ARTICLES

Genome sequence, comparative analysis and haplotype structure of the domestic dog

Kerstin Lindblad-Toh[1], Claire M Wade[1,2], Tarjei S. Mikkelsen[1,3], Elinor K. Karlsson[1,4], David B. Jaffe[1], Michael Kamal[1], Michele Clamp[1], Jean L. Chang[1], Edward J. Kulbokas III[1], Michael C. Zody[1], Evan Mauceli[1], Xiaohui Xie[1], Matthew Breen[5], Robert K. Wayne[6], Elaine A. Ostrander[7], Chris P. Ponting[8], Francis Galibert[9], Douglas R. Smith[10], Pieter J. deJong[11], Ewen Kirkness[12], Pablo Alvarez[1], Tara Biagi[1], William Brockman[1], Jonathan Butler[1], Chee-Wye Chin[1], April Cook[1], James Cuff[1], Mark J. Daly[1,2], David DeCaprio[1], Sante Gnerre[1], Manfred Grabherr[1], Manolis Kellis[1,13], Michael Kleber[1], Carolyne Bardeleben[6], Leo Goodstadt[8], Andreas Heger[8], Christophe Hitte[9], Lisa Kim[7], Klaus-Peter Koepfli[6], Heidi G. Parker[7], John P. Pollinger[6], Stephen M. J. Searle[14], Nathan B. Sutter[7], Rachael Thomas[5], Caleb Webber[8], Broad Institute Genome Sequencing Platform* & Eric S. Lander[1,15]

Here we report a high-quality draft genome sequence of the domestic dog (*Canis familiaris*), together with a dense map of single nucleotide polymorphisms (SNPs) across breeds. The dog is of particular interest because it provides important evolutionary information and because existing breeds show great phenotypic diversity for morphological, physiological and behavioural traits. We use sequence comparison with the primate and rodent lineages to shed light on the structure and evolution of genomes and genes. Notably, the majority of the most highly conserved non-coding sequences in mammalian genomes are clustered near a small subset of genes with important roles in development. Analysis of SNPs reveals long-range haplotypes across the entire dog genome, and defines the nature of genetic diversity within and across breeds. The current SNP map now makes it possible for genome-wide association studies to identify genes responsible for diseases and traits, with important consequences for human and companion animal health.

Man's best friend, *Canis familiaris*, occupies a special niche in genomics. The unique breeding history of the domestic dog provides an unparalleled opportunity to explore the genetic basis of disease susceptibility, morphological variation and behavioural traits. The position of the dog within the mammalian evolutionary tree also makes it an important guide for comparative analysis of the human genome.

The history of the domestic dog traces back at least 15,000 years, and possibly as far back as 100,000 years, to its original domestication from the grey wolf in East Asia[1–4]. Dogs evolved through a mutually beneficial relationship with humans, sharing living space and food sources. In recent centuries, humans have selectively bred dogs that excel at herding, hunting and obedience, and in this process have created breeds rich in behaviours that both mimic human behaviours and support our needs. Dogs have also been bred for desired physical characteristics such as size, skull shape, coat colour and texture[5], producing breeds with closely delineated morphologies. This evolutionary experiment has produced diverse domestic species, harbouring more morphological diversity than exists within the remainder of the family Canidae[6].

As a consequence of these stringent breeding programmes and periodic population bottlenecks (for example, during the World Wars), many of the ~400 modern dog breeds also show a high prevalence of specific diseases, including cancers, blindness, heart disease, cataracts, epilepsy, hip dysplasia and deafness[7,8]. Most of these diseases are also commonly seen in the human population, and clinical manifestations in the two species are often similar[9]. The high prevalence of specific diseases within certain breeds suggests that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans[10].

Genetic analysis of traits in dogs is enhanced by the close relationship between humans and canines in modern society.

[1]Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [3]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Program in Bioinformatics, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA. [5]Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. [6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA. [7]National Human Genome Research Institute, National Institutes of Health, 50 South Drive, MSC 8000, Building 50, Bethesda, Maryland 20892-8000, USA. [8]MRC Functional Genetics, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. [9]UMR 6061 Genetique et Developpement, CNRS—Université de Rennes 1, Faculté de Médecine, 2, Avenue Leon Bernard, 35043 Rennes Cedex, France. [10]Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. [11]Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, California 94609, USA. [12]The Institute for Genomic Research, Rockville, Maryland 20850, USA. [13]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [14]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [15]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.
*A list of participants and affiliations appears at the end of the paper.

Through the efforts of the American Kennel Club (AKC) and similar organizations worldwide, extensive genealogies are easily accessible for most purebred dogs. With the exception of human, dog is the most intensely studied animal in medical practice, with detailed family history and pathology data often available[8]. Using genetic resources developed over the past 15 years[11–16], researchers have already identified mutations in genes underlying ~25 mendelian diseases[17,18]. There are also growing efforts to understand the genetic basis of phenotypic variation such as skeletal morphology[10,19].

The dog is similarly important for the comparative analysis of mammalian genome biology and evolution. The four mammalian genomes that have been intensely analysed to date (human[20–22], chimpanzee[23], mouse[24] and rat[25]) represent only one clade (Euarchontoglires) out of the four clades of placental mammals. The dog represents the neighbouring clade, Laurasiatheria[26]. It thus serves as an outgroup to the Euarchontoglires and increases the total branch length of the current tree of fully sequenced mammalian genomes, thereby providing additional statistical power to search for conserved functional elements in the human genome[24,27–33]. It also helps us to draw inferences about the common ancestor of the two clades, called the boreoeutherian ancestor, and provides a bridge to the two remaining clades (Afrotheria and Xenarthra) that should be helpful for anchoring low-coverage genome sequence currently being produced from species such as elephant and armadillo[28].

Here we report a high-quality draft sequence of the dog genome covering ~99% of the euchromatic genome. The completeness, nucleotide accuracy, sequence continuity and long-range connectivity are extremely high, exceeding the values calculated for the recent draft sequence of the mouse genome[24] and reflecting improved algorithms, higher-quality data, deeper coverage and intrinsic genome properties. We have also created a tool for the formal assessment of assembly accuracy, and estimate that >99% of the draft sequence is correctly assembled.

We also report an initial compendium of SNPs for the dog population, containing >2.5 million SNPs derived primarily from partial sequence comparison of 11 dog breeds to a reference sequence. We characterized the polymorphism rate of the SNPs across breeds and the long-range linkage disequilibrium (LD) of the SNPs within and across breeds.

We have analysed these data to study genome structure, gene evolution, haplotype structure and phylogenetics of the dog. Our key findings include:

- The evolutionary forces molding the mammalian genome differ among lineages, with the average transposon insertion rate being lowest in dog, the deletion rate being highest in mouse and the nucleotide substitution rate being lowest in human.
- Comparison between human and dog shows that ~5.3% of the human genome contains functional elements that have been under purifying selection in both lineages. Nearly all of these elements are confined to regions that have been retained in mouse, indicating that they represent a common set of functional elements across mammals.
- Fifty per cent of the most highly conserved non-coding sequence in the genome shows striking clustering in ~200 gene-poor regions, most of which contain genes with key roles in establishing or maintaining cellular identity, such as transcription factors or axon guidance receptors.
- Sets of functionally related genes show highly similar patterns of evolution in the human and dog lineages. This suggests that we should be careful about interpreting accelerated evolution in human relative to mouse as representing human-specific innovations (for example, in genes involved in brain development), because comparable acceleration is often seen in the dog lineage.
- Analysis across the entire genome of the sequenced boxer and across 6% of the genome in ten additional breeds shows that linkage disequilibrium (LD) within breeds extends over distances of several megabases, but LD across breeds only extends over tens of kilobases. These LD patterns reflect two principal bottlenecks in dog history: early domestication and recent breed creation.
- Haplotypes within breeds extend over long distances, with ~3–5 alleles at each locus. Portions of these haplotypes, as large as 100 kilobases (kb), are shared across multiple breeds, although they are present at widely varying frequencies. The haplotype structure suggests that genetic risk factors may be shared across breeds.
- The current SNP map has sufficient density and an adequate within-breed polymorphism rate (~1/900 base pairs (bp) between breeds and ~1/1,500 bp within breeds) to enable systematic association studies to map genes affecting traits of interest. Genotyping of ~10,000 SNPs should suffice for most purposes.
- The genome sequence can be used to select a small collection of rapidly evolving sequences, which allows nearly complete resolution of the evolutionary tree of nearly all living species of Canidae.

Generating a draft genome sequence

We sequenced the genome of a female boxer using the whole-genome shotgun (WGS) approach[22,24] (see Methods and Supplementary Table S1). A total of 31.5 million sequence reads, providing ~7.5-fold sequence redundancy, were assembled with an improved version of the ARACHNE program[34], resulting in an initial assembly (CanFam1.0) used for much of the analysis below, and an updated assembly (CanFam2.0) containing minor improvements (Table 1 and Supplementary Table S2).

Genome assembly. The recent genome assembly spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly has extremely high continuity. The N50 contig size is 180 kb (that is, half of all bases reside in a contiguous sequence of 180 kb or more) and the N50 supercontig size is 45.0 Mb (Table 1). In particular, this means that most genes should contain no sequence gaps and that most canine chromosomes (mean size 61 Mb) have nearly all of their sequence ordered and oriented within one or two supercontigs (Supplementary Table S2). Notably, the sequence contigs are ~50-fold larger than the earlier survey sequence of the standard poodle[16].

The assembly was anchored to the canine chromosomes using data from both radiation hybrid and cytogenetic maps[11,13,14]. Roughly 97% of the assembled sequence was ordered and oriented on the chromosomes, showing an excellent agreement with the two maps. There were only three discrepancies, which were resolved by obtaining additional fluorescence *in situ* hybridization (FISH) data from the sequenced boxer. The 3% of the assembly that could not be anchored consists largely of highly repetitive sequence, including eight supercontigs of 0.5–1.0 Mb composed almost entirely of satellite sequence.

The nucleotide accuracy and genome coverage of the assembly is high (Supplementary Table S3). Of the bases in the assembly, 98% have quality scores exceeding 40, corresponding to an error rate of less than 10^{-4} and comparable to the standard for the finished human sequence[35]. When we directly compared the assembly to 760 kb of finished sequence (in regions where the boxer is

Table 1 | Assembly statistics for CanFam1.0 and 2.0

	CanFam1.0	CanFam2.0
N50 contig size	123 kb	180 kb
N50 supercontig size	41.2 Mb	45.0 Mb
Assembly size (total bases)	2.360 Gb	2.385 Gb
Number of anchored supercontigs	86	87
Percentage of genome in anchored supercontigs	96	97
Sequence in anchored bases	2.290 Gb	2.309 Gb
Percentage of assembly in gaps	0.9	0.8
Estimated genome size*	2.411 Gb	2.445 Gb
Percentage of assembly in 'certified regions', without assembly inconsistency	99.3	99.6

*Includes anchored bases, spanned gaps (21 Mb in CanFam1.0, 18 Mb in CanFam2.0) and centromeric sequence (3 Mb for each chromosome).

homozygous, to eliminate differences attributable to polymorphisms; see below), we found that the draft genome sequence covers 99.8% of the finished sequence and that bases with quality scores exceeding 40 have an empirical error rate of 2×10^{-5} (Supplementary Table S3).

Explaining the high sequence continuity. The dog genome assembly has superior sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago[24]. At least three factors contribute to the higher connectivity of the dog assembly (see Supplementary Information). First, we used a new version of ARACHNE with improved algorithms. Assembling the dog genome with the previous software version decreased N50 contig size from 180 kb to 61 kb, and assembling the mouse genome with the new version increased N50 contig size from 25 kb to 35 kb. Second, the amount of recently duplicated sequence is roughly twofold lower in dog than mouse (Supplementary Table S4); this improves contiguity because sequence gaps in both organisms tend to occur in recently duplicated sequence. Third, the dog sequence data has both higher redundancy (7.5-fold versus 6.5-fold) and higher quality (in terms of read length, pairing rate and tight distribution of insert sizes) compared with mouse. The contig size for the dog genome drops by about 32% when the data redundancy is decreased from 7.5-fold to 6.5-fold. A countervailing influence is that the dog genome contains polymorphism, whereas the laboratory mouse is completely inbred.

Assembly certification. Although 'quality scores' have been developed to indicate the nucleotide accuracy of a draft genome sequence[36], no analogous measures have been developed to reflect the long-range assembly accuracy. We therefore sought to develop such a measure on the basis of two types of internal inconsistencies (see Supplementary Information). The first is haplotype inconsistency, involving clear evidence of three or more distinct haplotypes within an assembled region from a single diploid individual. The second is linkage inconsistency, involving a cluster of reads for which the placement of the paired-end reads is illogical. This includes cases in which: (1) one end cannot be mapped to the region, (2) the linkage relationships are inconsistent with the sequence within contigs, or (3) distance constraints imply overlap between non-overlapping sequence contigs. The linkage inconsistency tests are most powerful when read pairs are derived from clone libraries with tight constraints on insert size. A region of assembly is defined as 'certified' if it is free of inconsistencies, and is otherwise 'questionable'.

Approximately 99.6% of the assembly resides in certified regions, with the N50 size of certified regions being ~12 Mb or about one-fifth of a chromosome. The remaining questionable regions are typically small (most are less than 40 kb), although there are a handful of regions of several hundred kilobases (Supplementary Fig. S1 and Supplementary Tables S5, S6). The questionable regions typically contain many inconsistencies, probably reflecting misassembly or overcollapse owing to segmental duplication. Chromosomes 2, 11 and 16 have 1.0–2.0% of their sequence in questionable regions. The certified and questionable regions are annotated in the public release of the dog genome assembly. With the concept of assembly certification, the scientific community can have appropriate levels of confidence in the draft genome sequence.

Genome landscape and evolution

Our understanding of the evolutionary processes that shape mammalian genomes has greatly benefited from the comparative analysis of sequenced primate[21,23] and rodent[24,25] genomes. However, the rodent genome is highly derived relative to that of the common ancestor of the eutherian mammals. As the first extensive sequence from an outgroup to the clade that includes primates and rodents, the dog genome offers a fresh perspective on mammalian genome evolution. Accordingly, we examined the rates and correlations of large-scale rearrangement, transposon insertion, deletion and nucleotide divergence across three major mammalian orders (primates, rodents and carnivores).

Conserved synteny and large-scale rearrangements. We created multi-species synteny maps from anchors of unique, unambiguously aligned sequences (see Supplementary Information), showing regions of conserved synteny among dog, human, mouse and rat genomes. Approximately 94% of the dog genome lies in regions of conserved synteny with the three other species (Supplementary Figs S2–S4 and Supplementary Table S7).

Given a pair of genomes, we refer to a 'syntenic segment' as a region that runs continuously without alterations of order and orientation, and a 'syntenic block' as a region that is contiguous in two genomes but may have undergone internal rearrangements. Syntenic breakpoints between blocks reflect primarily interchromosomal exchanges, and breakpoints between syntenic segments reflect intrachromosomal rearrangements. In the analysis below, we focus on syntenic segments of at least 500 kb.

We identified a total of 391 syntenic breakpoints across dog, human, mouse and rat genomes (Fig. 1 and Supplementary Figs S2, S5). With data for multiple species, it is possible to assign events to specific lineages (Fig. 1 and Supplementary Table S8). We counted the total number of breakpoints along the human, dog, mouse and rat lineages, with the values for each rodent lineage reflecting all breakpoints since the common ancestor with human (Fig. 1). The total number of breakpoints in the human lineage is substantially smaller than in the dog, mouse or rat lineages (83 versus 100, 161 or 176, respectively). However, there are more intrachromosomal breakpoints in the human lineage than in dog (52 versus 33).

Although the overall level of genomic rearrangement has been much higher in rodent than in human, comparison with dog shows that there are regions where the opposite is true. In particular, of the many intrachromosomal rearrangements previously observed between human chromosome 17 and the orthologous mouse

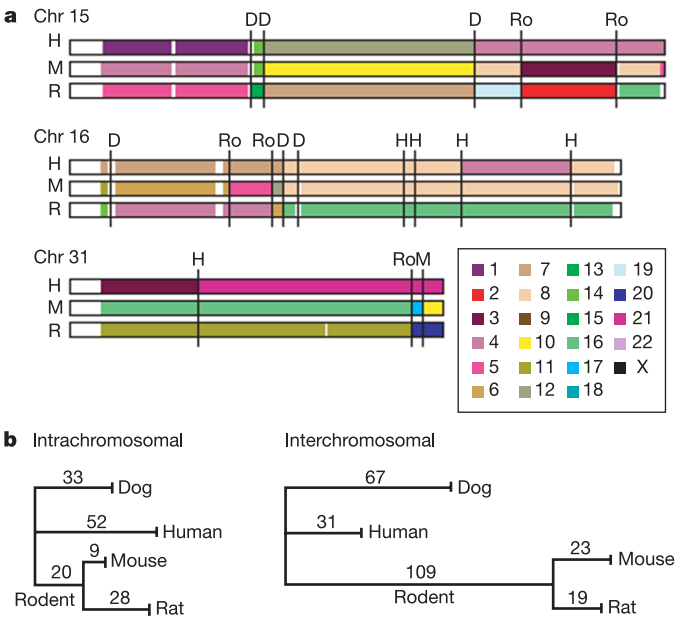

Figure 1 | Conserved synteny among the human, dog, mouse and rat genomes. a, Diagram of syntenic blocks (>500 kb) along dog chromosomes (Chr) 15, 16 and 31, with colours indicating the chromosome containing the syntenic region in other species. Synteny breakpoints were assigned to one of five lineages: dog (D), human (H), mouse (M), rat (R) or the common rodent ancestor (Ro). **b**, Lineage-specific intrachromosomal and interchromosomal breaks displayed on phylogenetic trees. Intrachromosomal breaks are seen more frequently in the human lineage than in mouse and rat, whereas interchromosomal breaks are somewhat more common in dog and considerably more common in rodents than in humans.

sequence[24], most have occurred in the human lineage (see Supplementary Information). Human chromosome 17 is rich in segmental duplications and gene families[21], which may contribute to its genomic fragility[37,38].

Genomic insertion and deletion. The euchromatic genome of the dog is ~150 Mb smaller than in mouse, and ~500 Mb smaller than in human. The smaller total size is reflected at the local level, with 100-kb blocks of conserved synteny in dog corresponding to regions for which the median size is ~3% larger in mouse and ~15% larger in human.

To understand the balance of forces that determine genome size, we studied the alignments of the human, mouse and dog genomes (Fig. 2). In particular, we identified the lineage-specific interspersed repeats within each genome, which consist of particular families of short interspersed elements (SINEs), long interspersed elements (LINEs) and other transposable elements that are readily recognized by sequence analysis (Supplementary Tables S9, S10). The remaining sequence was annotated as 'ancestral', consisting of both ancestral unique sequence and ancestral repeat sequence; these two categories were combined because the power to recognize ancient transposon-derived sequences degrades with repeat age, particularly in the rapidly diverging mouse lineage[24].

This comparative analysis indicates that different forces account for the smaller genome sizes in dog and mouse relative to human. The smaller size of the dog genome is primarily due to the presence of substantially less lineage-specific repeat sequence in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects a lower activity of endogenous retroviral and DNA transposons (~26,000 extant copies in dog versus ~183,000 in human), as well as the fact that the SINE element in dog is smaller than in human (although of similar length to that in mouse). As a consequence, the total proportion of repetitive elements (both lineage-specific and ancestral) recognizable in the genome is lower for dog (34%) than for mouse (40%) or human (46%). In contrast, the smaller size of the mouse genome is primarily due to a higher deletion rate. Specifically, the amount of extant 'ancestral sequence' is much lower in mouse (1,474 Mb) than in human (2,216 Mb) or dog (1,997 Mb). Assuming an ancestral genome size of 2.8 Gb (ref. 24) and also that deletions occur continuously, we suggest that the rate of genomic deletion in the rodent lineage has been approximately 2.5-fold higher than in the dog and human lineages (see Supplementary Information). As a consequence, the human genome shares ~650 Mb more ancestral sequence with dog than with mouse, despite our more recent common ancestor with the latter.

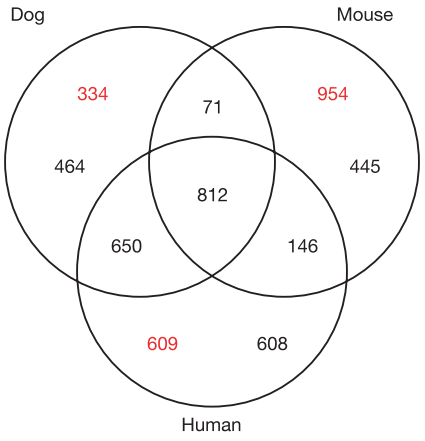

Figure 2 | Venn diagram showing the total lengths of aligned and unique sequences in the euchromatic portions of the dog, human and mouse genomes. Lengths shown in Mb, as inferred from genome-wide BLASTZ alignments (see Methods and Supplementary information). Overlapping partitions represent orthologous ancestral sequences. Each lineage-specific partition is further split into the total length of sequence classified as either lineage-specific interspersed repeats (red) or ancestral sequence (black). The latter is assumed to primarily represent ancestral sequences deleted in the two other species.

Active SINE family. Despite its relatively low proportion of transposable element-derived sequence, the dog genome contains a highly active carnivore-specific SINE family (defined as SINEC_Cf; RepBase release 7.11)[16]. The element is so active that many insertion sites are still segregating polymorphisms that have not yet reached fixation. Of ~87,000 young SINEC_Cf elements (defined by low divergence from the consensus sequence), nearly 8% are heterozygous within the draft genome sequence of the boxer. Moreover, comparison of the boxer and standard poodle genome sequences reveals more than 10,000 insertion sites that are bimorphic, with thousands more certain to be segregating in the dog population[16,39]. In contrast, the number of polymorphic SINE insertions in the human genome is estimated to be fewer than 1,000 (ref. 40).

The biological effect of these segregating SINE insertions is unknown. SINE insertions can be mutagenic through direct disruption of coding regions or through indirect effects on regulation and processing of messenger RNAs[39]. Such SINE insertions have already been shown to be responsible for two diseases in dog: narcolepsy and centronuclear myopathy[41,42]. It is conceivable that the genetic variation resulting from these segregating SINE elements has provided important raw material for the selective breeding programmes that have produced the wide phenotypic variations among modern dog breeds[16,43].

Sequence composition. The human and mouse genomes differ markedly in sequence composition, with the human genome having slightly lower average G+C content (41% versus 42% in mouse) but much greater variation across the genome. The dog genome closely resembles the human genome in its distribution of G+C content (Fig. 3a; Spearman's *rho* = 0.85 for dog–human and 0.76 for dog–mouse comparisons), even if we consider only nucleotides that can be aligned across all three species (Supplementary Fig. S6). The wider distribution of G+C content in human and dog is thus likely to reflect the boreoeutherian ancestor[44,45], with the more homogeneous composition in rodents having arisen primarily through lineage-specific changes in substitution patterns[46,47] rather than deletion of sequences with high G+C content.

Rate of nucleotide divergence. We estimated the mean nucleotide divergence rates in 1-Mb windows along the dog, human and mouse lineages on the basis of alignments of all ancestral repeats, using the consensus sequence for the repeats as a surrogate outgroup (Fig. 3b; see also Supplementary Information).

The dog lineage has diverged more rapidly than the human lineage (median relative divergence rate of 1.18, longer branch length in 95% of windows), but at only half the rate of the mouse lineage (median relative rate of 0.48, shorter branch length in 100% of windows). The absolute divergence rates are somewhat sensitive to the evolutionary model used and the filtering of alignment artefacts (data not shown), but the relative rates appear to be robust and are consistent with estimates from smaller sequence samples with multiple outgroups[28,48,49]. The lineage-specific divergence rates (human < dog < mouse) are probably explained by differences in metabolic rates[50,51] or generation times[52,53], but the relative contributions of these factors remain unclear[49].

Correlation in nucleotide divergence. As seen in other mammalian genomes[23–25], the average nucleotide divergence rate across 1-Mb windows varies significantly across the dog genome (coefficient of variation 0.11, compared with 0.024 expected under a uniform distribution). This regional variation shows significant correlation in orthologous windows across the dog, human and mouse genomes, but the strength of the correlation seems to decrease with total branch length (pair-wise correlation for orthologous 1-Mb windows: Spearman's *rho* = 0.49 for dog–human and 0.24 for dog–mouse comparisons). Lineage-specific variation in the regional divergence

rates may be coupled with changes in factors such as sequence composition or chromosomal position[23,54]. Consistent with this, the ratios of lineage-specific divergence rates in orthologous windows are positively correlated with the ratios of current G+C content in the same windows (Spearman's *rho* = 0.16 for dog–human, 0.24 for dog–mouse).

Male mutation bias. Comparison of autosomal and X chromosome substitution rates can be used to estimate the relative mutation rates in the male and female germ lines (α), because the X chromosome is present in females twice as often as in males. Using the lineage-specific rates from ancestral repeats, we estimate α as 4.8 for the lineage leading to human, and 2.8 for the lineages leading to both mouse and dog. These values fall between recent estimates from murids[24,25] and from hominids[23], and suggest that male mutation bias may have increased in the lineage leading to humans.

Mutational hotspots and chromosomal fission. Genome comparisons of human with both chicken[55] and chimpanzee[23] have previously revealed that sequences close to a telomere tend to have increased divergence rates and G + C content relative to interstitial sequences. It has been unclear whether these increases are inherent characteristics of the subtelomeric sequence itself or derived characteristics causally connected with its chromosomal position. We find a similar increase in both divergence (median increase 15%, $P < 10^{-5}$; Mann-Whitney *U*-test) and G+C content (median increase 9%, $P < 10^{-9}$) for subtelomeric regions along the dog lineage, with a sharp increase towards the telomeres (Supplementary Fig. S7).

This phenomenon is manifested at other synteny breaks, not only those at telomeres. We also observed a significant increase in divergence and G+C content in interstitial regions that are sites of syntenic breakpoints[54,56] (Supplementary Fig. S7). These properties therefore seem correlated with the susceptibility of regions to chromosomal breakage.

Proportion of genome under purifying selection

One of the striking discoveries to emerge from the comparison of the human and mouse genomes[21,24] was the inference that ~5.2% of the human genome shows greater-than-expected evolutionary conservation (compared with the background rate seen in ancestral repeat elements, which are presumed to be nonfunctional). This proportion greatly exceeds the 1–2% that can be explained by protein-coding regions alone. The extent and function of the large fraction of non-coding conserved sequence remain unclear[57], but this sequence is likely to include regulatory elements, structural elements and RNA genes.

Low turnover of conserved elements. We repeated the analysis of conserved elements using the human and dog genomes. Briefly, the analysis involves calculating a conservation score S_{HD}, normalized by the regional divergence rate, for every 50-bp window in the human genome that can be aligned to dog. The distribution of conservation scores for all genomic sequences is compared to the distribution in ancestral repeat sequences (which are presumed to diverge at the local neutral rate), showing a clear excess of sequences with high conservation scores. By subtracting a scaled neutral distribution from the total distribution, one can estimate the distribution of conservation scores for sequences under purifying selection. Moreover, for a given sequence with conservation score S_{HD}, one can also assign a probability $P_{selection}(S_{HD})$ that the sequence is under purifying selection (see ref. 24 and Supplementary Information).

The human–dog genome comparison indicates that ~5.3% of the human genome is under purifying selection (Fig. 4a), which is equivalent to the proportion estimated from human–rodent analysis. The obvious question is whether the bases conserved between human and dog coincide with the bases conserved between humans and rodents[25,58]. Because the conservation scores do not unambiguously assign sequences as either selected or neutral (but instead only assign probability scores for selection), we cannot directly compare the conserved bases. We therefore devised the following alternative approach.

We repeated the human–dog analysis, dividing the 1462 Mb of orthologous sequence between human and dog into those regions with (812 Mb) or without (650 Mb) orthologous sequence in mouse (Fig. 2). The first set shows a clear excess of conservation relative to background, corresponding to ~5.2% of the human genome (Fig. 4b). In contrast, the second set shows little or no excess conservation, corresponding to at most 0.1% of the human genome (Fig. 4c). This implies that hardly any of the functional elements conserved between human and dog have been deleted in the mouse lineage (see also Supplementary Information).

The results strongly suggest that there is a common set of functional elements across all three mammalian species, corresponding to ~5% of the human genome (~150 Mb). These functional elements reside largely within the 812 Mb of ancestral sequence common to human, mouse and dog. If we eliminate ancestral repeat elements within this shared sequence as largely non-functional, most functional elements can be localized to 634 Mb, and constitute approximately 24% of this sequence.

It should be noted that the estimate of ~5% pertains to conserved elements across distantly related mammals. It is possible that there are additional weakly constrained or recently evolved elements within narrow clades (for example, primates) that can only be detected by genomic sequencing of more closely related species[29].

Clustering of highly conserved non-coding elements. We next

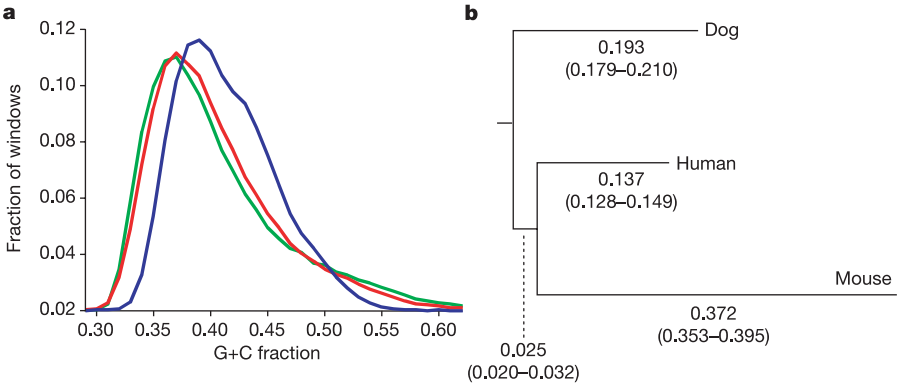

Figure 3 | Sequence composition and divergence rates. a, Distribution of G + C content in 10-kb windows across the genome in dog (green), human (red) and mouse (blue). **b**, Median lineage-specific substitution rates based on analysis of ancestral repeats aligning across all three genomes. Analysis was performed in non-overlapping 1-Mb windows across the dog genome that contained at least 2 kb of aligned ancestral repeat sequence (median 8.8 kb). The tree was rooted with the consensus sequences from the ancestral repeats. Numbers in parentheses give the 20–80th percentile range across the windows studied.

explored the distribution of conserved non-coding elements (CNEs) across mammalian genomes. For this purpose, we calculated a conservation score S_{HMD} based on simultaneous conservation across all three species (see Methods). We defined highly conserved non-coding elements (HCNEs) to be 50-bp windows that do not overlap coding regions and for which $P_{selection}(S_{HMD})$, the probability of being under purifying selection given the conservation score, is at least 95%. We identified ~140,000 such windows (6.5 Mb total sequence), comprising ~0.2% of the human genome and representing the most conserved ~5% of all mammalian CNEs.

The density of HCNEs shows striking peaks when plotted in 1-Mb windows across the genome (Fig. 4d and Supplementary Figs S8 and S9), with 50% lying in 204 regions that span less than 14% of the human genome (Supplementary Table S11). These regions are generally gene-poor, together containing only ~6% of all protein-coding sequence.

The genes contained within these gene-poor regions are of particular interest. At least 182 of the 204 regions contain genes with key roles in establishing or maintaining cellular 'state'. At least 156 of the regions contain one or, in a few cases, several transcription factors involved in differentiation and development[59]. Another 26 regions contain a gene important for neuronal specialization and growth, including several axon guidance receptors. The proportion of developmental regulators is far greater than expected by chance ($P < 10^{-31}$; see Supplementary Information).

We then tested whether the HCNEs within these regions tend to cluster around the genes encoding regulators of development. Analysis of the density of HCNEs in the intronic and intergenic sequences flanking every gene in the 204 regions revealed that the 197 genes encoding developmental regulators show an average of ~10-fold enrichment for HCNEs relative to the full set of 1,285 genes in the regions (Fig. 4e and Supplementary Fig. S10). The enrichment sometimes extends into the immediately flanking genes.

We note that the 204 regions include nearly all of the recently identified clusters of conserved elements between distantly related vertebrates such as chicken and pufferfish[55,59–62]. For example, they overlap 56 of the 57 large intervals containing conserved non-coding sequence identified between human and chicken[55]. The mammalian analysis, however, detects vastly more CNEs (>100-fold more sequence than with pufferfish[59] and 2–3-fold more than with chicken) and identifies many more clusters. The limited sensitivity of these more distant vertebrate comparisons may reflect the difficulty of aligning short orthologous elements across such large evolutionary distances or the emergence of mammal-specific regulatory elements. In any case, mammalian comparative analysis may be a more powerful tool for elucidating the regulatory controls across these important regions.

Although the function of conserved non-coding elements is unknown, on the basis of recent studies[59,63–66] it seems likely that many regulate gene expression. If so, the above results suggest that ~50% of all mammalian HCNEs may be devoted to regulating ~1% of all genes. In fact, the distribution may be even more skewed, as there are additional genomic regions with only slightly lower HCNE density than the 204 studied above (Supplementary Fig. S8). All of these regions clearly merit intensive investigation to assess indicators of regulatory function. We speculate that these regions may harbour characteristic chromatin structure and modifications that are potentially involved in the establishment or maintenance of cellular state.

Genes

Accurate identification of the protein-coding genes in mammalian genomes is essential for understanding the human genome, including its cellular components, regulatory controls and evolutionary

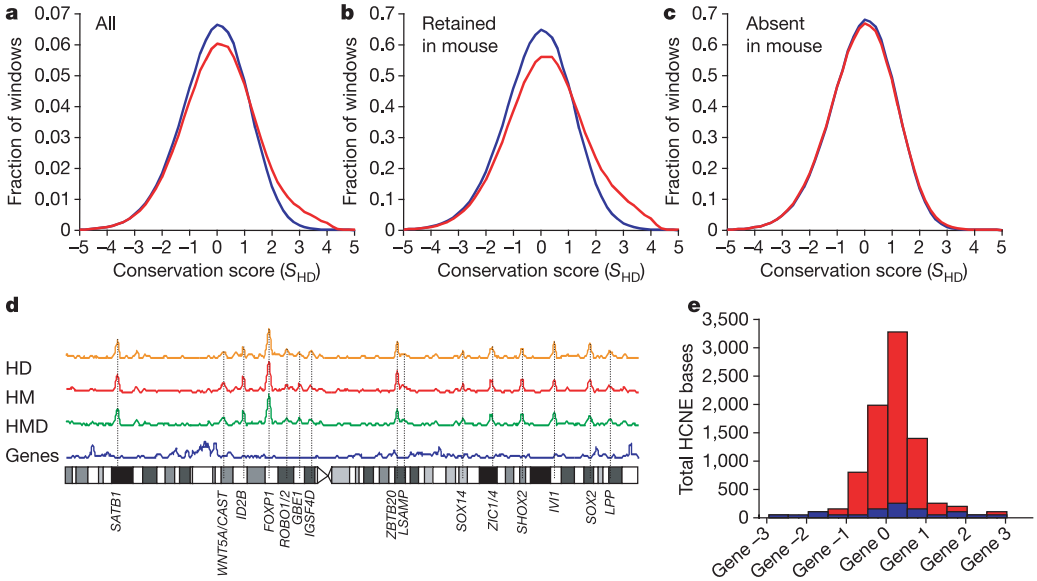

Figure 4 | Conservation of orthologous sequence between human and dog. **a**, Histogram of conservation scores, *S*, for all 50-bp windows across the human genome with at least 20 bases of orthologous sequence aligning to the dog genome, for all aligning sequences (red) and for ancestral repeat sequence only (blue). **b**, Conservation scores for the subset of windows that also have at least 20 bases of orthologous sequence aligning to the mouse genome. **c**, Conservation scores of the complementary subset of windows lacking such orthologous sequence in mouse. **d**, Density of 50-bp windows not overlapping known coding regions, for which $P_{selection}(S) > 95\%$, based on comparisons between human and dog (HD), human and mouse (HM), or between human, mouse and dog (HMD), and the density of known genes, all in 1-Mb sliding windows across human chromosome 3. **e**, Enrichment of HCNEs in the immediate neighbourhood of genes encoding developmental regulators in the 204 highly conserved regions. The histogram shows the median number of HCNE bases in the intronic and surrounding intergenic sequence, for the 197 known or putative development regulators (indicated by top of red bar) and for all of the 1,285 genes (blue bar). The histogram is centred at the 5′-end of the gene (marked 0) and each bin corresponds to half of the normalized distance to the flanking consecutive upstream genes (marked −1, −2 and −3) or consecutive downstream genes (1, 2 and 3) as indicated. The sequences surrounding the developmental genes are typically longer, have more HCNE sequence and have a higher density of HNCE sequence than other genes in the regions (see Supplementary Information).

constraints. The number of protein-coding genes in human has been a topic of considerable debate, with estimates steadily falling from ~100,000 to 20,000–25,000 over the past decade[21,22,67–70]. We analysed the dog genome in order to refine the human gene catalogue and to assess the evolutionary forces shaping mammals. (In the Genes section, 'gene' refers only to a protein-coding gene.)

Gene predictions in dog and human. We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains 19,300 dog gene predictions, with nearly all being clear homologues of known human genes.

The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (EnsEMBL build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable to spurious gene predictions in the human genome (M. Clamp, personal communication).

Gene duplications. Gene duplication is thought to contribute substantially to functional innovation[69,71]. We identified 216 gene duplications that are specific to the dog lineage and 574 that are specific to the human lineage, using the synonymous substitution rate K_S as a distance metric and taking care to discard likely pseudogenes. (The CanFam 2.0 assembly contains approximately 24 additional gene duplications, mostly olfactory receptors.) Human genes are thus 2.7-fold more likely to have undergone duplication than are dog genes over the same time period. This may reflect increased repeat-mediated segmental duplication in the human lineage[72].

Although gene duplication has been less frequent in dog than human, the affected gene classes are very similar. Prominent among the lineage-specific duplicated genes are genes that function in adaptive immunity, innate immunity, chemosensation and reproduction, as has been seen for other mammalian genomes[24,25,69,71]. Reproductive competition within the species and competition against parasites have thus been major driving forces in gene family expansion.

The two gene families with the largest numbers of dog-specific genes are the histone H2B family and the α-interferons, which cluster in monophyletic clades when compared to their human homologues. This is particularly notable for the α-interferons, for which the gene families within the six species (human, mouse, rat, dog, cat and horse) are apparently monophyletic. This may be due either to coincidental independent gene duplication in each of the six lineages or to ongoing gene conversion events that have homogenized ancestral gene duplicates[73].

Evolution of orthologous genes across three species. The dog genome sequence allows us for the first time to characterize the large-scale patterns of evolution in protein-coding genes across three major mammalian orders. We focused on a subset of 13,816 human, mouse and dog genes with 1:1:1 orthology. For each, we inferred the number of lineage-specific synonymous (K_S) and non-synonymous (K_A) substitutions along each lineage and calculated the K_A/K_S ratio (Table 2 and Supplementary Information), a traditional measure of the strength of selection (both purifying and directional) on proteins[74].

The median K_A/K_S ratio differs sharply across the three lineages ($P < 10^{-44}$, Mann-Whitney *U*-test), with the dog lineage falling between mouse and human. Population genetic theory predicts[75] that the strength of purifying selection should increase with effective population size (N_e). The observed relationship (mouse < dog < human) is thus consistent with the evolutionary prediction, given the expectation that smaller mammals tend to have larger effective population sizes[76].

We next searched for particular classes of genes showing deviations from the expected rate of evolution for a species. Such variation in rate (heterotachy) may point to lineage-specific positive selection or relaxation of evolutionary constraints[77]. We developed a statistical method similar to the recently described Gene Set Enrichment Analysis (GSEA)[78–80] to detect evidence of heterotachy for sets of functionally related genes (see Supplementary Information). Briefly, the approach involves ranking all genes by K_A/K_S ratio, testing whether the set is randomly distributed along the list and assessing the significance of the observed deviations by comparison with randomly permuted gene sets. In contrast to previous studies, which focused on small numbers of genes with prior hypotheses of selection, this approach detects signals of lineage-specific evolution in a relatively unbiased manner and can provide context to the results of more limited studies.

A total of 4,950 overlapping gene sets were studied, defined by such criteria as biological function, cellular location or co-expression (see Supplementary Information). Overall, the deviations between the three lineages are small, and median K_A/K_S ratios for particular gene sets are highly correlated for each pair of species (Supplementary Fig. S11). However, there is greater relative variation in human–mouse and dog–mouse comparisons than in human–dog comparisons (Supplementary Fig. S12).

This suggests that observed heterotachy between human and mouse must be interpreted with caution. For example, there is a great interest in the identification of genetic changes underlying the unique evolution of the human brain. A recent study[81] highlighted 24 genes involved in brain development and physiology that show signs of accelerated evolution in the lineage leading from ancestral primates to humans when compared to their rodent orthologues. We observe the same trend for the 18 human genes that overlap with the genes studied here, but find at least as many genes with higher relative acceleration in the dog lineage (see Supplementary Information). Heterotachy relative to mouse therefore does not appear to be a distinctive feature of the human lineage. It may reflect decelerated evolution in the rodent lineage, or possibly independent adaptive evolution in the human and dog lineages[82].

A small number of gene sets show evidence of significantly accelerated evolution in the human lineage, relative to both mouse and dog (32 sets at $z \geq 5.0$ versus zero sets expected by chance, $P < 10^{-4}$; Fig. 5a). These sets fall into two categories: genes expressed exclusively in testis, and (nuclear) genes encoding subunits of the mitochondrial electron transport chain (ETC) complexes. The former are believed to undergo rapid evolution as a consequence of sperm competition across a wide range of species[83–85], and lineage-specific acceleration suggests that sexual selection may have been a particularly strong force in primate evolution. The selective forces acting on the latter category are less obvious. Because of the importance of mitochondrial ATP generation for sperm motility[86], and the potentially antagonistic co-evolution of these genes with maternally inherited mitochondrial DNA-encoded subunits[87], we

Table 2 | Evolutionary rates for 1:1:1 orthologues among dog, mouse and human

	Median (20–80th percentile range)			Spearman's *rho*		
	Dog*	Mouse	Human	Dog–human	Dog–mouse	Human–mouse
K_S	0.210 (0.138–0.322)	0.416 (0.310–0.558)	0.139 (0.0928–0.214)	0.47	0.50	0.52
K_A	0.021 (0.006–0.051)	0.038 (0.013–0.087)	0.017 (0.005–0.040)	0.87	0.87	0.86
K_A/K_S	0.095 (0.030–0.221)	0.088 (0.031–0.197)	0.112 (0.034–0.272)	0.80	0.85	0.82

*Estimates are based on unrooted tree. The dog branch thus includes the branch from the boreoeutherian ancestor to the primate–rodent split.

propose that sexual selection may also be the primary force behind the rapid evolution of the primate ETC genes. Given the ubiquitous role of mitochondrial function, however, such sexual selection may have led to profound secondary effects on physiology[88].

We found no gene sets with comparably strong evidence for dog-specific accelerated evolution. There is, however, a small excess of sets with moderately high acceleration scores (19 sets at $z \geq 3.0$ versus 5 sets expected by chance, $P < 0.02$; Fig. 5b). These sets, which are primarily related to metabolism, may contain promising candidates for follow-up studies of molecular adaptation in carnivores.

Polymorphism and haplotype structure in the domestic dog

The modern dog has a distinct population structure with hundreds of genetically isolated breeds, widely varying disease incidence and distinctive morphological and behavioural traits[89,90]. Unlocking the full potential of the dog genome for genetic analysis requires a dense SNP map and an understanding of the structure of genetic variation both within and among breeds.

Generating a SNP map. We generated a SNP map of the dog genome containing >2.5 million distinct SNPs mapped to the draft genome sequence, corresponding to an average density of approximately one SNP per kb (Table 3). The SNPs were discovered in three complementary ways (see Supplementary Information). (1) We identified SNPs within the sequenced boxer genome (set 1; ~770,000 SNPs) by searching for sites at which alternative alleles are supported by at least two independent reads each. We tested a subset ($n = 40$ SNPs) by genotyping and confirmed all as heterozygous sites. (2) We compared the 1.5× sequence from the standard poodle[16] with the draft genome sequence from the boxer (set 2; ~1,460,000 SNPs). (3) We generated shotgun sequence data from nine diverse dog breeds (~100,000 reads each, 0.02× coverage), four grey wolves and one coyote (~22,000 reads each, 0.004× coverage) and compared it to the boxer (set 3; ~440,000 SNPs). We tested a subset ($n = 1,283$ SNPs) by genotyping and confirmed 96% as true polymorphisms.

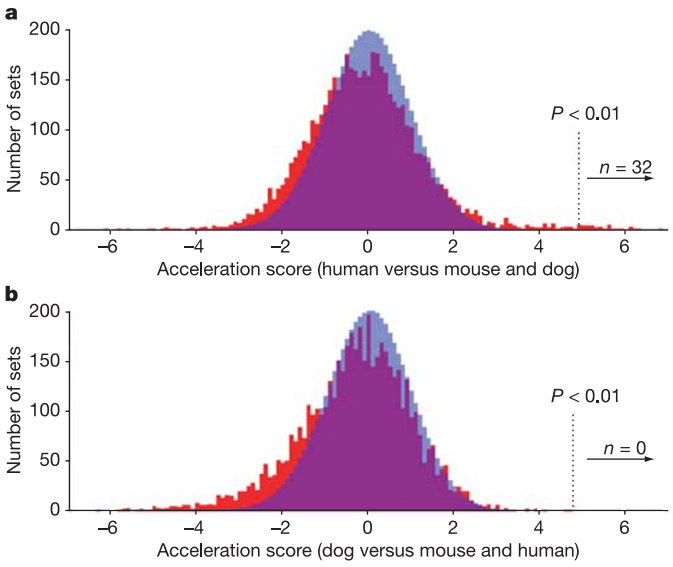

Figure 5 | Gene sets showing accelerated evolution along the human and dog lineages. a, Distribution of acceleration scores along the human lineage relative to both mouse and dog, observed for 4,950 gene sets (red). The expected distribution based on 10,000 randomized trials is shown in blue. The dotted line shows the acceleration score for which the probability of observing even a single set by random chance (out of the 4,950 sets tested) is less than 1%. In fact, 32 sets show acceleration scores on the human lineage exceeding this threshold. **b**, The observed (red) and expected (blue) distribution of acceleration scores for the dog lineage, relative to both human and mouse.

Table 3 | SNPs discovered in dogs, wolves and coyotes compared to the boxer assembly

Set number	Breed or species	Number of SNPs	SNP rate (one per *x* bases)
1	Boxer versus boxer	**768,948**	3,004 (observed) **1,637** (corrected)
2	Boxer versus poodle	**1,455,007**	**894**
3a	Boxer versus breeds*		
	German shepherd	45,271	900
	Rottweiler	44,097	917
	Bedlington terrier	44,168	913
	Beagle	42,572	903
	Labrador retriever	40,730	926
	English shepherd	40,935	907
	Italian greyhound	39,390	954
	Alaskan malamute	45,103	787
	Portuguese water dog	45,457	896
	Total distinct SNPs	**373,382**	**900**
3b	Boxer versus Canids†		
	China grey wolf	12,182	580
	Alaska grey wolf	13,888	572
	India grey wolf	14,510	573
	Spanish grey wolf	10,349	587
	California coyote	20,270	417
	Total distinct SNPs	**71,381**	
3	Set 3 total distinct SNPs	**441,441**	
Total	Total distinct SNPs	**2,559,519**	

*Based on ~100,000 sequence reads per breed.
†Based on ~20,000 sequence reads per wolf.

The SNP rate between the boxer and any of the different breeds is one SNP per ~900 bp, with little variation among breeds (Table 3). The only outlier (~1/790 bp) is the Alaskan malamute, which is the only breed studied that belongs to the Asian breed cluster[91]. The grey wolf (~1/580 bp) and coyote (~1/420 bp) show greater variation when compared with the boxer, supporting previous evidence of a bottleneck during dog domestication, whereas that the SNP rate is lower in the grey wolf than in the coyote reflects the closer relationship of the grey wolf to the domestic dog[1–3,92] (see section 'Resolving canid phylogeny').

The observed SNP rate within the sequenced boxer assembly is ~1/3,000 bp. This underestimates the true heterozygosity owing to the conservative criterion used for identifying SNPs within the boxer assembly (requiring two reads containing each allele); correcting for this leads to an estimate of ~1/1,600 bp (see Supplementary Information). This low rate reflects reduced polymorphism within a breed, compared with the greater variation of ~1/900 bp between breeds.

To assess the utility of the SNPs for dog genetics, we genotyped a subset from set 3a ($n = 1,283$) in 20 dogs from each of ten breeds (Supplementary Table S16). Within a typical breed, ~73% of the SNPs were polymorphic. The polymorphic SNPs have minor allele frequencies that are approximately evenly distributed between 5% and 50% (allele frequencies less than 5% are not reliable with only 40 chromosomes sampled). In addition, the SNPs from sets 2 and 3 have a roughly uniform distribution across the genome (Fig. 6a, see below concerning set 1). The SNP map thus has high density, even distribution and high cross-breed polymorphism, indicating that it should be valuable for genetic studies.

Expectations for linkage disequilibrium and haplotype structure. Modern dog breeds are the product of at least two population bottlenecks, the first associated with domestication from wolves (~7,000–50,000 generations ago) and the second resulting from intensive selection to create the breed (~50–100 generations ago). This population history should leave distinctive signatures on the patterns of genetic variation both within and across breeds. We might expect aspects of both the long-range LD seen in inbred mouse strains, with strain-specific haplotypes extending over multiple megabases, and the short-range LD seen in humans, with ancestral haplotype blocks typically extending over tens of kilobases. Specifically,

long-range LD would be expected within dog breeds and short-range LD across breeds.

Preliminary evidence of long-range LD within breeds has been reported[90]. Five genome regions were examined (~1% of the genome) in five breeds using ~200 SNPs with high minor allele frequency. LD seemed to extend 10–100-fold further in dog than in human, with relatively few haplotypes per breed.

With the availability of a genome sequence and a SNP map, we sought to undertake a systematic analysis of LD and haplotype structure in the dog genome.

Haplotype structure within the boxer assembly. We first analysed the structure of genetic variation within the sequenced boxer genome by examining the distribution of the ~770,000 SNPs detected between homologous chromosomes. Strikingly, the genome is a mosaic of long, alternating regions of near-total homozygosity and high heterozygosity (Fig. 6b, c), with observed SNP rates of ~14 per Mb and ~850 per Mb, respectively. (The latter is close to that seen within breeds and is indistinguishable when one corrects for the conservative criterion used to identify SNPs within the boxer assembly; see Supplementary Information.) The homozygous regions have an N50 size of 6.9 Mb and cover 62% of the genome, and the heterozygous regions have an N50 size of 1.1 Mb and cover 38% of the genome. The results imply that the boxer genome is largely comprised of vast haplotype blocks. The long stretches of homozygosity indicate regions in which the sequenced boxer genome carries the same haplotype on both chromosomes. The proportion of homozygosity (~62%) reflects the limited haplotype diversity within breeds.

Long-range haplotypes in different breeds. We sought to determine whether the striking haplotype structure seen in the boxer genome is representative of most dog breeds. To this end, we randomly selected ten regions of 15 Mb each (~6% of the genome) and examined linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of ten breeds and one dog from each of 24 additional breeds (see Supplementary Tables S17–S19).

The ten breeds were chosen to represent all four clusters described in ref. 91. The selected breeds have diverse histories, with varying population size and bottleneck severity. For example, the Basenji is an ancient breed from Africa that has a small breeding population in the United States descending from dogs imported in the 1930s–1940s (refs 93, 94). The Irish wolfhound suffered a severe bottleneck two centuries ago, with most dogs today being descendents of a single dog in the early 1800s (refs 5, 94). In contrast, the Labrador retriever and golden retriever have long been, and remain, extremely popular dogs

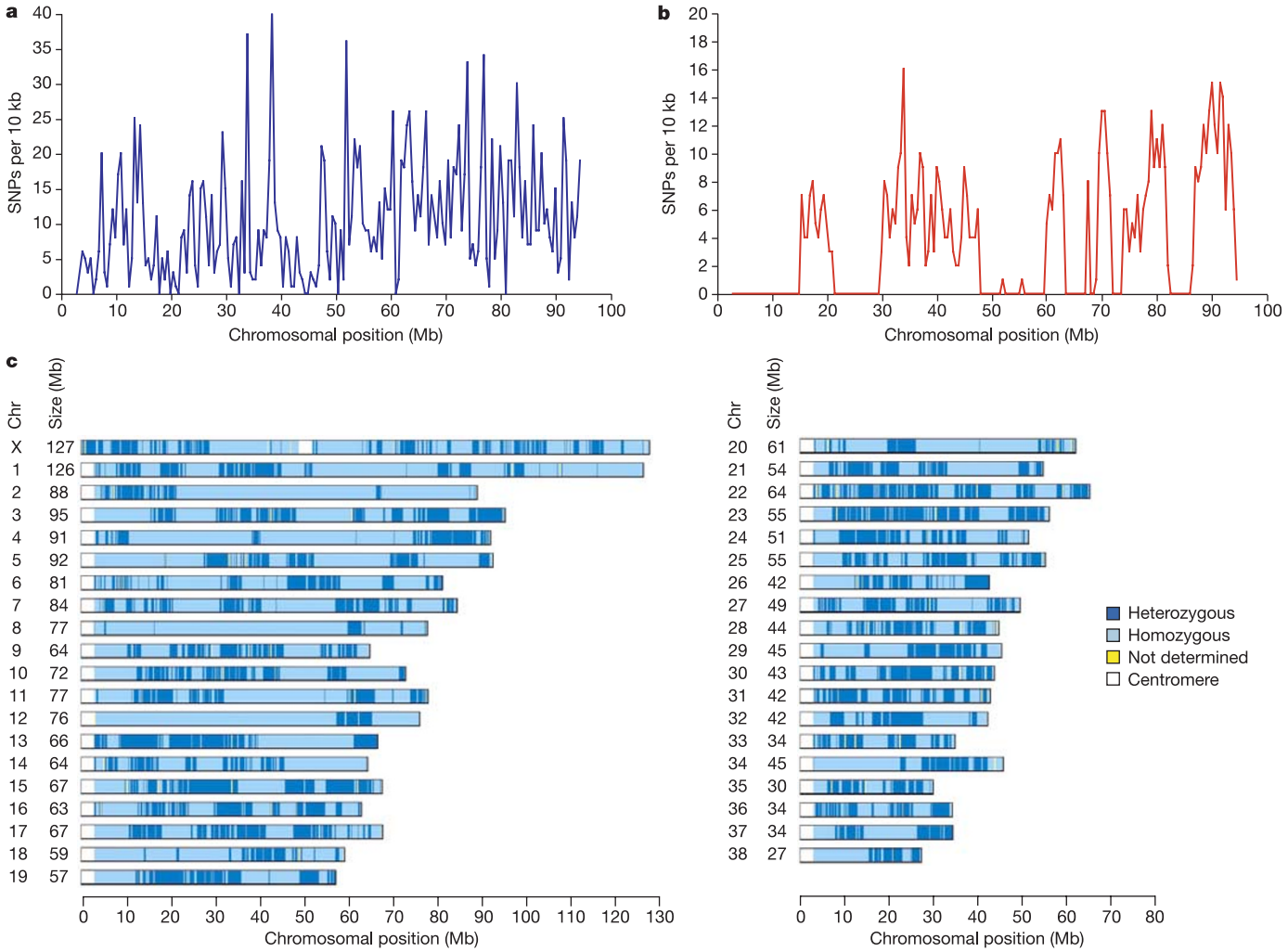

Figure 6 | The distribution of SNPs is fairly uniform across breeds, but non-uniform within the sequenced boxer assembly. a, SNPs across chromosome 3, generated by comparing the boxer assembly with WGS reads from nine breeds. **b**, The SNPs on chromosome 3 of the boxer assembly show an uneven distribution (plotted in 500-kb windows). Note that boxer SNPs were identified using a more conservative method, lowering the observed SNP rate by roughly twofold. **c**, An alternating pattern of large homozygous (light blue, ~62% of genome; N50 size 6.9 Mb) and large heterozygous (dark blue ~38% of genome; N50 size 1.1 Mb) blocks indicates large identical or divergent haplotypes across the boxer genome. White indicates centromeric sequence.

(with ~150,000 and ~50,000 new puppies registered annually, respectively). They have not undergone such recent severe bottlenecks, but some lines have lost diversity because of the repeated use of popular sires[89]. The Glen of Imaal terrier represents the opposite end of the popularity spectrum, with fewer than 100 new puppies registered with the American kennel Club each year.

The 224 dogs were genotyped for SNPs across each of the ten regions, providing 2,240 cases in which to assess long-range LD. The SNPs ($n = 1{,}219$; Supplementary Table S19) were distributed along the regions to measure the fall-off of genetic correlation, with higher density at the start of the region and lower densities at further distances (Fig. 7a). In 645 cases, we also examined the first 10 kb in

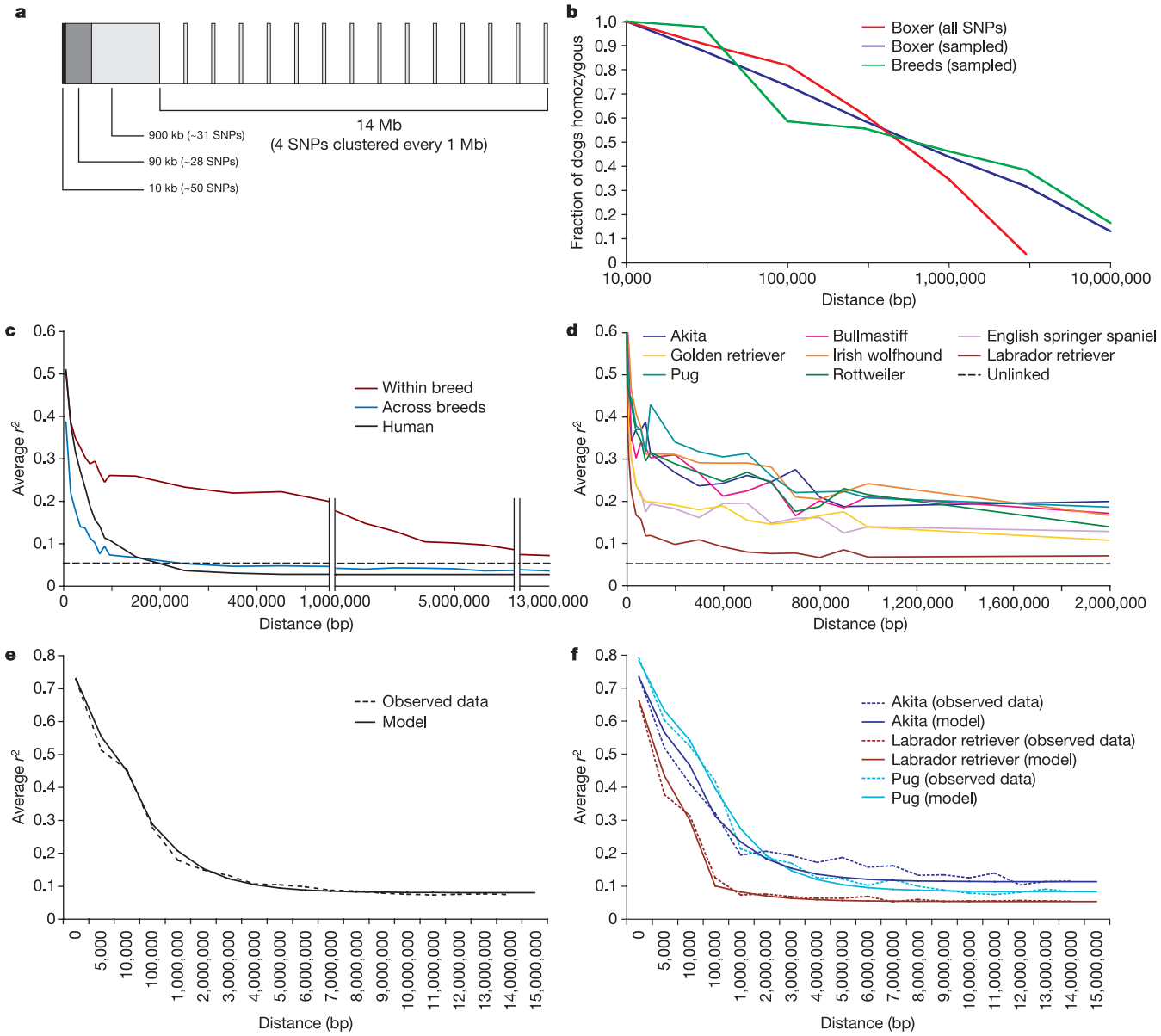

Figure 7 | Homozygous regions and linkage disequilibrium are nearly 100-fold longer within dog breeds than across the dog or human populations. **a**, Sampling design for ten random regions of 15 Mb each, used to assess the haplotype structure of ~6% of the genome (see Supplementary Information). For each region, we examined the first 10 kb through resequencing and dense genotyping. To detect long haplotypes, we genotyped SNPs distributed throughout the next 1 Mb and sampled SNPs at intervals of 1 Mb for the next 14 Mb. In total we genotyped 1,219 SNPs across the ten regions in a collection of 224 dogs (20 dogs from each of 10 breeds and one dog from each of 24 breeds). **b**, Conditional on a dog being homozygous for the initial 10-kb region ($n = 245$), we assessed the probability that the dog was homozygous for all SNPs within a given distance. The average proportion remaining homozygous is compared for the various breeds (green), for the boxer when sampled in the same ways as the breeds (blue) and for the boxer using all SNPs found in the genome sequence (red). About 50% of the individuals seem to be homozygous throughout 1 Mb both in the boxer and other breeds, indicating that other breeds have comparable long-range homozygosity. **c**, Linkage disequilibrium (LD) as a function of distance is shown as the r^2 statistic within individual breeds (red), across various breeds (blue) and a human population (black) taken from the CEPH collection genotyped as part of the ENCODE component of the International HapMap Project[118]. For the overall dog and human populations, LD falls rapidly, reaching the baseline level seen for unlinked loci by ~200 kb. In contrast, LD for individual breeds falls initially but then stays at a moderately high level across several megabases. **d**, The LD curves are broadly similar for most breeds, but the proportion of long-range LD is correlated with known breed history. **e**, The observed within-breed LD curve (averaged across breeds) is well fitted by a simple model with a domestication bottleneck 10,500 generations ago and a breed-creation bottleneck occurring 50 generations ago (see Supplementary Information). **f**, LD curves for individual dog breeds can be fitted by models with different breed-creation bottlenecks. The poorest fit is obtained for the akita, the breeding history of which is known to involve two separate breed-creation bottlenecks.

greater detail by denser genotyping (with ~2 SNPs per kb) in 405 cases and complete resequencing in 240 cases. The resequencing data yielded a heterozygosity rate of ~1 SNP per 1,500 bp, essentially equivalent to the rate seen in the sequenced boxer genome.

On the basis of examining the first 10 kb, we found that ~38% of instances seem to be completely homozygous and that all dogs seem to be homozygous for at least one of the ten regions. We then measured the distance over which homozygosity persisted. Of instances homozygous in the initial 10-kb segment, 46% were homozygous across 1 Mb and 17% were still homozygous across 10 Mb (Fig. 7b). The fall-off in homozygosity is essentially identical to that seen in the boxer genome, provided that the boxer data are sampled in an equivalent manner (see Supplementary Information). This indicates that the long-range haplotype structure seen in the boxer is typical of most dog breeds, although the precise haplotypes vary with breed and the locations of homozygous regions vary between individuals.

We also assessed long-range correlations by calculating r^2, a traditional measure of LD, across the 15-Mb regions. The r^2 curve representing the overall dog population (one dog from each of 24 breeds) drops rapidly to background levels. This is in sharp contrast to the r^2 curves within each breed. Within breeds, LD is biphasic, showing a sharp initial drop within ~90 kb followed by an extended shoulder that gradually declines to the background (unlinked) level by 5–15 Mb in most breeds (Fig. 7c). The basic pattern is similar in all ten regions (Supplementary Fig. S13) and in all breeds (Fig. 7d). (Labrador retrievers show the shortest LD, probably due to their mixed aetiology and large population size.)

The biphasic r^2 curves within each breed thus consist of two components (Fig. 7e), at scales differing by ~100-fold. The first component matches the fall-off in the general dog population and is likely to represent the short-range de-correlation of local haplotype blocks in the ancestral dog population. The second component represents long-range breed-specific haplotypes (Fig. 8a). Notably, the first component falls off nearly twice as quickly as the LD in the human population (~200 kb), and the second component falls off slightly slower than seen in laboratory mouse strains[95].

Modelling the effects of population history. We tested this interpretation by performing mathematical simulations on a dog population that underwent an ancient bottleneck and recent breed-creation bottlenecks, using the coalescent approach[96] (see Supplementary Information). Our experimental results were well fitted by models assuming an ancient bottleneck (effective domesticated population size 13,000, inbreeding coefficient $F = 0.12$) occurring ~9,000 generations ago (corresponding to ~27,000 years) and subsequent breed-creation bottlenecks of varying intensities occurring 30–90 generations ago[97] (Supplementary Fig. S14). The model closely reproduces the observed r^2 curves and the observed polymorphism rates within breeds, among breeds and between dog and grey wolf. The model also yields estimates of breed-specific bottlenecks that are broadly consistent with known breed histories. For example, Labrador retrievers, and to a lesser extent golden retrievers and English springer spaniels, show less severe bottlenecks.

Deterministically modelled results (Fig. 7e, f) indicate that a simple, two-bottleneck model provides a close fit to the data for the breeds. They do not rule out a more complex population history, such as multiple domestication events, low levels of continuing gene flow between domestic dog and grey wolf[97,98] or multiple bottlenecks within breeds. Notably, the akita yields the poorest fit to the model, with an r^2 curve that appears to be triphasic. This may reflect the initial creation of the breed as a hunting dog in Japan ~450 generations ago, and a consecutive bottleneck associated with its introduction into the United States during the 1940s (ref. 99).

Haplotype diversity. We next studied haplotype diversity within and among breeds, using the dense genotypes from the 10-kb regions. Across the 645 cases examined, there is an average of ~10 distinct haplotypes per region. Within a breed, we typically see four of these haplotypes, with the average frequency of the most common haplotype being 55% and the average frequency of the two most common being 80% (Fig. 8c and Supplementary Fig. S18). The haplotypes and their frequencies differ sharply across breeds. Nonetheless, 80% of the haplotypes seen with a frequency of at least 5% in one breed are found in other breeds as well (Supplementary Table S26). This extends previous observations of haplotype sharing across breeds[90]. In particular, the inclusion of all SNPs with a minor allele frequency ≥5% across all breeds provides a more accurate picture of haplotype sharing, because the analysis includes haplotypes that are rare within a single breed but more common across the population.

We then inferred the ancestral haplotype block structure in the ancestral dog population (before the creation of modern breeds) by combining the data across breeds and applying methods similar to those used for haplotype analysis in the human genome[100] (see Supplementary Information). In the 10-kb regions studied, one or two haplotype blocks were typically observed. Additional data across 100-kb regions suggest that the ancestral blocks have an average size of ~10 kb. The blocks typically have ~4–5 distinct haplotypes across the entire dog population (Fig. 8b). The overall situation closely resembles the structure for the human genome, although with slightly smaller block size (Supplementary Figs S15–S19 and Supplementary Table S24–26).

Ancestral and breed-specific haplotypes. A clear picture of the population genetic history of dogs emerges from the results detailed above:

- The ancestral dog population had short-range LD. The haplotype blocks were somewhat shorter than in modern humans (~10 kb versus ~20 kb in human), consistent with the dog population being somewhat older than the human population (~9,000 generations versus ~4,000 generations). Haplotype blocks at large distances were essentially uncorrelated (Fig. 8a).
- Breed creation introduced tight breed-specific bottlenecks, at least for the breeds examined. From the great diversity of long-range haplotype combinations carried in the ancestral population, the founding chromosomes emerging from the bottleneck represented only a small subset. These became long-range breed-specific haplotypes (Fig. 8a).
- Although the breed-specific bottlenecks were tight, they did not cause massive random fixation of individual haplotypes. Only 13% of the small ancestral haplotypes are monomorphic within a typical breed, consistent with the estimated inbreeding coefficient of ~12%. Across larger regions (≥100 kb), we observed no cases of complete fixation within a breed (Supplementary Fig. S20).
- There is notable sharing of 100-kb haplotypes across breeds, with ~60% seen in multiple breeds although with different frequencies. On average, the probability of sampling the same haplotype on two chromosomes chosen from different breeds is roughly twofold lower than for chromosomes chosen within a single breed (Supplementary Fig. S21).

Implications for genetic mapping. These results have important implications for the design of dog genetic studies. Although early efforts focused on cross-breeding of dogs for linkage analysis[101–103], it is now clear that within-breed association studies offer specific advantages in the study of both monogenic and polygenic diseases. First, they use existing dogs coming to medical attention and do not require the sampling of families with large numbers of affected individuals. Such studies should be highly informative, because dog breeds have retained substantial genetic diversity. Moreover, they will require a much lower density of SNPs than comparable human association studies, because the long-range LD within breeds extends ~50-fold further than in humans[90,104,105].

Whereas human association studies require >300,000 evenly spaced SNPs[100,106,107], the fact that LD extends over at least 50-fold greater distances in dog suggests that dog association studies would require perhaps ~10,000 evenly spaced SNPs. To estimate the

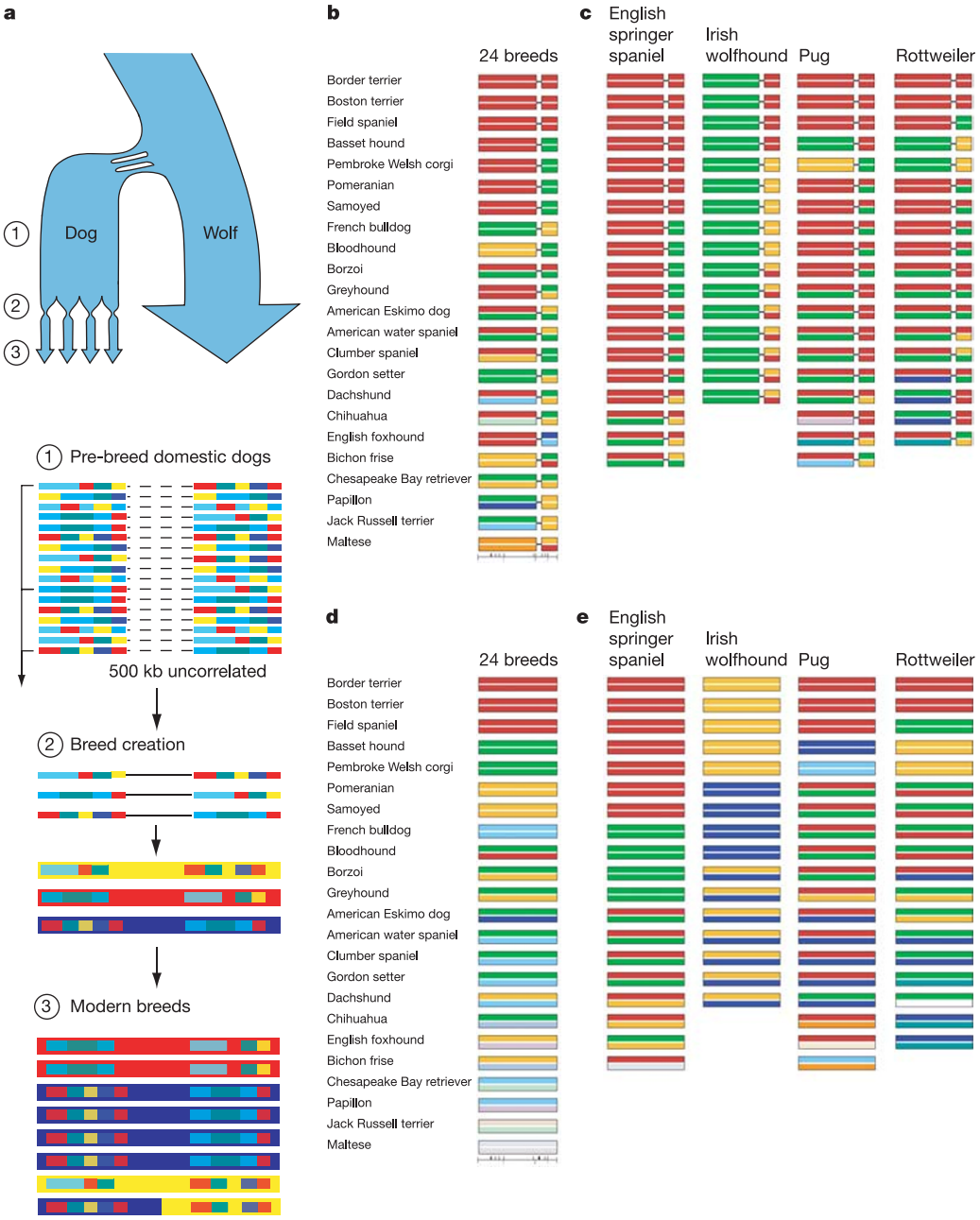

Figure 8 | Two bottlenecks, one old and one recent, have shaped the haplotype structure and linkage disequilibrium of canine breeds. **a**, Modern haplotype structure arose from key events in dog breeding history. The domestic dog diverged from wolves 15,000–100,000 years ago[97,119], probably through multiple domestication events[98]. Recent dog breeds have been created within the past few hundred years. Both bottlenecks have influenced the haplotype pattern and LD of current breeds. (1) Before the creation of modern breeds, the dog population had the short-range LD expected on the basis of its large size and time since the domestication bottleneck. (2) In the creation of modern breeds, a small subset of chromosomes was selected from the pool of domestic dogs. The long-range patterns that happened to be carried on these chromosomes became common within the breed, thereby creating long-range LD. (3) In the short time since breed creation, these long-range patterns have not yet been substantially broken down by recombination. Long-range haplotypes, however, still retain the underlying short-range ancestral haplotype blocks from the domestic dog population, and these are revealed when one examines chromosomes across many breeds. **b**, **c**, Distribution of ancestral haplotype blocks in a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**b**) and within four breeds (**c**). Ancestral haplotype blocks are 5–15 kb in size (which is shorter than the ~25-kb blocks seen in humans) and are shared across breeds. Typical blocks show a spectrum of ~5 haplotypes, with one common major haplotype. Blocks were defined using the modified four-gamete rule (see Supplementary Information) and each haplotype (minor allele frequency (maf) > 3%) within a block was given a unique colour. **d**, **e**, Distribution of breed-derived haplotypes across a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**d**) and within four breeds (**e**). Each colour denotes a distinct haplotype (maf > 3%) across 11 SNPs in the 10-kb window for each of the analysed dogs. Pairs of haplotypes have an average of 3.7 differences. Most haplotypes can be definitively identified on the basis of homozygosity within individual dogs. Grey denotes haplotypes that cannot be unambiguously phased owing to rare alleles or missing data. Within each of the four breeds shown, there are 2–5 haplotypes, with one or two major haplotypes accounting for the majority of the chromosomes. Across the 24 breeds, there are a total of seven haplotypes. All but three are seen in multiple breeds, although at varying frequencies.

number of SNPs required, we generated SNP sets from ten 1-Mb regions by coalescent simulations using the bottleneck parameters that generate SNP rates and LD curves equivalent to the actual data (Supplementary Fig. S14 and Supplementary Table S20). We then selected individual SNPs as 'disease alleles' and tested our ability to map them by association analysis with various marker densities (Fig. 9a).

For disease alleles causing a simple mendelian dominant trait with high penetrance and no phenocopies, there is overwhelming power to map the locus (Fig. 9a). Using ~15,000 evenly spaced SNPs and a log likelihood odds ratio (LOD score) score threshold of 5, the probability of detecting the locus is over 99% given a collection of 100 affected and 100 unaffected dogs. (The LOD score threshold corresponds to a false positive rate of 3% loci per genome.)

For a multigenic trait, the power to detect disease alleles depends on several factors, including the relative risk conferred by the allele, the allele frequency and the interaction with other alleles. We investigated a simple model of an allele that increases risk by a multiplicative factor (λ) of 2 or 5 (see Supplementary Information). Using the above SNP density and LOD score threshold, the power to detect a locus with a sample of 100 affected and 100 unaffected dogs is 97% for $\lambda = 5$ and 50% for $\lambda = 2$ (Fig. 9b, c). Although initial mapping will be best done by association within breeds, subsequent fine-structure mapping to pinpoint the disease gene will probably benefit from cross-breed comparison. Given the genetic relationships across breeds described above, it is likely that the same risk allele will be carried in multiple breeds. By comparing risk-associated haplotypes in multiple breeds, it should be possible to substantially narrow the region containing the gene.

Resolving canid phylogeny

The dog family, Canidae, contains 34 closely related species that diverged within the last ~10 million years[1]. Resolving the evolutionary relationships of such closely related taxa has been difficult because a great quantity of genomic sequence is typically required to yield enough informative nucleotide sites for the unambiguous reconstruction of phylogenetic trees. We sought to streamline the process of evolutionary reconstruction by exploiting our knowledge of the dog genome to select genomic regions that would maximize the amount of phylogenetic signal per sequenced base. Specifically, we sought regions of rapidly evolving, unique sequence.

We first compared the coding regions of 13,816 dog genes with human–dog–mouse 1:1:1 orthologues to find those with high neutral evolutionary divergence (comparing K_S and K_A/K_S). We selected 12 exons (8,080 bp) for sequencing, based on the criteria that their sequences (1) are consistent with the known phylogeny of human, dog, mouse and rat, (2) have a high percentage of bases (≥15%) that are informative for phylogenetic reconstruction in the human, dog, mouse and rat phylogenies, and (3) could be successfully amplified in all canids. The chosen exons contain 3.3-fold more substitutions than random exonic sequence. Using our SNP database, we also evaluated introns to identify those with high variation between dog and coyote. We selected four introns (3,029 bp) that contained ~5-fold more SNPs than the background frequency. We sequenced these exons and introns (11,109 bp) in 30 out of 34 living wild canids, and we combined the data with additional sequences (3,839 bp) from recent studies[3,92].

The resulting evolutionary tree has a high degree of statistical support (Fig. 10), and uniquely resolves the topology of the dog's closest relatives. Grey wolf and dog are most closely related (0.04% and 0.21% sequence divergence in nuclear exon and intron sequences, respectively), followed by a close affiliation with coyote, golden jackal and Ethiopian wolf, three species that can hybridize with dogs in the wild (Fig. 10). Closest to this group are the dhole and African wild dog, two species with a uniquely structured meat-slicing tooth, suggesting that this adaptation was later lost. The molecular tree supports an African origin for the wolf-like canids, as the two African jackals are the most basal members of this clade. The two other large groupings of canids are (1) the South American canids, which are clearly rooted by the two most morphologically divergent canids, the maned wolf and bush dog; and (2) the red fox-like canids, which are rooted by the fennec fox and Blanford's fox, but now also include the raccoon dog and bat-eared fox with higher support. Together, these three clades contain 93% of all living canids. The grey fox lineage seems to be the most primitive and suggests a North American origin of the living canids about 10 million years ago[1].

These results demonstrate the close kinship of canids. Their limited sequence divergence suggests that many molecular tools developed for the dog (for example, expression microarrays) will be useful for exploring adaptation and evolutionary divergence in other canids as well.

Conclusions

Genome comparison is a powerful tool for discovery. It can reveal unknown—and even unsuspected—biological functions, by sifting the records of evolutionary experiments that have occurred over 100 years or over 100 million years. The dog genome sequence illustrates the range of information that can be gleaned from such studies.

Mammalian genome analysis is helping to develop a global picture of gene regulation in the human genome. Initial comparison with rodents revealed that ~5% of the human genome is under purifying selection, and that the majority of this sequence is not protein-

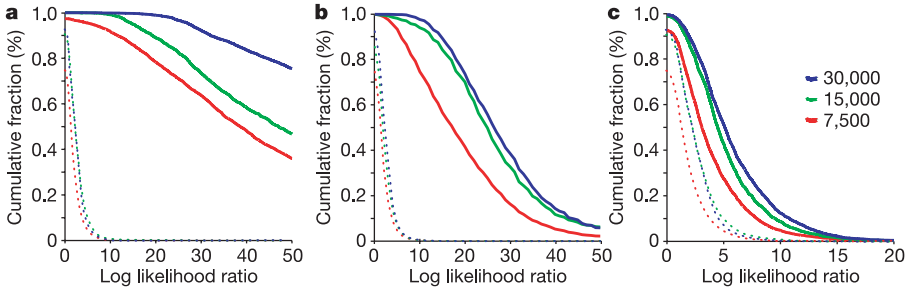

Figure 9 | Power to detect a disease locus by association mapping. One SNP was designated as a disease allele under one of three genetic models: (**a**) simple mendelian dominant, (**b**) fivefold multiplicative increase in risk and (**c**) twofold multiplicative increase in risk. SNP genotypes across surrounding chromosomal regions of 1 Mb were simulated, using the coalescent model corresponding to observed within-breed variation (see text). Diploid genotypes across the chromosomal region were then generated for 100 affected and 100 unaffected dogs, based on the disease model, and association analysis was performed to detect the presence of the disease allele. The distribution of the maximum LOD score across the 1-Mb region is shown for analyses based on multi-SNP haplotypes (solid lines) with SNP densities equivalent to a genome-wide map with a total of 7,500 (red), 15,000 (green) or 30,000 (blue) SNPs. Dotted curves show the null distribution for a genome-wide search in which no disease locus is present (see Supplementary Information). A LOD score of 5 corresponds to <3% chance of a false positive across the genome. For this threshold, the power to detect a disease allele that increases risk by twofold using haplotype analysis and a map with 15,000 SNPs is ~50%.

coding. The dog genome is now further clarifying this picture, as our data suggest that this ~5% represents functional elements common to all mammals. The distribution of these elements relative to genes is highly heterogeneous, with roughly half of the most highly conserved non-coding elements apparently devoted to regulating ~1% of human genes; these genes have important roles in development, and understanding the regulatory clusters that surround them may reveal how cellular states are established and maintained. In recent papers[32,108], the dog genome sequence has been used to greatly expand the catalogue of mammalian regulatory motifs in promoters and 3′-untranslated regions. The dog genome sequence is also being used to substantially revise the human gene catalogue. Despite these advances, it is clear that mammalian comparative genomics is still in its early stages. Progress will be markedly accelerated by the availability of many additional mammalian genome sequences, initially with light coverage[28] but eventually with near-complete coverage.

In addition to its role in studies of mammalian evolution, the dog has a special role in genomic studies because of the unparalleled phenotypic diversity among closely related breeds. The dog is a testament to the power of breeding programmes to select naturally occurring genetic variants with the ability to shape morphology, physiology and behaviour. Genome comparison within and across breeds can reveal the genes that underlie such traits, informing basic research on development and neurobiology. It can also identify disease genes that were carried along in breeding programmes. Potential benefits include insights into disease mechanism, and the possibility of clinical trials in disease-affected dogs to accelerate new therapeutics that would improve health in both dogs and humans. The SNP map of the dog genome confirms that dog breeds show the long-range haplotype structure expected from recent intensive breeding. Moreover, our analysis shows that the current collection of >2.5 million SNPs should be sufficient to allow association studies of

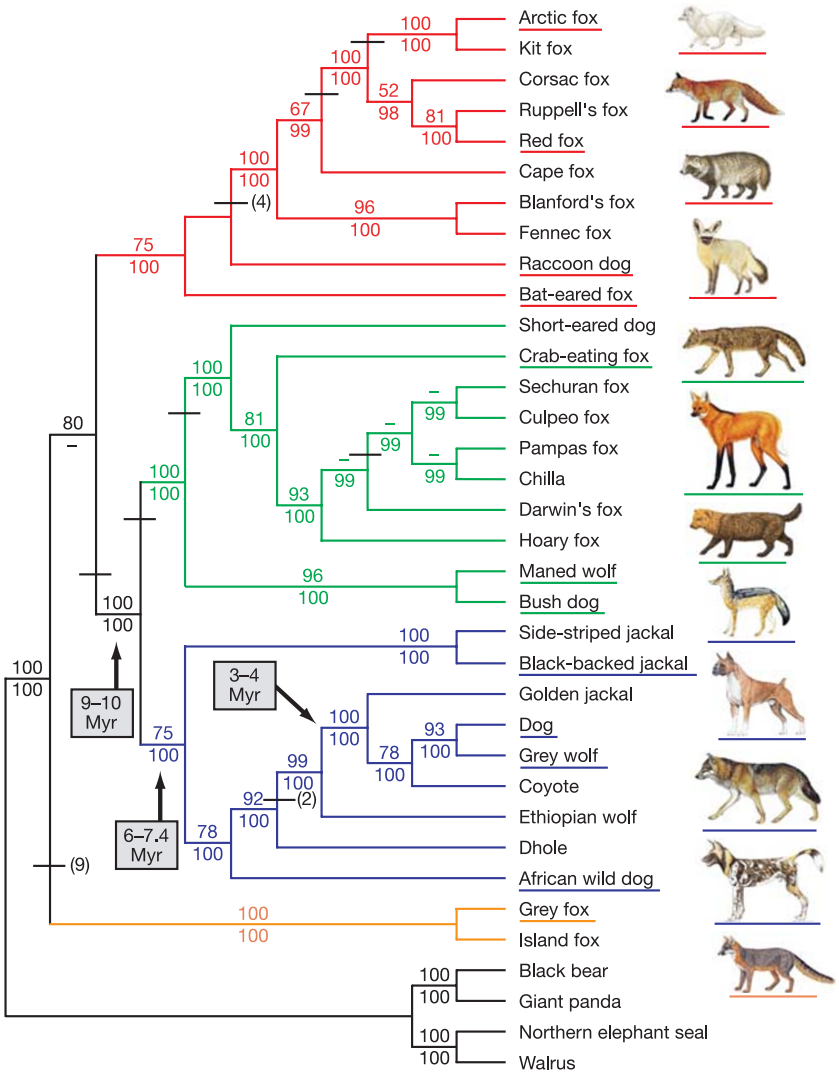

Figure 10 | Phylogeny of canid species. The phylogenetic tree is based on ~15 kb of exon and intron sequence (see text). Branch colours identify the red-fox-like clade (red), the South American clade (green), the wolf-like clade (blue) and the grey and island fox clade (orange). The tree shown was constructed using maximum parsimony as the optimality criterion and is the single most parsimonious tree. Bootstrap values and bayesian posterior probability values are listed above and below the internodes, respectively; dashes indicate bootstrap values below 50% or bayesian posterior probability values below 95%. Horizontal bars indicate indels, with the number of indels shown in parentheses if greater than one. Underlined species names are represented with corresponding illustrations. (Copyright permissions for illustrations are listed in the Supplementary Information.) Divergence time, in millions of years (Myr), is indicated for three nodes as discussed in ref. 1. For scientific names and species descriptions of canids, see ref. 119. A tree based on bayesian inference differs from the tree shown in two respects: it groups the raccoon dog and bat-eared fox as sister taxa, and groups the grey fox and island fox as basal to the clade containing these sister taxa. However, neither of these topological differences is strongly supported (see text and Supplementary Information).

nearly any trait in any breed. Realizing the full power of dog genetics now awaits the development of appropriate genotyping tools, such as multiplex 'SNP chips'[109]—this is already underway. For millennia, dogs have accompanied humans on their travels. It is only fitting that the dog should also be a valued companion on our journeys of scientific discovery.

METHODS

Detailed descriptions of all methods are provided in the Supplementary Information. Links to all of the data can be obtained via the Broad Institute website (http://www.broad.mit.edu/tools/data.html).

WGS sequencing and assembly. Approximately 31.5 million sequence reads were derived from both ends of inserts (paired-end reads) from 4-, 10-, 40- and 200-kb clones, all prepared from primary blood lymphocyte DNA from a single female boxer. This particular animal was chosen for sequencing because it had the lowest heterozygosity rate among ~120 dogs tested at a limited set of loci; subsequent analysis showed that the genome-wide heterozygosity rate in this boxer is not substantially different from other breeds[91]. The assembly was carried out using an interim version of ARACHNE2+ (http://www.broad.mit.edu/wga/).

Genome alignment and comparison. Synteny maps were generated using standard methods[24] from pair-wise alignments of repeat masked assemblies using PatternHunter[110] on CanFam2.0. All other comparative analyses were performed on BLASTZ/MULTIZ[111,112] genome-wide alignments obtained from the UCSC genome browser (http://genome.ucsc.edu), based on CanFam1.0. Known interspersed repeats were identified and dated using RepeatMasker and DateRepeats[113]. The numbers of orthologous nucleotides were counted directly from the alignments using human (hg17) as the reference sequence for all overlaps except the dog–mouse overlap, for which pair-wise (CanFam1.0, mm5) alignments were used.

Divergence rate estimates. Orthologous ancestral repeats were excised from the genome alignment and realigned with the corresponding RepBase consensus using ClustalW. Nucleotide divergence rates were estimated from concatenated repeat alignments using baseml with the REV substitution model[114]. Orthologous coding regions were excised from the genome alignments using the annotated human coding sequences (CDS) from Ensembl and the UCSC browser Known Genes track (October 2004) as reference. K_A and K_S were estimated for each orthologue triplet using codeml with the F3 × 4 codon frequency model and no additional constraints.

Detection and clustering of sequence conservation. Pair-wise conservation scores and the fraction of orthologous sequences under purifying selection were estimated as in ref. 24. The three-way conservation score S_{HMD} was defined as $S_{HMD} = (p - u)/\sqrt{(u(1-u)/n}$, where n is the number of nucleotides aligned across all three genomes (human, mouse, dog) for each non-overlapping 50-bp window with more than 20 aligned bases, p is the fraction of nucleotides identical across all three genomes, and u is the mean identity of ancestral repeats within 500 kb of the window. HCNEs were defined as windows with $S_{HMD} > 5.4$ that did not overlap a coding exon, as defined by the UCSC Known Genes track, and HCNE clusters were defined as all runs of overlapping 1-Mb intervals (50-kb step size) across the human genome with HCNE densities in the 90th percentile.

Gene set acceleration scores. Gene annotation was performed on CanFam1.0. A set of 13,816 orthologous human, mouse and dog genes were identified and compiled into 4,950 gene sets containing genes related by functional annotations or microarray gene expression data. For each gene set S, the acceleration score $A(S)$ along a lineage is defined by (1) ranking all genes based on K_A/K_S within a lineage, (2) calculating the rank-sum statistic for the set along each lineage (denoted $a_{dog}(S)$, $a_{mouse}(S)$, $a_{human}(S)$), (3) calculating the rank-sum for the lineage minus the maximum rank-sum the other lineages, for example, $a_{human}(S)$–max($a_{dog}(S)$, $a_{mouse}(S)$) and (4) converting this rank-sum difference to a z-score by comparing it to the mean and standard deviation observed in 10,000 random sets of the same size. The expected number of sets at a given z-score threshold was estimated by repeating steps (1)–(4) 10,000 times for groups of 4,950 randomly permuted gene sets.

SNP discovery. The SNP discovery was performed on CanFam2.0. Set 1 SNPs were discovered by comparison of the two haplotypes derived from the boxer assembly using only high-quality discrepancies supported by two reads. SNPs in sets 2 and 3 were discovered by aligning reads or contigs to the boxer assembly and using the SSAHA SNP algorithm[115].

Haplotype structure. The SNPs within the sequenced boxer genome (CanFam2.0) were assigned to homozygous or heterozygous regions using a Viterbi algorithm[116]. To determine whether the haplotype structure seen in the boxer is representative of most dog breeds, we randomly selected ten regions of 15 Mb each (~6% of the CanFam2.0 genome) and examined the extent of homozygosity and linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of 10 breeds (akita, basenji, bullmastiff, English springer spaniel, Glen of Imaal terrier, golden retriever, Irish wolfhound, Labrador retriever, pug and rottweiler) and one dog from each of 24 additional breeds (see Supplementary Information). For each instance in which a dog was homozygous in a particular 10-kb region, we measured the distance from the beginning of the 10-kb region to the first heterozygous SNP in the adjoining 100-kb, 1-Mb and 15-Mb data. This distance was used as the extent of homozygosity. The boxer sequence was sampled in an identical manner to the actual breed data. Linkage disequilibrium (represented by r^2) across the ten 15-Mb regions was assessed using Haploview[117].

Received 9 August; accepted 11 October 2005.

1. Wayne, R. K. *et al.* Molecular systematics of the Canidae. *Syst. Biol.* **46,** 622–653 (1997).
2. Vila, C. *et al.* Multiple and ancient origins of the domestic dog. *Science* **276,** 1687–1689 (1997).
3. Bardeleben, C., Moore, R. L. & Wayne, R. K. Isolation and molecular evolution of the selenocysteine tRNA (*Cf TRSP*) and RNase P RNA (*Cf RPPH1*) genes in the dog family, Canidae. *Mol. Biol. Evol.* **22,** 347–359 (2005).
4. Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298,** 1610–1613 (2002).
5. American Kennel Club. *The Complete Dog Book* (eds Crowley, J. & Adelman, B.) (Howell Book House, New York, 1998).
6. Wayne, R. K. Limb morphology of domestic and wild canids: the influence of development on morphologic change. *J. Morphol.* **187,** 301–319 (1986).
7. Ostrander, E. A., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends Genet.* **16,** 117–123 (2000).
8. Patterson, D. Companion animal medicine in the age of medical genetics. *J. Vet. Intern. Med.* **14,** 1–9 (2000).
9. Sargan, D. R. IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm. Genome* **15,** 503–506 (2004).
10. Chase, K. *et al.* Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl Acad. Sci. USA* **99,** 9930–9935 (2002).
11. Breen, M. *et al.* Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res.* **11,** 1784–1795 (2001).
12. Breen, M., Bullerdiek, J. & Langford, C. F. The DAPI banded karyotype of the domestic dog (Canis familiaris) generated using chromosome-specific paint probes. *Chromosome Res.* **7,** 401–406 (1999).
13. Breen, M. *et al.* An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5,** 65 (2004).
14. Hitte, C. *et al.* Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nature Rev. Genet.* **6,** 643–648 (2005).
15. Li, R. *et al.* Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58,** 9–17 (1999).
16. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301,** 1898–1903 (2003).
17. Sutter, N. & Ostrander, E. Dog star rising: The canine genetic system. *Nature Rev. Genet.* **5,** 900–910 (2004).
18. Galibert, F., Andre, C. & Hitte, C. Dog as a mammalian genetic model [in French]. *Med. Sci. (Paris)* **20,** 761–766 (2004).
19. Pollinger, J. P. *et al.* Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* doi:10.1101/gr.4374505 (in the press).
20. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
21. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
22. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
23. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
24. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
25. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
26. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614–618 (2001).
27. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
28. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102,** 4795–4800 (2005).
29. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).

30. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
31. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** e10 (2005).
32. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434,** 338–345 (2005).
33. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14,** 852–859 (2004).
34. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
35. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
36. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8,** 251–259 (1998).
37. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5,** R23 (2004).
38. Andelfinger, G. *et al.* Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* **83,** 1053–1062 (2004).
39. Wang, W. & Kirkness, E. F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* doi:10.1101/gr.3765505 (in the press).
40. Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B. & Sverdlov, E. D. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res.* **33,** e16 (2005).
41. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin* (*orexin*) *receptor 2* gene. *Cell* **98,** 365–376 (1999).
42. Pele, M., Tiret, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14,** 1417–1427 (2005).
43. Fondon, J. W. III & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA* **101,** 18058–18063 (2004).
44. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* **150,** 1577–1584 (1998).
45. Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58,** 653–660 (2004).
46. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8,** 1499–1504 (1980).
47. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90,** 11995–11999 (1993).
48. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13,** 813–820 (2003).
49. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101,** 13994–14001 (2004).
50. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90,** 4087–4091 (1993).
51. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102,** 140–145 (2005).
52. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224,** 149–154 (1969).
53. Li, W. H., Tanimura, M. & Sharp, P. M. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25,** 330–342 (1987).
54. Webber, C. & Ponting, C. P. Hot spots of mutation and breakage in dog and human chromosomes. *Genome Res.* doi:10.1101/gr.3896805 (in the press).
55. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
56. Marques-Bonet, T. & Navarro, A. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* **353,** 147–154 (2005).
57. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56 (2004).
58. Smith, N. G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84,** 806–813 (2004).
59. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
60. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145 (2005).
61. Walter, K., Abnizova, I., Elgar, G. & Gilks, W. R. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.* **21,** 436–440 (2005).
62. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
63. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
64. Kimura-Yoshida, C. *et al.* Characterization of the pufferfish *Otx2* cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131,** 57–71 (2004).
65. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4,** 509–519 (2003).
66. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res.* **15,** 1061–1072 (2005).
67. Daly, M. J. Estimating the human gene count. *Cell* **109,** 283–284 (2002).
68. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106,** 413–415 (2001).
69. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12,** 701–709 (2003).
70. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234 (2000).
71. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nature Genet.* **33** (suppl.), 255–265 (2003).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73,** 823–834 (2003).
73. Hughes, A. L. The evolution of the type I interferon gene family in mammals. *J. Mol. Evol.* **41,** 539–548 (1995).
74. Hurst, L. D. The *Ka/Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18,** 486 (2002).
75. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99,** 16134–16137 (2002).
76. Demetrius, L. Directionality theory and the evolution of body size. *Proc. Biol. Sci.* **267,** 2385–2391 (2000).
77. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* **4,** 213–235 (2003).
78. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101,** 4164–4169 (2004).
79. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34,** 267–273 (2003).
80. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
81. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119,** 1027–1040 (2004).
82. Saetre, P. *et al.* From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res. Mol. Brain Res.* **126,** 198–206 (2004).
83. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403,** 304–309 (2000).
84. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3,** 262–273 (2002).
85. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. *Nature Genet.* **36,** 1326–1329 (2004).
86. Ruiz-Pesini, E. *et al.* Correlation of sperm motility with mitochondrial enzymatic activities. *Clin. Chem.* **44,** 1616–1620 (1998).
87. Zeh, J. A. & Zeh, D. W. Maternal inheritance, sexual conflict and the maladapted male. *Trends Genet.* **21,** 281–286 (2005).
88. Grossman, L. I., Wildman, D. E., Schmidt, T. R. & Goodman, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet.* **20,** 578–585 (2004).
89. Ostrander, E. A. & Kruglyak, L. Unleashing the canine genome. *Genome Res.* **10,** 1271–1274 (2000).
90. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **12,** 2388–2396 (2004).
91. Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304,** 1160–1164 (2004).
92. Bardeleben, C., Moore, R. L. & Wayne, R. K. A molecular phylogeny of the Canidae based on six nuclear loci. *Mol. Phylogenet. Evol.* **37,** 815–831 (2005).
93. Fogel, B. *The Encyclopedia of the Dog* (D.K. Publishing, New York, 1995).
94. Wilcox, B. & Walkowicz, C. *The Atlas of Dog Breeds of the World* (T.H.F. Publications, Neptune City, New York, 1995).
95. Frazer, K. A. *et al.* Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res.* **14,** 1493–1500 (2004).
96. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
97. Vila, C., Seddon, J. & Ellegren, H. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet.* **21,** 214–218 (2005).

98. Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298,** 1613–1616 (2002).
99. Kajiwara, N. & Japanese Kennel Club in *Akita* (eds Kariyabu, T. & Kaluzniacki, S.) 1–103 (Japan Kennel Club, Tokyo, 1998).
100. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
101. Werner, P., Raducha, M. G., Prociuk, U., Henthorn, P. S. & Patterson, D. F. Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. *Genomics* **42,** 74–82 (1997).
102. Todhunter, R. J. *et al.* Power of a Labrador Retriever-Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. *Am. J. Vet. Res.* **64,** 418–424 (2003).
103. Sidjanin, D. J. *et al.* Canine *CNGB3* mutations establish cone degeneration as orthologous to the human achromatopsia locus *ACHM3*. *Hum. Mol. Genet.* **11,** 1823–1833 (2002).
104. Lou, X. Y. *et al.* The extent and distribution of linkage disequilibrium in a multi-hierarchic outbred canine pedigree. *Mamm. Genome* **14,** 555–564 (2003).
105. Hyun, C. *et al.* Prospects for whole genome linkage disequilibrium mapping in domestic dog breeds. *Mamm. Genome* **14,** 640–649 (2003).
106. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19,** 135–140 (2003).
107. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31,** 4910–4916 (2003).
108. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
109. Syvanen, A. C. Toward genome-wide SNP genotyping. *Nature Genet.* **37** (suppl.), S5–10 (2005).
110. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18,** 440–445 (2002).
111. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
112. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
113. Smit, A. F. A. & Green, P. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html).
114. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11,** 316–324 (1994).
115. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
116. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Process.* **13,** 260–269 (1967).
117. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
118. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
119. Macdonald, D. W. & Sillero-Zubiri, C. in *Biology and Conservation of Canids* (eds Macdonald, D. W. & Sillero-Zubiri, C.) 1–30 (Oxford Univ. Press, Oxford, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the canine research community, and in particular D. Patterson, G. Acland and K. G. Lark, whose vision and research convinced the NIH of the importance of generating a canine genome sequence. We also thank all those who shared insights at the Dog Genome Community meetings, including G. Acland, G. D. Aguirre, M. Binns, U. Giger, P. Henthorn, F. Lingaas, K. Murphy and P. Werner. We thank our many colleagues (G. Acland, G. D. Aguirre, C. Andre, N. Fretwell, G. Johnson, K. G. Lark and J. Modiano), as well as the dog owners and breeders who provided us with samples. We thank colleagues at the UCSC browser for providing data (such as BLASTZ alignments), A. Smit for providing the RepeatMasker annotations used in our analyses and N. Manoukis for providing Unix machines for the phylogenetic analyses. Finally, we thank L. Gaffney and K. Siang Toh for editorial and graphical assistance. The genome sequence and analysis was supported in part by the National Human Genome Research Institute. The radiation hybrid map was supported in part by the Canine Health Foundation. Sample collection was supported in part by the Intramural Research Program of the National Human Genome Research Institute and the Canine Health Foundation.

Author Information The draft genome sequence has been deposited in public databases under NCBI accession codes AAEX01000000 (CanFam1.0) and AAEX02000000 (CanFam2.0). SNPs have been deposited in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L.T. (kersli@broad.mit.edu) or E.S.L. (lander@broad.mit.edu).

Broad Sequencing Platform members Jennifer Baldwin[1], Adal Abebe[1], Amr Abouelleil[1], Lynne Aftuck[1], Mostafa Ait-zahra[1], Tyler Aldredge[1], Nicole Allen[1], Peter An[1], Scott Anderson[1], Claudel Antoine[1], Harindra Arachchi[1], Ali Aslam[1], Laura Ayotte[1], Pasang Bachantsang[1], Andrew Barry[1], Tashi Bayul[1], Mostafa Benamara[1], Aaron Berlin[1], Daniel Bessette[1], Berta Blitshteyn[1], Toby Bloom[1], Jason Blye[1], Leonid Boguslavskiy[1], Claude Bonnet[1], Boris Boukhgalter[1], Adam Brown[1], Patrick Cahill[1], Nadia Calixte[1], Jody Camarata[1], Yama Cheshatsang[1], Jeffrey Chu[1], Mieke Citroen[1], Alville Collymore[1], Patrick Cooke[1], Tenzin Dawoe[1], Riza Daza[1], Karin Decktor[1], Stuart DeGray[1], Norbu Dhargay[1], Kimberly Dooley[1], Kathleen Dooley[1], Passang Dorje[1], Kunsang Dorjee[1], Lester Dorris[1], Noah Duffey[1], Alan Dupes[1], Osebhajajeme Egbiremolen[1], Richard Elong[1], Jill Falk[1], Abderrahim Farina[1], Susan Faro[1], Diallo Ferguson[1], Patricia Ferreira[1], Sheila Fisher[1], Mike FitzGerald[1], Karen Foley[1], Chelsea Foley[1], Alicia Franke[1], Dennis Friedrich[1], Diane Gage[1], Manuel Garber[1], Gary Gearin[1], Georgia Giannoukos[1], Tina Goode[1], Audra Goyette[1], Joseph Graham[1], Edward Grandbois[1], Kunsang Gyaltsen[1], Nabil Hafez[1], Daniel Hagopian[1], Birhane Hagos[1], Jennifer Hall[1], Claire Healy[1], Ryan Hegarty[1], Tracey Honan[1], Andrea Horn[1], Nathan Houde[1], Leanne Hughes[1], Leigh Hunnicutt[1], M. Husby[1], Benjamin Jester[1], Charlien Jones[1], Asha Kamat[1], Ben Kanga[1], Cristyn Kells[1], Dmitry Khazanovich[1], Alix Chinh Kieu[1], Peter Kisner[1], Mayank Kumar[1], Krista Lance[1], Thomas Landers[1], Marcia Lara[1], William Lee[1], Jean-Pierre Leger[1], Niall Lennon[1], Lisa Leuper[1], Sarah LeVine[1], Jinlei Liu[1], Xiaohong Liu[1], Yeshi Lokyitsang[1], Tashi Lokyitsang[1], Annie Lui[1], Jan Macdonald[1], John Major[1], Richard Marabella[1], Kebede Maru[1], Charles Matthews[1], Susan McDonough[1], Teena Mehta[1], James Meldrim[1], Alexandre Melnikov[1], Louis Meneus[1], Atanas Mihalev[1], Tanya Mihova[1], Karen Miller[1], Rachel Mittelman[1], Valentine Mlenga[1], Leonidas Mulrain[1], Glen Munson[1], Adam Navidi[1], Jerome Naylor[1], Tuyen Nguyen[1], Nga Nguyen[1], Cindy Nguyen[1], Thu Nguyen[1], Robert Nicol[1], Nyima Norbu[1], Choe Norbu[1], Nathaniel Novod[1], Tenchoe Nyima[1], Peter Olandt[1], Barry O'Neill[1], Keith O'Neill[1], Sahal Osman[1], Lucien Oyono[1], Christopher Patti[1], Danielle Perrin[1], Pema Phunkhang[1], Fritz Pierre[1], Margaret Priest[1], Anthony Rachupka[1], Sujaa Raghuraman[1], Rayale Rameau[1], Verneda Ray[1], Christina Raymond[1], Filip Rege[1], Cecil Rise[1], Julie Rogers[1], Peter Rogov[1], Julie Sahalie[1], Sampath Settipalli[1], Theodore Sharpe[1], Terrance Shea[1], Mechele Sheehan[1], Ngawang Sherpa[1], Jianying Shi[1], Diana Shih[1], Jessie Sloan[1], Cherylyn Smith[1], Todd Sparrow[1], John Stalker[1], Nicole Stange-Thomann[1], Sharon Stavropoulos[1], Catherine Stone[1], Sabrina Stone[1], Sean Sykes[1], Pierre Tchuinga[1], Pema Tenzing[1], Senait Tesfaye[1], Dawa Thoulutsang[1], Yama Thoulutsang[1], Kerri Topham[1], Ira Topping[1], Tsamla Tsamla[1], Helen Vassiliev[1], Vijay Venkataraman[1], Andy Vo[1], Tsering Wangchuk[1], Tsering Wangdi[1], Michael Weiand[1], Jane Wilkinson[1], Adam Wilson[1], Shailendra Yadav[1], Shuli Yang[1], Xiaoping Yang[1], Geneva Young[1], Qing Yu[1], Joanne Zainoun[1], Lisa Zembek[1] & Andrew Zimmer[1]

Affiliation for participants: [1]Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA.