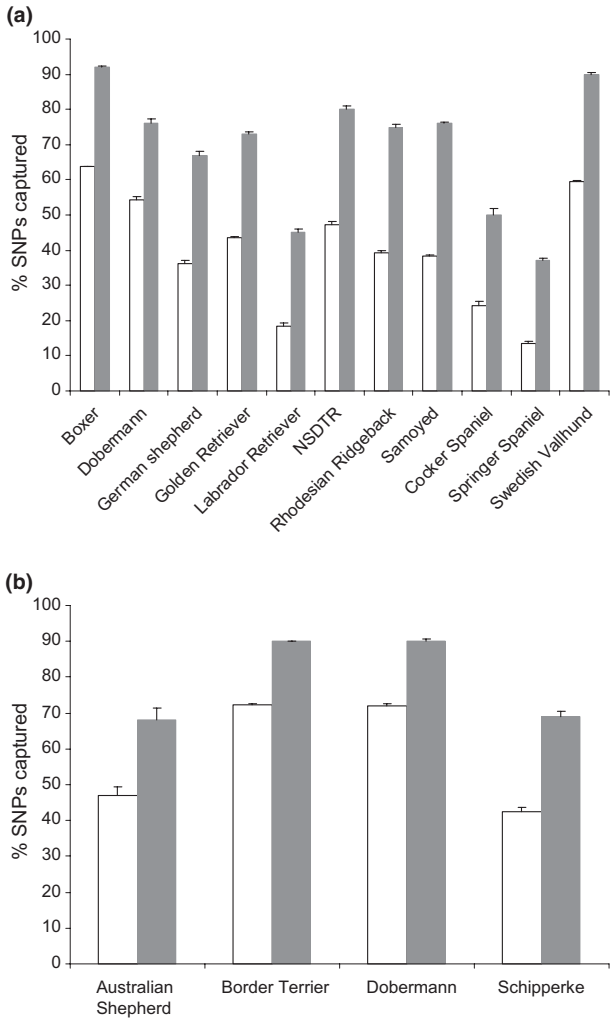

Figure 4 Tagging capability and genome-wide coverage of Illumina 22K and Affymetrix 50K arrays. A 'nongenotyped' SNP was regarded as captured if the maximum r^2 between it and other SNPs in the array was ≥0.8. White bars denote pairwise tagging where each 'nongenotyped' SNP has a window of 100 SNPs and 10 Mb either side; grey bars denote results of multi-marker haplotype tagging where a sliding window of three consecutive SNPs in the same chromosome as the 'nongenotyped' SNP was used. The sample size was 20 in each sample set and only polymorphic SNPs were used. (a) Breeds genotyped by the Illumina 22K array. (b) Breeds genotyped by the Affymetrix 50K array. Results were the averages of five repeated experiments, with standard errors shown on the error bars.

dropped to 48% and 50%, respectively, when these SNPs in the shared set were used as the only tagging SNPs to tag other SNPs in the full array.

Multi-marker haplotypes are often used in place of single SNPs to increase power and coverage (de Bakker *et al.* 2006; Karlsson *et al.* 2007). In this study, sliding windows of three SNPs across the genome were used to examine the efficiency of pairwise tagging vs. haplotype tagging. In all cases, tagging efficiency (and thus genome-wide coverage of both arrays) was dramatically increased by the use of haplotype tagging (Fig. 4a,b). For the Boxers and Swedish Vallhunds genotyped using the Illumina array and the Border Terriers and Dobermanns genotyped using the Affymetrix array, approximately 90% of 'nongenotyped' SNPs can be captured. For the Springer Spaniels, however, coverage remained low, with <50% of 'nongenotyped' SNPs captured by the Illumina array.

Discussion

This study is a dedicated large-scale assessment of the functionalities of canine genome-wide SNP arrays in domestic dog breeds and wild canids. The two popular canine genome-wide SNP arrays (Illumina 22K SNP array and Affymetrix 50K SNP array) were investigated for their capability to detect potential causal variants in the context of a genome-wide case–control association study. The results demonstrated that this capability could vary from breed to breed and that such differences should be carefully considered at the study design stage. Although the current study did not provide a detailed and fine-level picture of LD structure for individual regions across the genome because of lack of equivalents of a HapMap for the individual canid populations and the relatively low SNP density of the arrays, we believe that knowledge of LD difference between breeds gained from the present study is important, as has been previously demonstrated for human studies (Gabriel *et al.* 2002; The International HapMap Consortium 2005). Unlike in human SNP arrays, where SNPs have been generally validated across platforms, SNP genotype calling inaccuracy may be a problem for the canine SNP arrays. Such data quality problems can affect the accuracies of LD block boundaries and sizes. In this study, it was the relative difference of LD block coverage and tagging performance of the SNP arrays, rather than the exact block structure, that were of primary interest. We believe that such between-breed comparisons were valid on the assumption that the data quality problems were limited and common to all breeds.

First, the results demonstrated that the two genome-wide SNP arrays showed reasonable LD coverage for most of the domestic dog breeds examined. The majority of SNPs (>65%) in the Illumina array were also found to be polymorphic in the Grey Wolf and Ethiopian Wolf samples, demonstrating the potential usage of such arrays in other canids. As expected, the figure was lower for Coyotes, which represent an out-bred population. Significant differences in LD were observed between domestic dog breeds, for example, the Boxers, Bedlington Terriers and the Swedish Vallhunds were found to have the highest LD block coverage, while the Springer Spaniels and Labrador Retrievers were among breeds having the lowest LD block coverage. LD differences between breeds have been reported in previous studies where specific regions were selected for SNP discovery and high density genotyping (Sutter *et al.* 2004; Gray *et al.* 2009). The results here further demonstrate that large differences in LD exist at a genome-wide level.

As a consequence of LD variability from breed to breed, the tagging capacity and genome-wide coverage of both the Illumina 22K and Affymetrix 50K SNP arrays varied widely across breeds. It should be noted that high tagging capacity can be achieved by using multi-marker haplotype tagging by the two SNP arrays for most of the breeds examined. The result suggested that in case–control studies, larger sample sizes are needed for breeds with lower levels of LD to maintain similar levels of power to those studies done with breeds with higher levels of LD. An important approach to improve power is to increase SNP density in a genome-wide SNP array, as observed in this study. New canine SNP arrays, such as the recently introduced 185K Illumina array, will undoubtedly make significant contributions not only in improving power for individual genome-wide case–control association studies but also in understanding LD structure at a finer level than is currently possible. The high density array may also make it possible to impute SNPs for association studies, for example for SNPs that were not present in the Illumina 22K and Affymetrix 50K arrays.

Although many dog breeds are generally of very recent origin (50–100 generations) and tend to be highly inbred, hidden population structure in dog breeds has recently been reported by analysis of pedigree structure (Calboli *et al.* 2008), investigation into microsatellite genotypes on poodles (Björnerfeldt *et al.* 2008), and genome-wide SNP genotype data on four dog breeds (Chang *et al.* 2009). Our present study confirmed the presence of varying degrees of population structure within dog breeds. The presence of population structure has several important implications. First, it should be acknowledged that the LD and tagging analysis in this study was confounded to some extent by the presence of within-breed population structures. This study therefore serves only as a crude survey of LD and tagging power of the two genome-wide SNP arrays in dog breeds. However, such surveys are still important, as these were representative of the sample collections commonly obtained by individual research groups, and study designs are often based on information from such collections. Second, the impact of population structure on case–control association studies should be carefully investigated and controlled. Population structure and stratification has long been regarded as an important confounder for genetic association studies (Cardon & Palmer 2003; Balding 2006; Amos 2007). Although correction of population stratification can be achieved using approaches such as genomic control (Bacanu *et al.* 2002), principal component analysis (Price *et al.* 2006), multidimensional scaling (Miclaus *et al.* 2009) and logistic regression (Setakis *et al.* 2006), this often comes at the cost of reduced power (e.g. because of extreme imbalance of case vs. controls in a cluster or missing values of covariates). Other efforts can be put into practice during the study design stage. Geographical location and breed selection regimes have been demonstrated to be important sources of population substructure (Quignon *et al.* 2007). Rather than resorting to adjustment of such structural covariates in a logistic regression model during the data analysis stage, it is generally preferable to have as balanced matching as possible of cases and controls for these factors at the outset of a study. Knowledge of LD for breed collections, broken down to the substructure, will undoubtedly also be very valuable for both study design and data analysis.

Finally, it should be noted that besides SNPs, other types of polymorphisms, such as copy number variations (CNVs), can also play important roles in disease aetiology and complex traits, as reported for humans (Estivill & Armengol 2007; De Cid *et al.* 2009). Recent investigations into structure variations and associated copy number variations in domestic dogs (Chen *et al.* 2009; Nicholas *et al.* 2009) have not only improved our understanding about the mechanisms of canine genome evolution, but have also provided valuable resources for future canine genome-wide array design to improve both coverage and power for canine disease studies.

Acknowledgements

We thank: Joseph A. Cook, Professor and Curator of Mammals, University of New Mexico for providing the Coyote samples; Karen Laurenson (Frankfurt Zoological Society, Regional Office in the Serengeti) for the Ethiopian Wolf samples and Lindsey Carmichael (Department of Biological Sciences, University of Alberta, Edmonton, Canada) for some of the Grey Wolf samples. Part of the UK work was supported by American Kennel Club Canine Health Foundation grants: A-0859, A-0871 and A-0876. The Finnish study was partly supported by the Sigrid Juselius Foundation, Biocentrum Helsinki, the Academy of Finland, and the Jane and Aatos Erkko Foundation and partly funded by research grants of Dr. Leena Peltonen and the Center of Excellence of Complex Disease Genetics of the Academy of Finland.

References

Altshuler D., Daly M.J. & Lander E.S. (2008) Genetic mapping in human disease. *Science* **322,** 881–8.

Amos C.I. (2007) Successful design and conduct of genome-wide association studies. *Human Molecular Genetics.* **16:** R220–5.

Amos C.I., Wu X., Broderick P. *et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40,** 616–22.

Andersson L. (2009) Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. *Genetica* **136,** 341–9.

Bacanu S., Devlin B. & Roeder K. (2002) Association Studies for Quantitative Traits in Structured Populations. *Genetic Epidemiology,* **22,** 78–93.

de Bakker P.I., Burtt N.P., Graham R.R., Guiducci C., Yelensky R. *et al.* (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genetics* **38,** 1298–303.

Balding D.J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics* **7,** 781–91.

Barrett J.C. & Cardon L.R. (2006) Evaluating coverage of genome-wide association studies. *Nature Genetics* **38,** 659–62.

Björnerfeldt S., Hailer F., Nord M. & Vilà C. (2008) Assortative mating and fragmentation within dog breeds. *BMC Evolutionary Biology* **8,** 28.

Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R. & Cavalli-Sforza L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368,** 455–457.

Calboli F.C.F., Sampson J., Fretwell N. & Balding D.J. (2008) Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics.* **179,** 593–601.

Cardon L.R. & Palmer L.J. (2003) Population stratification and spurious allelic association. *Lancet* **361,** 598–604.

Chang M.L., Yokoyama J.S., Branson N., Dyer D.J., Hitte C., Overall K.L. & Hamilton S.P. (2009) Intrabreed stratification related to divergent selection regimes in purebred dogs may affect the interpretation of genetic association studies. *Journal of Heredity.* **100**(Suppl. 1): S28–36.

Chase K., Lawler D.F., Carrier D.R. & Lark K.G. (2005) Genetic regulation of osteoarthritis: a QTL regulating cranial and caudal acetabular osteophyte formation in the hip joint of the dog (*Canis familiaris*). *American Journal of Medical Genetics* **135A,** 334–5.

Chen W.K., Swartz J.D., Rush L.J. & Alvarez C.Z. (2009) Mapping DNA structural variation in dogs. *Genome Research* **19,** 500–9.

De Cid R., Riveira-Munoz E., Zeeuwen P.L., Robarge J., Liao W. *et al.* (2009) Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nature Genetics* **41,** 211–5.

Drögmüller C., Karlsson E.K., Hytönen M.K., Perloski M., Dolf G. *et al.* (2008) A mutation in hairless dogs implicates *FOXI3* in ectodermal development. *Science* **321,** 1462.

Estivill X. & Armengol L. (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics* **3,** 1787–99.

Gabriel S.B., Schaffner S.F., Nguyen H. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–9.

Gershwin L.J. (2007) Veterinary autoimmunity: autoimmune diseases in domestic animals. *Annals of the New York Academy of Sciences* **1109,** 109–16.

Glickman L., Glickman N. & Thorpe R. (1999) The Golden Retriever Club of America National Health Survey 1998-1999. http://www.grca.org/pdf/health/healthsurvey.pdf.

Gray M.M., Granka J., Bustamante C.D., Sutter N.B., Boyko A.R., Zhu L., Ostrander E.A. & Wayne R.K. (2009) Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics.* **181,** 1493–505.

Karlsson E.K. & Lindblad-Toh K. (2008) Leader of the pack: gene mapping in dogs and other model organisms. *Nature Reviews. Genetics* **9,** 713–25.

Karlsson E.K., Baranowska I., Wade C.M., Salmon Hillbertz N.H., Zody M.C. *et al.* (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics* **39,** 1321–8.

Ke X., Durrant C., Morris A., Hunt S., Bentley D.R., Deloukas P. & Cardon L.R. (2004) Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Human Molecular Genetics* **13,** 2557–65.

Kennedy L.J., Angles J.M., Barnes A., Carmichael L.E., Radford A.D., Ollier W.E. & Happ G.M. (2007) *DLA-DRB1, DQA1,* and *DQB1* Alleles and Haplotypes in North American Gray Wolves. *Journal of Heredity* **98,** 491–9.

Khanna C., Lindblad-Toh K., Vail D., London C., Bergman P. *et al.* (2006) The dog as a cancer model. *Nature Biotechnology* **24,** 1065–6.

Kijas J.W., Townley D., Dalrymple B.P. *et al.* for the International Sheep Genomic Consortium. (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PlosOne.* **4,** e4668.

Knobel D.L., Fooks A.R., Brookes S.M., Randall D.A., Williams S.D., Argaw K., Shiferaw F., Tallents L.A. & Laurenson M.K. (2008) Trapping and vaccination of endangered Ethiopian Wolves to control an outbreak of rabies. *The Journal of Applied Ecology* **45,** 109–16.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., Karlsson E.K., Jaffe D.B. *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–19.

Loscher W., Schwartz-Porsche D., Frey H.H. & Schmidt D. (1985) Evaluation of epileptic dogs as an animal model of human epilepsy. *Arzneimittel-Forschung* **35,** 82–7.

Miclaus K., Wolfinger R. & Czika W. (2009) SNP selection and multidimensional scaling to quantify population structure. *Genetic Epidemiology* **33,** 488–96.

Nicholas T.J., Cheng Z., Ventura M., Mealey K., Eichler E.E. & Akey J.M. (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research* **19,** 491–9.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A. & Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–9.

Purcell S., Neal B., Todd-Brown K. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81,** 559–75.

Quignon P., Herbin L., Cadieu E., Kirkness E.F., Hédan B., Mosher D.S., Galiber F., André C., Ostrander E.A. & Hitte C. (2007) Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. *PLoS ONE* **2,** e1324.

Salmon Hillbertz N.H., Isaksson M., Karlsson E.K., Hellmen E., Pielberg G.R. *et al.* (2007) Duplication of *FGF3, FGF4, FGF19* and *ORAOV1* causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nature Genetics* **39,** 1318–20.

Setakis E., Stirnadel H. & Balding D.J. (2006) Logistic regression protects against population structure in genetic association studies. *Genome Research* **16,** 290–6.

Sutter N.B. & Ostrander E.A. (2004) Dog star rising: the canine genetics system. *Nature Reviews. Genetics* **5,** 900–10.

Sutter N.B., Eberle M.A., Parker H.G., Pullar B.J., Kirkness E.F., Kruglyak L. & Ostrander E.A. (2004) Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Research* **14,** 2388–96.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature.* **437,** 1299–320.

The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **477**, 661–78.

Thomson W., Barton A., Ke X., Eyre S., Hinks A. *et al.* (2007) Rheumatoid arthritis association at 6q23. *Nature Genetics* **39**, 1431–3.

Wang N., Akey J.M., Zhang K., Chakraborty R. & Jin L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics* **71**, 1227–34.

Zeggini E., Weedon M.N., Lindgren C.M., Frayling T.M., Elliott K.S. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–41.