Improvement of the Dog Genome to Near Finished Quality

Kerstin Lindblad-Toh and Eric Lander
Broad Institute of MIT and Harvard

Summary
The dog (*Canis familiaris*) was chosen for sequencing both for its unusual potential for disease gene mapping and for its clear value for mammalian comparative sequence analysis. The current 7.5x assembly generated from a female boxer is of high quality and has already proven useful for identification of regulatory elements [1-3]. The SNP map accompanying the genome sequence has also been utilized to survey the haplotype structure and demonstrate the feasibility of disease gene mapping within dog breeds[2].

Despite the overall high quality of the dog genome, ~1% of the euchromatic genome resides in gaps and ~0.5% of the assembly shows evidence of potential assembly errors. In addition, ~20% of genes show some evidence of sequencing error such as a frameshift or an unusually short intron. To fully utilize the favorable evolutionary position of dog, we would propose to improve the genome sequence further by performing targeted finishing of problematic exons and genomic regions as well as to fill the majority of gaps, bringing the genome to near finished quality. Such a project would cost ~$3 million and could be performed in 12-18 months.

Scientific justification for improving the dog genome

The scientific justification for sequencing the dog genome is described in detail in the original Whitepaper. Here we emphasize the particular advantages that we see for further improving the quality of the dog genome.

The dog is an important reference genome for mammalian analysis based on its position in the mammalian tree[4,5]. The dog is the first sequenced mammal belonging to a major new clade, Laurasiatheria as compared to the previously sequenced mammals (human[6-7], mouse[8] and rat[9]). Adding a representative from this clade reduces the evolutionary distance from the common ancestor to the clades of Afrotheria and Xenarthra. Therefore, the ability to align reads from certain low coverage sequencing projects (such as elephant, tenrec and armadillo) increases when both human and dog are used as reference genomes for alignment. In fact, ~10% of reads from a 2x sequencing project cannot be assembled with the low-coverage data alone, but can be aligned using both dog and human as reference genomes. Furthermore, in our dog genome analysis[2] we have noted that the dog has few segmental duplications, less lineage specific repeat sequence, relatively few genes and less gene family expansions than human[6-7], mouse[8] and rat[9]. Thus, the dog appears to have retained many of the features of the "ancestral" eutherian genome, making it a good anchor for mammalian genomics. The structural accuracy of such a reference genome is crucial, because it influences both the assemblies and conclusions drawn from other genomes. For all of these regions, it is important to improve regions marked as questionable in the current assembly.

Dog is also a stellar organism for disease gene mapping[10-12]. The dog community is very supportive of an effort to improve the dog genome. They cite both the importance of dog as a model for human common diseases such as cancer and its frequent use by the pharmaceutical industry and in clinical trials. Many resources now exist to leverage this research. In particular, the boxer genome is complemented by 1.5x sequence from a Standard poodle[13] and a 2.1 million SNP map[2], as well as state of the art genomic tools such as expression arrays[14] and CGH arrays[15]. Of particular importance for disease gene mapping is the fact that LD is long within breeds (extending over megabases), but short across breeds[2,16]. The breed structure in dogs thus permits disease gene mapping using only ~10,000 SNPs for genome wide association mapping. For the dog genome community a complete gene set and a full comparison to the human genome would be very valuable. We therefore propose to specifically target faulty or missing genes for improvement.

Finally, relatively little improvement is necessary to substantially improve the genome assembly. As described below the dog genome is already of high quality and has been error-checked against the 10,000 marker RH map[17] and FISH data[18-20]. Our extensive evaluation of the dog genome as well as newly developed tools and processes for identification and finishing of targeted regions, will enable us to generate a near finished quality genome in a cost effective manner.

Current status of the dog genome

The current genome assembly**,** CanFam2.0, covers ~99% of the euchromatic portion of the dog genome as compared against a small amount of finished sequence (Table 1)[2]. An initial assembly (CanFam1.0), based on 7.5 fold sequence coverage of the genome from a female boxer was released in July 2004. The new assembly is an incremental update, based on extensive analysis and some correction to generate an assembly of overall higher contiguity and coverage, with a trustworthy contig and supercontig structure. As part of this effort, internal inconsistency checks and comparison to the 10,000 RH map was performed. Additional FISH data was generated to resolve the last few inconsistencies.

The assembly is of high quality on a fine-scale level with roughly 98% of bases in the assembly of Q40 or above.

Table 1: Assembly statistics (CanFam2.0) using ARACHNE[21]

Estimated genome size: anchored bases, spanned gaps (18Mb) and centromeric sequence (3Mb each)	2.445 Gb
N50 contig size	180 kb
N50 supercontig size	45.0 Mb
Assembly size, total bases	2.385 Gb
Number of anchored supercontigs	87
Portion of genome in anchored supercontigs	97%
Portion of assembly in gaps	1.0%
Portion of assembly with a quality score >Q40	98%
Portion of assembly in 'certified regions', without assembly inconsistency	99.3%

While 'quality scores' have been developed to indicate the nucleotide accuracy of a draft genome sequence, no analogous measures have been developed to reflect the long-range assembly accuracy. We therefore sought to develop such a measure based on the absence of two types of internal inconsistencies[2]. 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: (i) one end cannot be mapped to the region; (ii) the linkage relationships are inconsistent with the sequence within contigs; or (iii) 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 "questionable" otherwise. Roughly 99.5% 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 (Table 2).

Roughly 27,000 spanned gaps exist accounting for ~1% of the span of the assembly, the majority of these are small. In addition, the amount of missing sequence within 48 uncaptured gaps and 78 chromosomal ends is unknown.

Table 2. Size distribution of gaps and uncertified regions

	Gaps*		Uncertified regions	
Size distribution (kb)	**Number**	**(Mb)**	**Number**	**(Mb)**
All	26,726	26.4	919	10.6
<1	21,598	6.8	286	0.2
1-10	4,895	10.2	348	1.7
10-40	161	3.2	235	4.7
40-200	72	6.2	49	3.7
>200	0	0	1	0.3

*Gaps between supercontigs not included

Based on the current assembly we estimate the canine genome to contain ~19,500 genes. The gene set was identified using the Ensembl and Broad automated annotation pipelines followed by careful examination using conserved synteny analysis and manual curation of all genes that did not have a clear 1-1 orthologous relationship between human and dog[22]. Although the majority of genes are accurately represented in the dog genome, ~4,000 genes show evidence of some sort of sequencing error based on the presence of a frameshift or short intron. An example is the AP3S1 gene where an apparent frameshift (TT in human vs TTT in dog) is seen in the last coding exon. Additionally, ~ 1,000 genes are missing at least one exon. For example, in refseq NP_006124, exon 11 is present in mouse and human, but absent from dog. In this case the dog genome has a gap where the exon should reside. To create a well annotated mammalian gene set resolving these issues would be of high priority. Such corrections would also improve the ability to perform gene prediction across mammals.

Plans for improvement

To facilitate the use of the canine genome as a third mammalian reference genome, our goal is to produce a near finished genome with high structural integrity and complete gene content. To accomplish this we will perform finishing of ~80Mb of sequence, targeted to cover large gaps (both spanned and uncaptured), the regions marked as questionable in the assembly and the ENCODE regions. In addition, primer walks will be performed to validate genes with apparent errors, fill in missing exons and close the majority of small gaps. We would expect this to result in a genome with the following characteristics:

1. No global misassembly (including > 99.9% of assembly certified)
2. >99.5% coverage of the euchromatic portion of the genome
3. High base-quality: >99.5% of bases >Q40
4. All genes with 1-1 orthologs with human fully represented and accurate and most gene family members resolved
5. <10,000 gaps in genome (<4 gaps/Mb)
6. Finished quality ENCODE regions

The process for genome improvement would consist of computational and laboratory approaches to ensure the quality of the genome and to fill gaps or finish sequence. We anticipate the following main components:

1. Targeted finishing of ~80Mb of genome. We anticipate shotgun sequencing 400-600 clones to 10x coverage followed by traditional finishing. A selection of clones that correspond to ~80Mb territory will be identified, with the selection of a BAC or Fosmid determined by the size of the feature that needs to be covered. We will prioritize covering territory with the following characteristics:
    a. ENCODE regions
    b. Regions where the assembly integrity is questionable
    c. Large internal gaps (>10kb)
    d. Supercontig gaps and chromosome ends
    e. Clusters of mid size gaps (1-10 kb)

2. Gene validation. Genes with potential sequencing errors and gaps have been identified through our analysis of the dog gene content.
    a. For the ~4,000 genes with a potential error, the suspect exon with will be edited by a finisher either using only the original whole genome shotgun data or following the addition of paired primer walks. A pilot is ongoing to determine the need for primer walks.
    b. The ~1,000 exons with gaps in the dog sequence will be primer walked to fill the gaps.

3. Filling small gaps. Primer walks will be attempted on the ~27,000 gaps smaller than1kb in size. This will increase the overall contiguity of the assembly.

4. Filling mid sized gaps. For unclustered mid sized gaps (1-10 kb) we aim to use a combination of primer walking and novel computational approaches. Primer walks will be performed at the edge of the gap and used as bait to pull in existing unplaced reads or small contigs to extend into gaps.

5. Updating the genome assembly. Updated genome assemblies will be released periodically after integration of new data into the previous version of the assembly.

Cost and time frame

We estimate the cost of the project to be ~ $3 million, depending on the price of shotgun reads and finishing when the work is performed. The expected time frame for the project would be 12-18 months.

References:

1. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
2. Lindblad-Toh, K. Wade, C., Mikkelsen, T.S. Karlsson, E. et al Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog (submitted)
3. 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, 2004. **14**(5): p. 852-9.
4. Margulies EH, Blanchette M, Haussler D, Green ED; NISC Comparative Sequencing Program. Related Articles, Links Free Full Text Identification and characterization of multi-species conserved sequences.Genome Res. 2003 Dec;13(12):2507-18.
5. Murphy, W.J., et al., *Molecular phylogenetics and the origins of placental mammals.* Nature, 2001. **409**(6820): p. 614-8.
6. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
7. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
8. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
9. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
10. Ostrander, E.A., F. Galibert, and D.F. Patterson, *Canine genetics comes of age.* Trends in Genetics, 2000. **16**: p. 117-123.
11. Patterson, D., *Companion animal medicine in the age of medical genetics.* J Vet Internal Med, 2000. **14**(1): p. 1-9.
12. Chase, K., et al., *Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9930-5. Epub 2002 Jul 11.
13. Kirkness, E.F., et al., *The dog genome: survey sequencing and comparative analysis.* Science, 2003. **301**(5641): p. 1898-903.
14. Khanna, C, Wan X, Bose S, Cassaday R, Olomu O, Mendoza A, Yeung C, Gorlick R, Hewitt SM, Helman LJ. Related Articles, Links Abstract The membrane-cytoskeleton linker ezrin is necessary for osteosarcoma metastasis. Nat Med. 2004 10:182-6.
15. Thomas R, Fiegler H, Ostrander EA, Galibert F, Carter NP, Breen M. A canine cancer-gene microarray for CGH analysis of tumors. Cytogenet Genome Res. 2003 **102:**254-60.
16. Sutter, N.B., et al., *Extensive and breed specific linkage disequilibrium in Canis familiaris.* Genome Research, 2004. **12**: p. 2388-96.
17. Hitte, C., et al., *Opinion: Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping.* Nat Rev Genet, 2005.
18. Breen, M., J. Bullerdiek, and C.F. Langford, *The DAPI banded karyotype of the domestic dog (Canis familiaris) generated using chromosome-specific paint probes.* Chromosome Res, 1999. **7**(5): p. 401-406.

19. 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, 2001. **11**(10): p. 1784-95.
20. Breen, M., et al., *An integrated 4249 marker FISH/RH map of the canine genome.* BMC Genomics, 2004. **5**(1): p. 65.
21. Jaffe, D.B., et al., *Whole-genome sequence assembly for mammalian genomes: Arachne 2.* Genome Res, 2003. **13**(1): p. 91-6.
22. Clamp, M., et al., *A revised human gene catalog.* (in preparation),