The complete genome sequence of a dog: a perspective

Soohyun Lee[1,3]* and Simon Kasif[1,2,3,4]

Summary

A complete, high-quality reference sequence of a dog genome was recently produced by a team of researchers led by the Broad Institute, achieving another major milestone in deciphering the genomic landscape of mammalian organisms. The genome sequence provides an indispensable resource for comparative analysis and novel insights into dog and human evolution and history. Together with the survey sequence of a poodle previously published in 2003, the two dog genome sequences allowed identification of more than 2.5 million single nucleotide polymorphisms within and between dog breeds, which can be used in evolutionary analysis, behavioral studies and disease gene mapping.[1] ***BioEssays*** **28:569–573, 2006.** **© 2006 Wiley Periodicals, Inc.**

"The dog represents all that is best in man"

—Etienne Charlet

Introduction

The continuing reductions in cost and complexity of sequencing due to improvements in biotechnology and computational methods allow rapid and high-quality production of large eukaryotic genomes. A mammalian genome can be completely sequenced for $40–50 million in about a year.[2] High-coverage sequencing of the canine genome was proposed in 2003 and the completed sequence became publicly available in July 2004 (Genbank (http://www.ncbi.nlm.nih.gov/) and UCSC genome browser (http://genome.ucsc.edu/)). An official report on the sequencing and analysis of this draft canine sequence was published December of last year in Nature.[1] The domestic dog (*Canis familiaris*) is the fifth completely sequenced mammal following the human,[3,4] mouse,[5] chimpanzee[6] and rat.[7]

A female boxer was chosen to be sequenced because of its apparent high homozygosity, and the sequencing was done using a whole genome shotgun approach (WGS), which was first deployed at such a large scale in the human genome sequencing by Celera.[3] WGS involves redundant sequencing of random fragments of the whole genome and assembly of the sequenced fragments, in contrast to hierarchical genome sequencing, which involves hierarchical fragmentation and assembly. The dog genome was sequenced to 7.5 × redundancy, which covered about 99% of the genome excluding highly repetitive heterochromatin regions.

The assembled sequence is of very high quality. The error rate is less than 10^{-4} and half of the bases belong to continuous scaffolds (supercontigs) larger than 45 Mb (N50 = 45 Mb). This is an enormous improvement compared to human (N50 = 8.4 Mb)[4] and mouse (N50 = 16.9 Mb).[5] Such a high-quality draft dog genome sequence can be used confidently in a variety of studies including identification of mammalian genomic features and their conservation, whole-genome association mapping of canine genetic diseases, history of dogs and comparative analysis of the human genome.

The dog sequence helps us learn more about us the human

Mouse has served as an excellent model organism for studying genetics and molecular biology of mammals and its genome sequence has facilitated understanding the human genome.[5,8–10] The dog provides additional information as a distinct group of mammal. As an outgroup to both human and mouse (Fig. 1), dogs can serve as a reference for comparison between human and mouse.[1,11] Dogs belong to a mammalian clade Laurasiatheria, whereas both human and mouse are Euarchontoglires. Laurasiatheria diverged from Euarchontoglires about 89–98 million years ago (MYA),[12] whereas the last shared common ancestor of human and mouse is dated roughly 75 MYA.[5] Since mouse has evolved relatively rapidly from the common ancestor, dogs share more similarity with human than mouse.

[1]Bioinformatics Program, Boston University, Boston, MA.
[2]Department of Biomedical Engineering, Boston University, Boston, MA.
[3]Center for Advanced Genomic Technology, Boston University, Boston, MA.
[4]Children's Hospital Informatics Program at Harvard-MIT Health Sciences and Technology, Boston, MA.
*Correspondence to: Soohyun Lee, Bioinformatics Program, Boston University, 24 Cummington St. Room 948, Cube 6, Boston, MA 02215.
E-mail: duplexa@bu.edu
DOI 10.1002/bies.20421
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: ETC, electron transport chain; LD, linkage disequilibrium; mtDNA, mitochondrial DNA; SNP, single nucleotide polymorphism; WGS, whole genome shotgun.

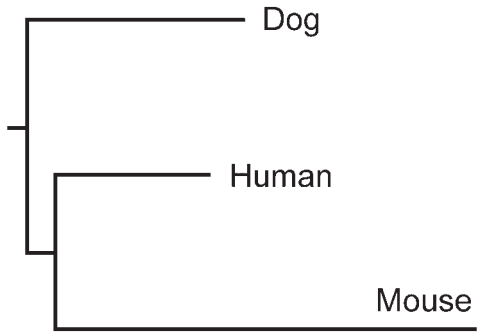

Figure 1. Phylogeny of dog,human and mouse. The length of each branch represents evolutionary rate. Human and mouse have a more recent common ancestor than dog, but the rodent lineage evolves more rapidly. Adapted by permission from Macmillan Publishers Ltd: Nature (1), copyright (2005).

A previous study had suggested that 24 brain development genes seem to have undergone accelerated evolution in humans compared to rodents.[13] A similar analysis with the dog genome sequence suggested that the evolution of at least 18 brain genes were accelerated in the dog lineage as well, compared to rodents. This may be due to a decelerated evolution of brain genes in rodents and suggests mouse only may provide a biased reference to human evolution.[1] Compared to both dog and mouse, changes in genes involved in reproductive competition (testis-specific genes and mitochondrial ETC subunits that provide the source of energy in sperm motility) seem to have been particularly accelerated in the lineage leading to humans, indicating that the major evolutionary drive for primates is sexual selection rather than intellectual improvement.

Alternatively, dogs and humans may have co-evolved, with similar genes changing together. Brain genes may have been co-accelerated in both human and dog compared to mouse. Many dog experts agree that dogs must have undergone selection to be compatible with humans.[14] Hare et al. observed that dogs, even those with minimal interaction with humans, could follow human clues such as finger pointing, whereas wolves and chimps could not.[15] Kukekova et al. compared tamed and aggressive silver foxes and identified serotonin receptor genes as part of their genetic differences.[16] The tamed foxes were also morphologically similar to dogs. Apparently, humans have affected dog evolution through both artificial selection and environmental sharing. In this sense, one must be careful in interpreting any dog–human similarity compared to other animals.

History of dogs and implication in phylogenetics

Dogs belong to a 50-million-year-old family Canidae, which also includes wolves, foxes, coyotes and jackals. It has been widely suggested that dogs have been domesticated from other canids, most likely the grey wolf.[17,18] A study suggests that this domestication occurred in East Asia as a common practice.[19] Participating closely in human activities, dogs have undergone population bottlenecks (shrinkage in population size) during the World Wars, and substantial artificial selection by human breeding programs. An analysis of the remains of an ancient dog suggests that humans brought dogs with them when they moved to America for the first time 20,000–25,000 years ago. These ancient dogs seem to have been mostly replaced by European dogs during or after the colonization in the 16th century.[20] Thus, examining dog's history is interesting in this sense—in that it is very closely related to our own history.

In contrast to the relatively early divergence from other carnivores, the canid family shows great familial kinship, with an estimated time of common ancestry of the extant species within the last 10 million years.[21] Evolutionary analyses based on molecular data among closely-related species or populations is often difficult, primarily because their DNA sequences are too similar. Mitochondrial DNA (mtDNA) sequences such as cytochrome b genes and 'control regions' are often used, since they evolve rapidly enough to retrieve information about the evolutionary or geneological history.[22] The previously studied history of dogs has also been almost exclusively based on mtDNAs.[17,19,20,23] However, mtDNA is of maternal inheritance and does not contain information about male-mediated evolutionary events. Nuclear DNAs contain both maternal and paternal information, but approaches using them have been limited.

A whole genome sequence enables a screen of rapidly evolving nuclear DNA regions. Lindblad-Toh et al.[1] identified twelve fast-evolving exons, based on K_a/K_s, a measure of acceleration in amino acid sequence change compared to silent nucleotide substitutions, and four fast-evolving introns with particularly high SNP rates (~5-fold) compared to background. The resulting phylogenetic tree was similar to those obtained in previous studies.[18,23] Dogs were grouped with grey wolves and then with coyotes. Golden jackals were the next most closely related canid, instead of simien jackals (Ethiopian wolf) as indicated by the mitochondrial data. Such a discrepancy may be explained by a paternal gene flow. African canids were the basal members of the wolf-like group (grey wolf, dog, coyote and some jackals) as in the previous analysis, suggesting an African origin of this clade. Considering possible hybridizations among canids,[18] a non-tree-based approach may also be interesting, such as a clustering method, which proved to work nicely in revealing dog breed history that cannot be explained well with a tree model.[24]

On a smaller scale, genealogical histories such as population bottlenecks can be analyzed using polymorphic loci and their non-random association in the population, called linkage disequilibrium (LD). Alleles of nearby loci tend to be inherited together and this tendency decreases with distance

between the loci. LD is the tendency of co-inheritance captured at the population level. LD patterns are affected by population size, selection and recombination rate and thus can be used to infer these variables. Lindblad-Toh et al.[1] discovered more than 2.5 million loci of single-base differences (SNPs) in the dog genome and used them to analyse the dog's LD and population history. Their results, which are based on both a stochastic (coalescent)[25] and a deterministic model, suggest that a two-bottleneck model shows a good fit to many dog breeds. The first ancient bottleneck occurred around 27,000 years ago, probably due to the domestication events, and the second was around the time when the practice of breeding began (Fig. 2). Some breed-specific bottleneck patterns were detected that were broadly consistent with each breed history. For example, a Japanese breed Akita shows an additional bottleneck about 20 generations ago (near 1940–1950), probably due to the World War II and introduction to the United States. The severity of bottleneck also varies among breeds. For example, Labrador retrievers, a North American breed with relatively continuous popularity showed less severe bottlenecks in the analysis.

Dog genetics

Knowledge of dog diseases and genes responsible for disease susceptibility can be greatly enhanced by the dense map of SNPs and the entire genome space that provide the potential for discovery of other genetic markers. Dogs are an excellent system for studying genetics. The unique breeding history of dogs greatly improves the statistical power of disease locus mapping.[26] Being highly inbred and homozygous, dogs have long-range linkage disequilibrium (LD) within each breed (50-fold larger than human).[1,27] This means that the dependency among SNPs spans wider distances, and therefore fewer representative SNPs are needed to cover the whole genome. Lindblad-Toh et al[1] estimates that about 10,000 evenly spaced SNPs will be required, compared to more than 300,000 SNPs in human. Using about 15,000 SNPs, the probability of detecting a simple Mendelian dominant trait locus is over 99%. Complex traits (e.g. quantitative traits or multi-gene traits) can also be effectively studied in dogs. In an analysis using a Portuguese water dog, two loci that determine skeletal shape were identified using a simple principal component analysis.[28]

Dog genetics help understand human genetic diseases as well as canine diseases. Many genetic diseases are shared between dogs and humans. The top canine diseases include cancer, heart disease and autoimmune diseases, as in human. Cloning and mapping narcolepsy,[29] hereditary kidney cancer[30] and blindness[31] genes in dog have accelerated the identification of similar genetic predispositions in human.

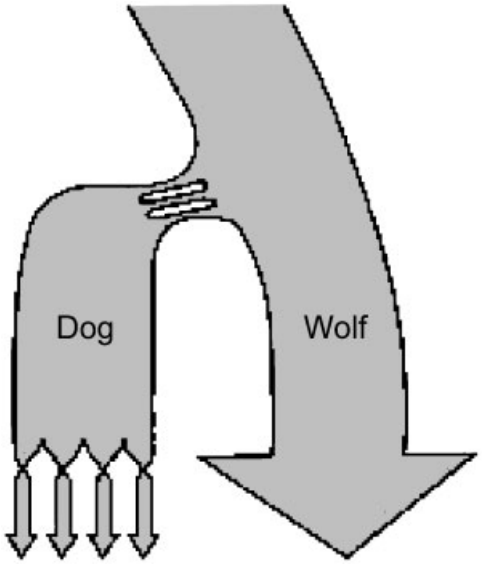

Figure 2. Two population bottlenecks of dog. A two-bottleneck model fits many dog breeds. The first ancient bottleneck occurred around the time of domestication from wolves, and the second one occurred when the practice of breeding began. Adapted by permission from Macmillan Publishers Ltd: Nature (1), copyright (2005).

Two dog genome papers

Prior to the release of the complete sequence, there was another independent dog genome sequencing project, which provided 77% (1.5 × coverage) of a male poodle genome sequence.[11] In addition to the obvious benefits of obtaining a full sequence, the comparison between the low-coverage survey and the high-coverage complete sequences allows complete analysis of the cost–utility trade-offs in producing a complete sequence. The majority of global genome-scale statistics had been estimated based on this survey sequence, and the complete genome sequence confirmed these estimates. For example, the canine genome spans about 2.4Gb, which is smaller than both human (2.9Gb) and mouse (2.5Gb) genomes. This is in part due to fewer repeat sequences in the dog genome compared to human and mouse (dog 34%, human 46% and mouse 40%). The survey sequence paper reconstructed a phylogeny of human, mouse and dog, assuming neutral evolution of repeat sequences and a similar tree was obtained based on the draft genome sequence. The nucleotide divergence rate in the dog was higher than human, but lower than mouse, partly due to generation times and metabolic rates. Conserved synteny blocks between dog, human and mouse had also been identified using the survey sequence, and the full genome sequence led to a higher resolution synteny map. Nearly 1 million poodle SNP sites were predicted based on the survey sequence, whereas the draft sequence provided a richer repository of SNPs, including within-breed, between-breed and between-canid differences, with the help of low coverage sequencing of other individuals. A

retrotransposon family SINEC_Cf was found to be particularly prevalent in the dog genome. The identification of genomic coordinates of these transposons in the draft sequence allowed a more intensive analysis. Wang & Kirkness[32] exploited this topic further and revealed that about half of the annotated canine genes contain SINEC_Cfs and may occasionally incorporate them as an alternative exon.

The survey paper reported more than 2000 human protein-coding genes whose homologs were missing in the dog genome. A complete genome sequence provided a better picture of these genes. Lindblad-Toh et al.[1] reported 19,300 predicted dog genes where more than 2000 genes were missing compared to 22,000 human genes catalogued in ENSEMBL build.26 (http://nov2004.archive.ensembl.org/Homo_sapiens/), indicating that this difference was not due to incomplete coverage of the survey sequence, but rather gene loss in the canine lineage or erroneous human gene prediction.

A low-cost, low-coverage survey sequence provided an accurate scaffold from which to draw broad comparative conclusions. The limitations of this survey were overcome by sequencing the entire genome. Most eukaryotic genes are large and few genes can be expected to have a complete genomic sequence in a low-coverage approach. Gene-based analyses can benefit greatly from the complete genome sequence. Genetic and evolutionary studies can also be assisted by large-scale SNP and LD maps. In comparative studies of regulatory regions on the genome, a large number of low-redundancy sequences can provide better statistical power in identifying conserved regulatory elements,[33] whereas in-depth studies on long-range functional regions or cooperation among regulatory elements requires a high-continuity genome sequence. The draft dog genome sequence has already been used in several studies on regulatory regions.[34–36] The increasing evidence about the important role of micro-regulation, alternative splicing, RNAi and other subtle mechanisms that control living cells and influence phenotype continues to support the current need for high-coverage sequencing of selected model organisms.

Conclusion

The dog has accompanied humans for at least 10,000–15,000 years.[14,20] It is the first domesticated animal in human history[19] and currently about 36% of households in the US have at least one dog.[26] Being such a close symbiont, dogs have served humans in countless ways, as companions, guards, assistants and entertainers. They have been shown to reduce stress and improve health in the elderly, the sick and the young. The release of the dog sequence allows us to rely on our best friends in new and currently unforeseen ways. Their genomic sequences provide some of the best targets for studying the genetics of behavioral and physical traits. They will hopefully lead to even a greater respect and affection than we already interject into our relationship with dogs. We also believe that the availability of a complete dog genome sequence will enable more preselection and care in the design of highly focused studies of dogs as models of human disease and genetic variation.

The complete genome sequencing has been funded by the National Institute of Health, National Human Genome Research Institute and other agencies, which all deserve endless credit for supporting this influential project.

Acknowledgments

We thank Michael Schaffer, Sharona Thompson, Howook Hwang, Elinor Karlsson, Art Delcher and Tarjei Mikkelsen for valuable discussions and suggestions. We also thank all the researchers that contributed to this remarkable new genomic resource enabling a deeper understanding of molecular function and genetics.

References

1. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.
2. Wayne RK, Ostrander EA. 2004. Out of the dog house: the emergence of the canine genome. Heredity 92:273–274.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. Science 291:1304–1351.
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
5. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
6. Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.
7. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521.
8. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, et al. 2003. Comparative gene prediction in human and mouse. Genome Res 13:108–117.
9. Zhang L, Pavlovic V, Cantor CR, Kasif S. 2003. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. Genome Res 13:1190–1202.
10. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26:225–228.
11. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, et al. 2003. The dog genome: survey sequencing and comparative analysis. Science 301:1898–1903.
12. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100:1056–1061.
13. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, et al. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell 119:1027–1040.
14. Pennisi E. 2002. Canine evolution. A shaggy dog history. Science 298:1540–1542.
15. Hare B, Brown M, Williamson C, Tomasello M. 2002. The domestication of social cognition in dogs. Science 298:1634–1636.
16. Kukekova AV, Trut LN, Oskina IN, Kharlamova AV, Shikhevich SG, et al. 2004. A marker set for construction of a genetic map of the silver fox (Vulpes vulpes). J Hered 95:185–194.
17. Vila C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, et al. 1997. Multiple and ancient origins of the domestic dog. Science 276:1687–1689.

18. Wayne RK. 1993. Molecular evolution of the dog family. Trends Genet 9:218–224.
19. Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T. 2002. Genetic evidence for an East Asian origin of domestic dogs. Science 298:1610–1613.
20. Leonard JA, Wayne RK, Wheeler J, Valadez R, Guillen S, et al. 2002. Ancient DNA evidence for Old World origin of New World dogs. Science 298(5598):1613–1616.
21. Vila C, Maldonado JE, Wayne RK. 1999. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. J Hered 90:71–77.
22. Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. Nat Rev Genet 4:900-910.
23. Wayne RK, Geffen E, Girman DJ, Koepfli KP, Lau LM, et al. 1997. Molecular systematics of the Canidae. Syst Biol 46:622–653.
24. Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, et al. 2004. Genetic structure of the purebred domestic dog. Science 304:1160–1164.
25. Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390.
26. Sutter NB, Ostrander EA. 2004. Dog star rising: the canine genetic system. Nat Rev Genet 5:900–910.
27. Sutter NB, Eberle MA, Parker HG, Pullar BJ, Kirkness EF, et al. 2004. Extensive and breed-specific linkage disequilibrium in Canis familiaris. Genome Res 14:2388–2396.
28. Chase K, Carrier DR, Adler FR, Jarvik T, Ostrander EA, et al. 2002. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. Proc Natl Acad Sci U S A 99:9930–9935.
29. Lin L, Faraco J, Li R, Kadotani H, Rogers W, et al. 1999. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. Cell 98:365–376.
30. Moe L, Lium B. 1997. Hereditary multifocal renal cystadenocarcinomas and nodular dermatofibrosis in 51 German shepherd dogs. J Small Anim Pract 38:498–505.
31. Acland GM, Ray K, Mellersh CS, Gu W, Langston AA, et al. 1998. Linkage analysis and comparative mapping of canine progressive rod-cone degeneration (prcd) establishes potential locus homology with retinitis pigmentosa (RP17) in humans. Proc Natl Acad Sci U S A 95: 3048–3053.
32. Wang W, Kirkness EF. 2005. Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res 15:1798–1808.
33. Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A 102:4795–4800.
34. Lee S, Kohane I, Kasif S. 2005. Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. BMC Genomics 6:168.
35. Kamal M, Xie X, Lander ES. 2006. A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A [In press].
36. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338–345.