(*23*); in contrast, A145R/Y87H dosed as high as 30 mg per kilogram of mouse body weight (along with native TNF at 30 μg/kg) resulted in no mortality or hepatotoxicity (Fig. 4A) (*24*). Similarly, lethal doses of native human TNF (30 μg/kg) mixed before injection with varying ratios of A145R/Y87H produced no TNF-induced damage. This protection was observed at native: variant ratios as low as 1:1 and with a superlethal dose of TNF (Fig. 4A). Further, sandwich ELISA analyses of serum samples indicated that a substantial portion (30 ng/ml) of administered A145R (3 mg/kg) was in heterotrimers with the endogenous mouse TNF at 1 hour.

A145R/Y87H was next assessed in a model of chronic disease as an initial test of the DN-TNF antagonism mechanism in a disease-relevant setting. We selected the rat 7-day established collagen-induced arthritis (CIA) model because it simulates chronic autoimmune joint disease and can be treated by TNF blockade (*25*). When dosed after the onset of symptoms, only interventions with rapid onset of action would be able to affect disease progression in this model, thus requiring rapid exchange in vivo of TNF variants with endogenous TNF. To ensure that there were no confounding in vivo effects of using affinity-tagged variants, we produced A145R/Y87H that lacked such tags. Further, to decrease in vivo clearance, we added one polyethylene glycol (PEG; ~5 kD/molecule) to each monomeric subunit of A145R/Y87H. This modification had no effect on the dominant-negative properties of the molecule in vitro (fig. S7). A145R/Y87H reduced joint swelling in the CIA model when dosed once daily at 2.0 mg/kg subcutaneously with a loading dose of 2.0 mg/kg and twice daily at 10 mg/kg intravenously (Fig. 4B). These results demonstrate the potential of DN-TNFs to inhibit TNF-mediated inflammation and verify that exchange occurs rapidly enough to affect progression of acute symptoms when dosed therapeutically.

Given their high-yield bacterial production, theoretical low immunogenicity, and unique mechanism of action, DN-TNFs show potential as a new class of anti-inflammatory therapy, particularly because existing methodologies (i.e., PEG modification) can be used to further enhance their pharmacokinetic properties (*26*, *27*). Further, we propose that this dominant-negative approach should be tested for its potential to create inhibitors of other multimeric extracellular signaling molecules, in particular other members of the TNF superfamily (e.g., RANKL, CD40L, and BAFF) that have been implicated in human pathophysiology (*28*, *29*).

References and Notes

1. B. B. Aggarwal, A. Samanta, M. Feldmann, in *Cytokine Reference*, J. J. Oppenheim, M. Feldmann, Eds. (Academic Press, London, 2000).
2. G. Chen, D. V. Goeddel, *Science* **296**, 1634 (2002).
3. D. J. MacEwan, *Cell Signal.* **14**, 477 (2002).
4. M. P. Boldin, T. M. Goncharov, Y. V. Goltsev, D. Wallach, *Cell* **85**, 803 (1996).
5. G. M. Cohen, *Biochem. J.* **326**, 1 (1997).
6. S. R. Ruuls *et al.*, *Immunity* **15**, 533 (2001).
7. M. Feldmann, R. N. Maini, *Annu. Rev. Immunol.* **19**, 163 (2001).
8. B. Beutler, *Immunity* **15**, 5 (2001).
9. M. Feldmann, *Nature Rev. Immunol.* **2**, 364 (2002).
10. R. Goldbach-Mansky, P. E. Lipsky, *Annu. Rev. Med.* **54**, 197 (2003).
11. R. J. Hayes *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15926 (2002).
12. A. V. Filikov *et al.*, *Protein Sci.* **11**, 1452 (2002).
13. P. Luo *et al.*, *Protein Sci.* **11**, 1218 (2002).
14. Materials and methods are available as supporting material on *Science* Online.
15. J. Yamagishi *et al.*, *Protein Eng.* **3**, 713 (1990).
16. X. M. Zhang, I. Weber, M. J. Chen, *J. Biol. Chem.* **267**, 24069 (1992).
17. G. Steiner *et al.*, *Rheumatology* **38**, 202 (1999).
18. T. Horiuchi *et al.*, *Endocr. J.* **46**, 643 (1999).
19. A. K. Ulfgren *et al.*, *Arthritis Rheum.* **43**, 2391 (2000).
20. P. Ameloot, W. Declercq, W. Fiers, P. Vandenabeele, P. Brouckaert, *J. Biol. Chem.* **276**, 27098 (2001).
21. I. Hishinuma *et al.*, *Hepatology* **12**, 1187 (1990).
22. S. Sakaguchi, S. Furusawa, K. Yokota, M. Takayanagi, Y. Takayanagi, *Int. J. Immunopharmacol.* **22**, 935 (2000).
23. P. Brouckaert, C. Libert, B. Everaerdt, W. Fiers, *Lymphokine Cytokine Res.* **11**, 193 (1992).
24. P. M. Steed *et al.*, data not shown.
25. A. M. Bendele *et al.*, *Arthritis Rheum.* **43**, 2648 (2000).
26. K. Sreekrishna *et al.*, *Biochemistry* **28**, 4117 (1989).
27. Y. P. Li *et al.*, *Biol. Pharm. Bull.* **24**, 666 (2001).
28. C. F. Ware, *J. Exp. Med.* **192**, F35-8 (2000).
29. R. M. Locksley, N. Killeen, M. J. Lenardo, *Cell* **104**, 487 (2001).

30. We thank M. Ary for technical assistance with the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/301/5641/1895/DC1
Materials and Methods
References
Figs. S1 to S7

9 December 2002; accepted 14 August 2003

The Dog Genome: Survey Sequencing and Comparative Analysis

Ewen F. Kirkness,[1] Vineet Bafna,[2]* Aaron L. Halpern,[2]* Samuel Levy,[2]* Karin Remington,[2]* Douglas B. Rusch,[2]* Arthur L. Delcher,[1] Mihai Pop,[1] Wei Wang,[1] Claire M. Fraser,[1] J. Craig Venter[2]

A survey of the dog genome sequence (6.22 million sequence reads; 1.5× coverage) demonstrates the power of sample sequencing for comparative analysis of mammalian genomes and the generation of species-specific resources. More than 650 million base pairs (>25%) of dog sequence align uniquely to the human genome, including fragments of putative orthologs for 18,473 of 24,567 annotated human genes. Mutation rates, conserved synteny, repeat content, and phylogeny can be compared among human, mouse, and dog. A variety of polymorphic elements are identified that will be valuable for mapping the genetic basis of diseases and traits in the dog.

Our understanding of how the human genome functions in health and disease will benefit from comparison of its structure with the genomes of other species (*1*, *2*). The domestic dog is a particularly good example, where an unusual population structure offers unique opportunities for understanding the genetic basis of morphology, behaviors, and disease susceptibility (*3*, *4*). The physical and behavioral characteristics of ~300 dog "breeds" are maintained by restricting gene flow between breeds. Many modern breeds are derived from few founders and have been inbred for desired characteristics. This has led to a species with enormous phenotypic diversity, but with significant homogenization of the gene pool within breeds. Many of the ~360 known genetic disorders in dogs resemble human conditions, and their causes may be more tractable in large dog pedigrees than in small, outbred human families (*4*, *5*). The combination of genetic homogeneity and phenotypic diversity also provides an opportunity to understand the genetic basis of many complex developmental processes in mammals (*6*).

Because of the costs of sequencing mammalian genomes to completion, these projects have been restricted to a few species that are considered to be of greatest value to biomedical research. The decision as to whether future projects should aim for complete sequence coverage of a few more genomes, or whether the existing "reference genomes" can be exploited to characterize a wider variety of genomes that are sequenced to a lower level of coverage, must be made. Here,

[1]The Institute for Genomic Research, Rockville, MD 20850, USA. [2]The Center for Advancement of Genomics, Rockville, MD 20850, USA.

*These authors contributed equally to this work.

we address this issue by exploring how much information can be extracted from 1.5× sequence coverage of the dog genome.

Assembly of 6.22 million sequence reads from the genomic DNA of a male standard poodle yielded 1.09 million contigs and 0.85 million singletons. The dog sequences described in this paper have been assigned GenBank accession numbers AACN010000001 to AACN011089636 and CE000001 to CE853796. Analysis of start position offsets for overlapping reads yielded estimates of the euchromatic genome size that ranged from 2.31 to 2.47 Gb [supporting online material (SOM) Text]. These values are similar to the estimated length of the mouse genome (2.5 Gb) (*7*) but smaller than for human (2.9 Gb) (*8*). The rank order of these values is consistent with a previous estimate of mammalian genome lengths (*9*). Assuming a haploid genome size of 2.4 Gb (i.e., 1.5× sequence coverage), the assembly output resembles a simple model for the assembly (40 base overlaps) of 6.22 million reads that lack repeats: 1.16 million contigs (mean length 1414 bases, 5.0 reads/contig) and 0.39 million singletons (*10*). An excess of residual singletons is the principal deviation from the model assembly and can be explained largely by their content of repetitive sequence.

The contigs and singletons were ordered and oriented with Bambus, a general purpose scaffolder (SOM Text). When single links were permitted, Bambus generated 522,101 scaffolds with a mean length of 3.8 kb and mean span of 8.6 kb. Because of a small but significant rate of mispaired reads in related studies (0.34%) (*11*) and uncertainty of contig ordering, the use of single-link scaffolds here was restricted to analyses where the results of their use could be validated by an independent measure (e.g., colocalization of adjacent markers on the dog and human genomes) (SOM Text).

To assess the randomness of the shotgun sequence data and the fidelity of assembly, we examined the coverage of dog genomic DNA that had been sequenced independently. The sequences of four overlapping bacterial artificial chromosomes (BACs) (GenBank accession numbers AC114891, AC114332, AC113570, and AC114890) provided a reference 512 kb region of the dog genome. Theoretically (*10*), random 1.5× coverage should provide sequence data for 78% of the region, with 297 gaps of mean length 385 bases. The actual coverage was 77%, with 265 gaps of mean length 452 bases and median length 256 bases. The 512 kb interval was covered by 314 sequences. Of these, 10 failed to align over their complete lengths. Five of the partial alignments were caused by differential insertions of short interspersed nuclear elements (SINEs). Four involved single reads that contain either insertions or deletions of more than 70 bases. These could be caused by polymorphisms, misassembly of the BAC sequences, or multiple copies of the aligned sequences within the dog genome. The remaining partial alignment involved a contig where the alignment terminated at a short overlap of the assembled reads. This contig (0.4% of all aligned assemblies) was most likely to be a partial misassembly.

Approximately 31% of the dog genomic sequence was identified as repetitive after comparison with RepeatMasker libraries of known vertebrate and carnivore-specific repeat elements (table S2). This value is smaller than the content of known repetitive elements in the human (46%) and mouse (38%) draft genomes (*7*). At least part of this difference is likely to be caused by lineage-specific repeats that have not yet been characterized in the dog. Indeed, a small sample of repeat-masked dog sequence was found to contain nine additional repetitive sequences that each cover >1 Mb of genomic sequence, representing an additional 4.4% of the dog genome (table S3). These repetitive elements were not detected in either the human or mouse genomes, and their low divergence (average 15%) indicates that they are lineage-specific.

Recently, a comparison of ancestral repeats in the human and mouse genomes indicated that, since divergence from a common ancestor, the two lineages have been subject to distinctive mutation rates (*7*). Here, we performed a similar analysis on all repeat elements that are common to human, mouse, and dog and that are represented in the dog by >0.75 Mb of sequence. This analysis confirmed a higher level of substitution in mouse than human (~1.6-fold) (table S4). In contrast, there was little difference between the substitution levels for dog and human.

The L1MA family is the youngest of the mammalian-wide long interspersed nuclear elements (LINE1) and gave rise to most of the species-specific LINEs that are currently active (*12*). In the mouse genome, RepeatMasker identified a substantial representation of the subfamilies L1MA10 (oldest) through L1MA5 (youngest). However, in the dog genome, relatively little DNA was classified as L1MA8, L1MA7, L1MA6, or L1MA5 (fig. S1). In contrast, a relatively large amount of the dog genome was identified as L1MA9; this may derive from early carnivore-specific LINEs that have not yet been classified but which arose from L1MA9, and with which they therefore share greatest sequence similarity. This supports the view the dog lineage was the first to diverge from the common ancestor of human, mouse, and dog (*13*, *14*).

The most abundant class of SINE, representing 7% of the dog genome, is thought to be derived from transfer RNA (tRNA)-Lys and has homologs throughout the carnivore lineage (*15*, *16*). In the dog, a subfamily of this element (defined as SINEC_Cf; RepBase release 7.11) could be distinguished from related SINEs by a two-base insertion (RG) at position 91. We estimate that this subfamily is represented by approximately 230,000 copies in the dog (RepeatMasker analysis). Despite such abundance, these elements display only 4.8% average divergence from their consensus sequence. This indicates a recent, large expansion of these elements in the canine lineage.

We aligned the 1.5× dog sequences to the draft human and mouse genomes [National Center for Biotechnology Information (NCBI) build 31 and build 3 respectively] using BLASTN, and we identified the best-scoring alignments to segments of dog sequence (SOM Text). For comparison, best alignments of mouse segments to human were determined similarly. Here, "alignment" refers to such best alignments unless otherwise noted. Almost twice as much unique human sequence could be aligned with the 1.5× collection of dog sequences than with the more complete 8× collection of mouse sequences (table S5).

The best alignments of dog and mouse genomic fragments with the human genome were compared for their content of genes and transcripts, as defined by the Ensembl annotation of the human genome (version 11.31.1). Alignments of mouse contigs to the human genome covered 80% (29,529) of all human transcripts and 75% (18,311) of all genes; 77% of the protein coding sequence from hit transcripts is covered. The fraction of genes covered is consistent with independent estimates for the number of genes that have 1:1 orthology between mouse and human (70 to 80%) (*7*, *17*). Despite much lower sequence coverage, the dog alignments covered a similar number of human transcripts (29,673) and genes (18,473), although only 61% of coding sequence from hit transcripts was covered.

Of the 29,673 human transcripts that aligned with dog sequences, 83% were aligned for more than 50% of their lengths, compared with 93% against the more complete mouse genome. The main difference between the two data sets is the distribution of values for fractional coverage. As expected, the draft mouse genome provides full-length coverage for a large proportion of the transcripts. In contrast, there is a broad distribution of coverage values, peaking at 70 to 80%, for transcripts that aligned with the dog sequences (Fig. 1).

There were 7292 human transcripts (6326 genes) that were not represented in the alignments of dog sequences. Most of these transcripts are relatively short (less than 600 bases) (Fig. 1), and the absence of orthologous dog sequences could be explained by gaps in the sequence coverage of the dog genome. However, a similar number of human transcripts (7436; 6461 genes) were not included in the

alignments with the mouse genome. Indeed, most of these transcripts (82%) failed to yield best alignments with either mouse or dog sequences. Further analysis of the 7292 "missing" transcripts indicated that 2136 did not display any significant similarity to dog sequence based on BLASTN and TBLASTN (peptide alignments). In addition to incomplete coverage, there are several possible explanations for the failure to detect homologous sequences in the dog genome. First, some of the human transcripts may be annotated incorrectly and may not actually represent functional genes. Second, homologous genes may have been lost from the dog genome since the divergence of dog and human lineages. Third, some genes may be evolving too rapidly to permit identification by sequence comparisons.

For the remaining 5156 transcripts, there were homologous dog sequences. However, these dog sequences had higher scoring alignments with other human transcripts. In most cases [4347 (84%)], the best alignments were to related human genes (i.e., members of the same Ensembl gene family). In addition to the previous explanations, a failure to detect orthologous dog sequences for these 4347 human transcripts includes the possibility of genes that have duplicated in the human lineage since divergence from the dog lineage. Such differential gene expansion, as has been observed between human and mouse (*7*), could account for most of the human transcripts that failed to provide best alignments in this study.

If we consider only those 29,529 human transcripts for which we found alignments to the 8× mouse genome assembly, 96% also align with the 1.5× dog genome assembly. Of the 29,673 human transcripts that aligned with dog sequences, there were 1319 (4%) that did not align with the mouse genome. It is likely that these consist largely of common ancestral genes that have been conserved in human and dog but have been lost or have mutated substantially in mouse (*18*). In addition to protein-coding sequence, the alignments to known genes included substantial coverage (18 to 42%) of noncoding elements (Table 1). Our data have already been very useful for designing probes and markers to rapidly characterize regions of the dog genome that have been associated with genetic disorders (*19*).

Comparative genomics and survey sequencing can be used to identify gene families that are relatively expanded or contracted. Where multiple dog sequences best align to a single coding segment of a human gene, we call this a "pile-up." When no dog sequences align to a coding segment of a human gene, we call this a "gap." In table S6, 1355 and 513 human genes are listed that have been identified as sites of pile-ups and gaps, respectively ($P < 0.01$) (SOM Text). Like mouse (*20*), dog appears to have a much larger complement of olfactory receptor genes than human, and several large pile-ups were observed for different subfamilies (table S7). However, the large repertoire of ~140 vomeronasal receptors in mouse (*21*) is not reflected in the dog 1.5× sequence, and only four members of this gene family were identified. As expected, families of cytochrome P-450 genes are represented by both gaps and pile-ups, indicating that dog, like mouse, has a unique repertoire of genes for oxidative metabolism of endogenous compounds and xenobiotics. There are also examples of multiple pile-ups, but no gaps, in different gene families that perform related metabolic functions, indicating areas of metabolism for which dog has an expanded repertoire of genes relative to human (e.g., peptide metabolism by dipeptidyl peptidases) (table S7). In addition to loss or duplication of single genes, 10 large (>500 kb) contiguous regions of the human genome that each contain at least 10 genes were devoid of best-hits for any dog sequences (table S8). Some of these involve clustered gene families (e.g., pregnancy-specific β-1 glycoproteins, defensins)

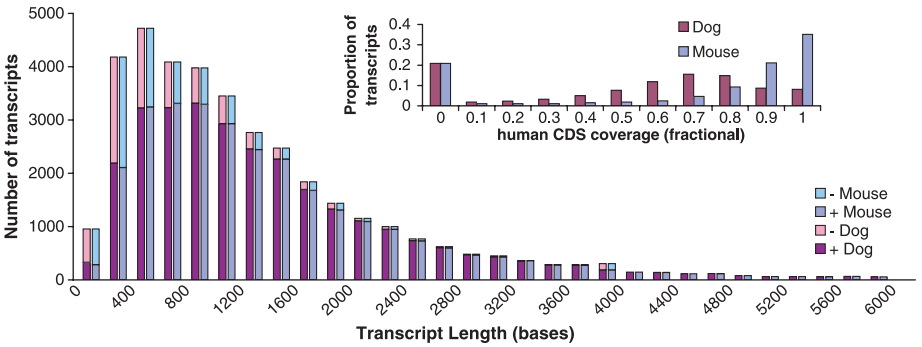

Fig. 1. Comparison of human transcripts that are represented by best alignments of dog or mouse genomic fragments with the human genome. Each bar indicates the number of human transcripts that have either no coverage (–) or partial or complete coverage (+) by best alignments with the dog or mouse sequences. For example, of 4150 total transcripts between 200 and 400 bases long, 48% have alignments with dog sequences; of transcripts 3000 to 3200 bases, 95% have such alignments. Note that many human transcripts (3800 to 4000 bases), did not align with either dog or mouse sequences. Most of these transcripts (>90%) are annotated as containing reverse transcriptase domains that were masked, along with other repetitive elements, prior to alignment of dog and mouse sequences with the human genome. (**Inset**) Fractional coverage of human transcripts by dog and mouse genomic fragments. For each human transcript, the fractional coverage of the protein-coding portion of the transcript (CDS) by best alignments with the dog or mouse sequences was calculated. The figure illustrates the proportion of all transcripts that fall into each bin of fractional coverage.

Table 1. Coverage of human transcripts and intergenic regions by alignments with dog and mouse genomic sequence. Nonredundant coverage of human genomic sequence by the best dog and mouse alignments and three-way alignments (COBs) were classified using Ensembl release 11.31.1. Values represent the total length of coverage for each class of sequence and the percentage of each class covered.

Sequence class	Dog best hits		Mouse best hits		COBs	
	Mb	%	Mb	%	Mb	%
5′-UTR	3.10	41.9	3.82	51.6	1.97	26.6
3′-UTR	10.60	52.2	10.51	50.9	5.66	27.5
CDS	20.58	60.7	26.00	76.6	17.08	50.4
Intron	192.93	26.0	109.91	14.8	46.53	6.8
Upstream (5 kb)	30.29	22.2	22.43	16.4	9.33	6.8
Downstream (5 kb)	35.04	25.9	26.27	19.4	11.75	8.7
Intergenic	360.47	18.3	179.13	9.1	77.05	3.9

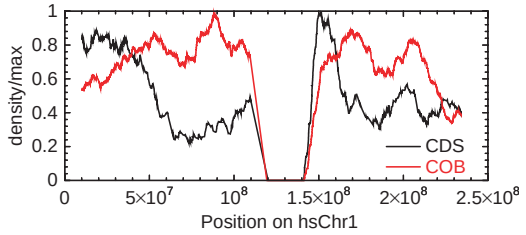

Fig. 2. The densities of COBs (red) and coding sequence (black) along human chromosome 1. They were computed as number of bases in sliding windows of 20 Mb and were normalized to the single largest value for each category.

that appear to have undergone significant expansion in human relative to dog.

On the basis of the collections of pairwise alignments defined above, we constructed a more restrictive set of three-way alignments that we term COBs (clusters of orthologous bases) by analogy to COGs (clusters of orthologous groups) (*22*). Each COB consists of sequence from dog, human and mouse, in which all pairwise alignments are mutually best matches [i.e., each pairwise alignment is a "syntenic anchor" in the sense of (*17*) (SOM Text)]. Like the analyses of pairwise BLASTN alignments, the COBs indicate that dog and human genome sequences are more similar to each other than either is to mouse. Also, mouse sequences are more similar to human than to dog (SOM Text), as previously indicated by an analysis of a 200 kb genomic region (*23*). COBs are enriched within the coding sequence of genes (Table 1). However, as reported previously in connection with human and mouse syntenic anchors (*17*), there are many COBs in intergenic regions, and the distribution of COBs along the genome is distinct from that of genes (Fig. 2). The enrichment in untranslated regions (UTR) relative to intronic, upstream and downstream regions suggests widespread conservation of potential regulatory signals in these regions.

Analysis of synonymous and nonsynonymous substitutions confirmed that mouse is considerably more divergent from both dog and human than the latter two are from each other (table S10; SOM Text). However, because of evidence of differing mutation rates in the mouse and human lineages (*7*), current levels of sequence similarity are insufficient to define the times since divergence of different lineages. We considered the possibility that consensus repeat patterns could serve to root a tree that is based on alignments of orthologous repeats in the three species (SOM Text). We chose to use orthologous repeats rather than orthologous genic sequence (coding or regulatory) because it is plausible that repeats evolve in a neutral (unselected) fashion. Additionally, the consensus sequence for a repeat family is a more accurate reconstruction of the ancestral sequence than is the consensus sequence for the alignment of a specific gene, because the repeat family consensus can be based on many instances from each species. The resulting tree (Fig. 3) has dog as the outgroup. Although there is only a short interior branch separating dog from mouse and human, bootstrap analysis on this data set gave 100/100 replicates, indicating that the tree is robust. Thus, this analysis is consistent with the content of LINE1 elements (fig. S1), implying that mouse and human share a common ancestor that is distinct from the dog lineage.

Since divergence, the common ancestral genome of human, mouse, and dog has undergone distinctive rearrangements in each lineage. However, there has been conservation of local gene order (synteny) within relatively large segments of each genome. Comparison of the human and dog genomes by reciprocal Zoo-FISH (fluorescent in situ hybridization) (*24*) has revealed approximately 70 conserved blocks on the dog genome. Recently, 85 orthologous regions were identified by comparison of 830 radiation hybrid (RH)–mapped markers from dog with the human genome sequence (*25*). However, the number of distinct segments within most syntenic blocks remains to be established. Although ~3200 RH markers have been characterized (*25*), most of these are microsatellites and their minimal content of unique dog sequence is insufficient to directly identify orthologous loci on other genomes. Therefore, we aligned these markers to contigs and scaffolds of the dog assembly, which allowed most to be mapped to the human and mouse genomes (SOM Text). Analysis of 2704 extended markers resulted in the clustering of best-hits for 2177 and 1766 on the human and mouse genomes, respectively. These clusters were then ordered on dog chromosomes (Fig. 4) (tables S11 and S12; SOM Text). The clusters confirmed 78 of the 85 conserved regions reported previously (*25*), and revealed an additional syntenic block (CFA2/HSA17). Of the seven that were not confirmed, six are currently supported by a single marker and, if real, are likely to represent only short regions of conserved synteny. More significantly, clustering of the markers resolved distinct

Table 2. Pros (possibilities) and cons (limitations) for survey sequencing of a mammalian genome.

Possibilities	Limitations
Rapid and economical path to genomic information	
• A variety of related genomes can be surveyed simultaneously, for same cost as the 'reference sequence' of any one genome	• Complete understanding will ultimately require reference-quality sequence from a broad spectrum of species
Genome characterization	
• Identification of exon sequence for most genes - Develop probes for isolating cDNAs and building comparative genome maps - Provide data for building expression arrays - Resolve functionally important elements common to multiple genomes • Extensive sampling of ancient repetitive elements - Estimation of neutral mutation rates - Construction of evolutionary trees • Broadly conserved non-coding sequence suggests regulatory signals, structural elements or non-protein coding genes	• Requirement for annotated reference genome sequence from related species • Missing genes due to incomplete coverage • Incomplete gene sequences and lack of context • Less accurate assignments of orthology, identification of pseudogenes, and analyses of syn/non-syn polymorphisms • Incomplete coverage limits ability to conclude a segment is not conserved
• Preliminary identification of gene family expansions/contractions	• Less definitive than alignment of completed genomes
• Identification of potentially polymorphic sequences (STRs, SNPs, SINEs) that are contiguous with genes/conserved segments important for detailed linkage and association studies	• Dependence on conservation of synteny with reference genome for application to whole genome linkage to detailed mapping of candidate regions
• Construction of a BAC-based physical map	• Requirement for long segments of conserved and ordered synteny, and a low frequency of micro-rearrangements
Preservation of investment	
• Survey sequence can be re-assembled with supplementary sequence data in future efforts to complete genome sequence	• Assumption that 0.5 to 1.0 kb reads will remain the fundamental unit of genome sequence assemblies

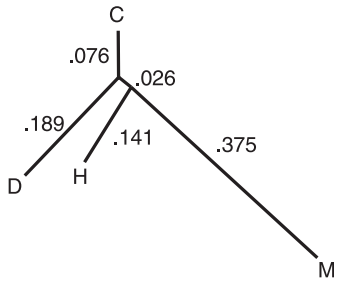

Fig. 3. Reconstruction of the dog-human-mouse divergence based on comparison to consensus ancestral repeats. A composite data set was constructed by concatenating alignments of dog, human, mouse, and consensus repeat sequences for repetitive elements present in all three species in positions, consistent with the three-way mapping induced by COBs. The resulting data set was used to determine a maximum-likelihood tree using PAML (baseml with the REV model). D, dog; H, human; M, mouse; C, repeat consensus. Values next to branches give the estimated branch lengths in units of expected numbers of substitutions per site. The tree is drawn to indicate the presumptive correspondence of the repeat consensus to a common ancestor of the sequences observed in dog, human, and mouse.

segments of conserved ordered synteny within the larger blocks. For example, CFA18 has orthology with only human chromosomes 7 and 11. However, rearrangements appear to have broken these two blocks into at least nine distinct segments of conserved ordered synteny. To declare such a segment, we required at least three consecutive markers in which the RH mapping and comparative mapping were consistent (SOM Text). In some cases, this filtering eliminates segments implied by previous studies (*24*, *25*); such segments will be a mixture of errors in the RH mapping and local rearrangements that are too small to identify confidently, given the current marker density and map precision. The resulting 159 segments of conserved synteny collectively span 2.2 Gb of the human genome. The mean length of these segments on the human genome was 13.2 Mb, and ranged from <1 Mb to 84 Mb. Despite the high number of chromosomes in the dog, the true number of syntenic breakpoints (excluding microrearrangements) is unlikely to be more than a few hundred. For the dog-human comparison, interchromosomal rearrangements are fewer, and represent a smaller proportion of the total rearrangements (78 of 159 segments) than for the dog-mouse comparison (130 of 205).

Chromosomal rearrangements that are unique to the dog lineage are indicated by coincident breakpoints in dog-human and dog-mouse synteny (e.g., CFA5). Similarly, lineage-specific rearrangements in the mouse genome are indicated when a single segment of conserved dog-human synteny corresponds to multiple segments of conserved dog-mouse synteny (e.g., CFA12). This comparative analysis of the three genomes identified 69 breakpoints that are consistent with rearrangements in human, 60 in dog, and 115 in mouse. These data support some predicted models of ancestral mammalian genomes and provide a detailed view of the locations of fission and fusion events that have contributed to the high number of dog chromosomes relative to human (SOM Text).

Contigs of the dog 1.5× assembly were examined for single nucleotide polymorphisms (SNPs) using the Bayesian SNP-detecting basecaller within the Celera Assembler (SOM Text). This basecaller evaluates the quality value data underlying the consensus sequence at each column of the contig multiple sequence alignment, determines the most likely dual haplotype call, and provides a quality value for this call. Setting a confidence threshold of 0.90 to obtain only high-quality calls, 974,400 putative SNPs were identified in the 1.5 Gb of assembled sequence (roughly 1/1500 bases), and a further 149,818 high-quality di-, tri-, and tetranucleotide polymorphisms were predicted. The putative SNPs have been assigned GenBank accession numbers ss8830321 to ss9805720. The sequence data was derived from a standard poodle with a Wright's inbreeding coefficient of 19.1% (5 generations). The density of predicted SNPs falls within the range of values that have been estimated to occur in other individual purebred dogs (1/800 to 1/5400

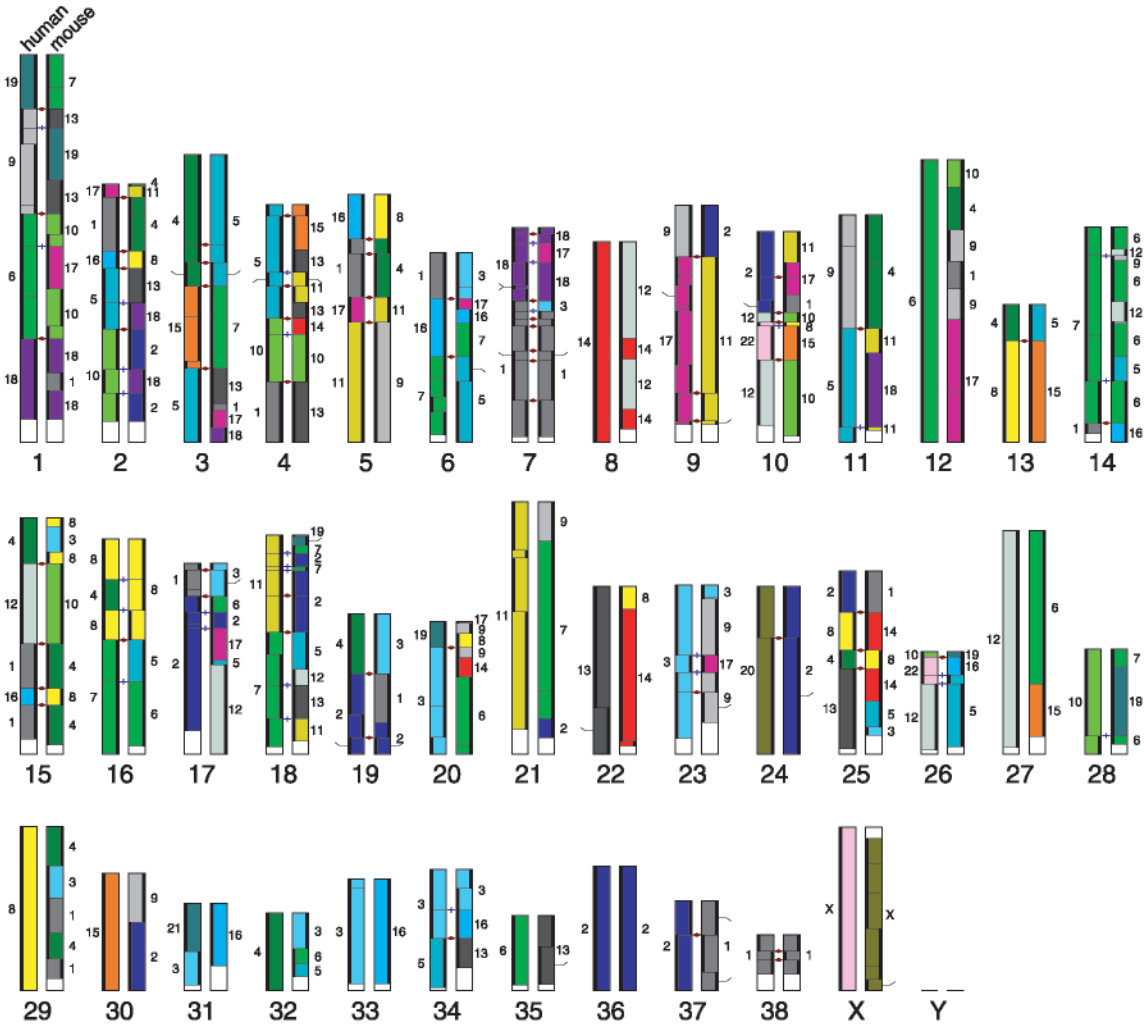

Fig. 4. Comparative map of the 40 dog chromosomes overlaid with mouse and human genomes. Each dog chromosome is represented twice, and overlaid with either human (**left**) or mouse (**right**) genomic segments. Map positions in dog (*25*) increase from bottom to top along each dog chromosome. Distinct segments of conserved synteny between mouse and human are depicted by variously colored and numbered blocks corresponding to the 22 autosomes and X in human and to the 19 autosomes and X in mouse. No systemic blocks were found for chromosome Y. The relative orientation of each block is indicated by a thick vertical line on either the left (human or mouse map coordinates increasing) or right (coordinates decreasing) of each block. Segments for which a simple inversion would remove either the upper (__) or lower (/‾‾) breakpoints are indicated. Segmental breakpoints that are coincident on the human and mouse genomes indicate rearrangements in the dog lineage (sideways ◊). Several coincident breakpoints can also be explained by independent rearrangements in each of the human and mouse genomes (+).

bases) (*26*). It is also similar to the estimated density of SNPs in human individuals (1/1000 to 1/2000 bases) (*27*). Of the putative SNPs, 295,178 represented base pair deletions, whereas 680,222 were base pair substitutions (tables S13 and S14). Of all the high quality SNPs, 268,482 (27%) are contained in contigs that were mapped to human chromosomes by BLASTN alignments. Of these, more than one-third mapped near or within the coding sequence of 14,679 distinct human genes (table S15).

Comparisons of the 1.5× data with dog BAC sequences in GenBank and comparisons of sequences from overlapping BAC clones revealed numerous examples of sequences that differ only by the presence or absence of a SINE insertion (tables S16 and S17). In almost all cases, the SINE most closely resembled the SINEC_Cf repeat (RepBase release 7.11). This element has undergone a relatively recent large expansion in the canine lineage. For most examples, polymerase chain reaction (PCR) amplification across the implied region of SINE insertion in different dogs verified the polymorphism (tables S16 and S17). To estimate the abundance of bimorphic SINEs in the sequenced poodle, a sample of 20,048 SINEC_Cf elements, each flanked by at least 60 bases of nonrepetitive sequence, was searched against the complete 1.5× data set. (The SINEC_Cf elements had an average length of 189 bases and average divergence of 3.4%.) For 709 (3.5%) of these, there were unique database matches in which the sequence of the SINE flanks are contiguous and the SINE is absent (SOM Text). When the same analysis was performed on an older family of SINEs (SINEC_Cf2, average length 182 bases, average divergence 7.8%), only 0.2% of the sample yielded such matches. The 1.5× data set is predicted to cover ~50% of a 4.8 Gb diploid genome. Consequently, these data indicate that approximately 7% of SINEC_Cf elements are bimorphic in the sequenced poodle (i.e. ~16,000 of 230,000 copies in the entire genome, as estimated above). Undoubtedly, there are many more bimorphic loci in the general dog population, and this genetic diversity is likely to be a valuable resource for identifying the ancestral relationships between different dog breeds and between dogs and related canids. For comparison, the number of bimorphic SINEs (Alus) in the human population is estimated to be only ~1200 (*28*). The estimated abundance of bimorphic SINEs was validated by amplifying all SINEC_Cf elements that could be identified in a contiguous 425 kb region of the dog genome. Of the 24 SINEs examined, six were bimorphic in a small sample of dogs, and two additional sites displayed variation in related canids (fig. S2; table S18). In the genome of the sequenced poodle, SINEs were absent from one or both alleles at 4 of the 24 loci. Five of the bimorphic loci are within the PFTAIRE-1 gene. The insertion of SINEs within genes can cause dramatic phenotypic effects (e.g., canine narcolepsy) (*29*), and many such insertions are likely to have at least subtle effects on gene expression patterns. Owing to the abundance of bimorphic SINEs in the dog, it is tempting to speculate that these elements contribute to the unusual phenotypic diversity of modern dog breeds.

Our work with 1.5× sequence coverage of the dog genome has highlighted some of the insights, potential applications, and limitations that derive from survey sequencing (Table 2) and is relevant for future decisions on how best to characterize large eukaryotic genomes. This depth of coverage led to only limited assembly (small contigs and short scaffolds) and thus is of limited value on its own. However, when used in conjunction with at least one related reference genome, it proved an economical way to obtain a large amount of functional annotation. The survey sequence permitted reliable estimates for several global parameters of the dog genome, such as its neutral mutation rate and repeat content. The coverage also includes partial sequence data for dog orthologs of most annotated human genes. An obvious limitation, relative to a high-quality draft, is that few dog genes are sequenced completely, most consist of multiple fragments, and a small fraction is likely to have been missed entirely. However, the gene fragments provide a valuable resource for rapidly developing short tandem repeat polymorphism (STR)- or SNP-based assays for resolution of linkage between a candidate gene and a specific phenotype, or reagents for mRNA expression studies. Because most genes are represented by multiple sequence fragments, associated cDNAs may be isolated or further sequencing performed to encompass all exons. This will be a relatively straightforward process if the survey sequencing project includes end-sequencing of a large-insert library. Indeed, our preliminary studies suggest that an extensive physical map of ordered dog BAC clones can be assembled on a platform of the human genome, using only BAC-end sequences and scaffolds of the 1.5× sequence assembly (SOM Text). The combined resources of survey sequence and physical clone coverage will permit the dog genome to be navigated with ease and allow any selected genomic regions to be rapidly characterized by more extensive sequencing.

Despite the fragmentary nature of the data, we were able to demonstrate that >4% of intergenic sequence is conserved among dog, human, and mouse. This extends recent findings that human and mouse have considerable conservation in noncoding regions (*7*, *17*). Each new genome that is sequenced (or randomly sampled) will further resolve those regions that are of critical functional importance.

There are two major motives for sequencing additional mammalian genomes. First, the sequence provides an infrastructure for genetic mapping studies that can identify genes that are responsible for specific traits and diseases in an organism. Second, comparative genome analysis can identify conserved genomic elements of general functional importance that are often otherwise overlooked. We have shown that 1.5× coverage of a genome provides a valuable, cost-effective resource for both organism-specific biology and comparative genomics.

References

1. S. J. O'Brien *et al.*, *Science* **286**, 458 (1999).
2. S. J. O'Brien *et al.*, *Science* **286**, 479 (1999).
3. E. A. Ostrander, F. Galibert, D. F. Patterson, *Trends Genet.* **16**, 117 (2000).
4. D. F. Patterson, *J. Vet. Intern. Med.* **14**, 1 (2000).
5. D. F. Patterson, *Canine Genetic Disease Information Systems: A Computerized Knowledge Base of Genetics Diseases in Dogs* (Mosby Yearbook, St. Louis, MO, 2001).
6. K. Chase *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9930 (2002).
7. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
8. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
9. A. E. Vinogradov, *Cytometry* **31**, 100 (1998).
10. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
11. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
12. A. F. Smit, G. Toth, A. D. Riggs, J. Jurka, *J. Mol. Biol.* **246**, 401 (1995).
13. O. Madsen *et al.*, *Nature* **409**, 610 (2001).
14. W. J. Murphy *et al.*, *Nature* **409**, 614 (2001).
15. M. F. Minnick, L. C. Stillwell, J. M. Heineman, G. L. Stiegler, *Gene* **110**, 235 (1992).
16. N. S. Vassetzky, D. A. Kramerov, *Mamm. Genome* **13**, 50 (2002).
17. R. J. Mural *et al.*, *Science* **296**, 1661 (2002).
18. P. A. Davies, W. Wang, T. G. Hales, E. F. Kirkness, *J. Biol. Chem.* **278**, 712 (2003).
19. J. K. Lowe *et al.*, *Genomics* **82**, 86 (2003).
20. X. Zhang, S. Firestein, *Nature Neurosci.* **5**, 124 (2002).
21. I. Rodriguez *et al.*, *Nature Neurosci.* **5**, 134 (2002).
22. R. L. Tatusov *et al.*, *Nucleic Acids Res.* **29**, 22 (2001).
23. I. Dubchak *et al.*, *Genome Res.* **10**, 1304 (2000).
24. M. Breen *et al.*, *Genome Res.* **11**, 1784 (2001).
25. R. Guyon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5296 (2003).
26. J. A. Brouillette, J. R. Andrew, P. J. Venta, *Mamm. Genome* **11**, 1079 (2000).
27. R. Sachidanandam *et al.*, *Nature* **409**, 928 (2001).
28. M. A. Batzer, P. L. Deininger, *Nature Rev. Genet.* **3**, 370 (2002).
29. L. Lin *et al.*, *Cell* **98**, 365 (1999).

30. We thank E. Ostrander, E. Mignot, R. Wayne, and J. Pollinger for providing samples of DNA from dog and other canids. We are grateful to E. Ostrander and F. Galibert for providing the sequences of RH-mapped markers. We thank C. Fosler and H. Koo for help with sequence submissions to GenBank. This work was supported by the J. Craig Venter Science Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/301/5641/1898/DC1
SOM Text
Figs. S1 and S2
Tables S1 to S18

5 May 2003; accepted 14 August 2003

Science
AAAS

The Dog Genome: Survey Sequencing and Comparative Analysis

Ewen F. Kirkness *et al.*

Science **301**, 1898 (2003);
DOI: 10.1126/science.1086432