Journal of Heredity Advance Access published July 25, 2007

Journal of Heredity
doi:10.1093/jhered/esm055

Published by Oxford University Press 2007.

Genome Annotation Resource Fields—GARFIELD: A Genome Browser for *Felis catus*

JOAN U. PONTIUS AND STEPHEN J. O'BRIEN

From the Laboratory of Genomic Diversity, Basic Research Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702 (Pontius); and the Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702 (O'Brien).

Address correspondence to J. U. Pontius at the address above, or e-mail: pontiusj@ncifcrf.gov.

Abstract

Annotation features from the 1.9-fold whole-genome shotgun (WGS) sequences of domestic cat have been organized into an interactive web application, Genome Annotation Resource Fields (GARFIELD) (http://lgd.abcc.ncifcrf.gov) at the Laboratory of Genomic Diversity and Advanced Biomedical Computing Center (ABCC) at The National Cancer Institute (NCI). The GARFIELD browser allows the user to view annotations on a per chromosome basis with unplaced contigs provided on placeholder chromosomes. Various tracks on the browser allow display of annotations. A Genes track on the browser includes 20 285 regions that align to genes annotated in other mammalian genomes: *Homo sapiens, Pan troglodytes, Mus musculus, Rattus norvegicus, Bos taurus*, and *Canis familiaris*. Also available are tracks that display the contigs that make up the chromosomes and representations of their GC content and repetitive elements as detected using the RepeatMasker (http://www.repeatmasker.org). Data from the browser can be downloaded in FASTA and GFF format, and users can upload their own data to the display. The *Felis catus* sequences and their chromosome assignments and additional annotations incorporate data analyzed and produced by a multicenter collaboration between NCI, ABCC, Agencourt Biosciences Corporation, Broad Institute of Harvard and Massachusetts Institute of Technology, National Human Genome Research Institute, National Center for Biotechnology and Information, and Texas A&M.

The Generic Genome Browser (Gbrowse; Stein et al. 2002) was employed to organize the annotations of the *Felis catus* genome into the online tool: Gene Annotation Resource Fields (GARFIELD). The Gbrowse interface is used by several other genome projects, including Mouse Genome Informatics (http://gbrowse.informatics.jax.org/cgi-bin/gbrowse/mouse_current/), the HapMap project (http://www.hapmap.org), WormBase (http://www.wormbase.org), and the Rat Genome Database (http://rgd.mcw.edu/). Gbrowse allows easy access to genomic data through the use of a graphical user interface that includes a chromosome view, a regional view of user-selected chromosomal regions, and lastly, a text view, detailing information related to individual features. It allows the user to download the DNA sequence and feature annotations of selected regions and allows users to upload their own annotations to display. The browser can be queried with key words such as a gene title or symbol or with terms describing gene function, as assigned by the Gene Ontology database (http://www.geneontology.org/).

The annotations available in GARFIELD summarize the work of a multicenter collaboration to annotate the whole-genome shotgun (WGS) sequence (GenBank accession AANG00000000) at 1.9-fold sequence density of the *F. catus* genome (Pontius et al., forthcoming). More than 6 million WGS reads were assembled into 817 956 contigs, and these were assigned chromosome positions by making use of 1680 RhMarkers (Murphy et al. 2007) as well as sequence alignment to the assembled dog and human genomes. Annotations include 20 285 putative genes, more than 300 000 single-nucleotide polymorphisms (SNPs), more than 200 000 short-tandem repeats (STRs), and dozens of integrated elements such as nuclear mitochondrial DNA and endogenous retroviruses (Table 1, Figure 1).

GARFIELD includes hyperlinks between the annotated features and related resources on the internet. The cat genome made extensive use of the Genomes, Genes, and HomoloGene databases at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov; Wheeler et al. 2005). The cat contigs were aligned to 6 mammalian genomes provided by NCBI (human, chimp, mouse, rat, cow, and dog) using MEGABLAST (Zhang et al. 2000), retaining only the reciprocal best matches (RBM): those alignments which represent, for each region of each genome, the best matched alignment between

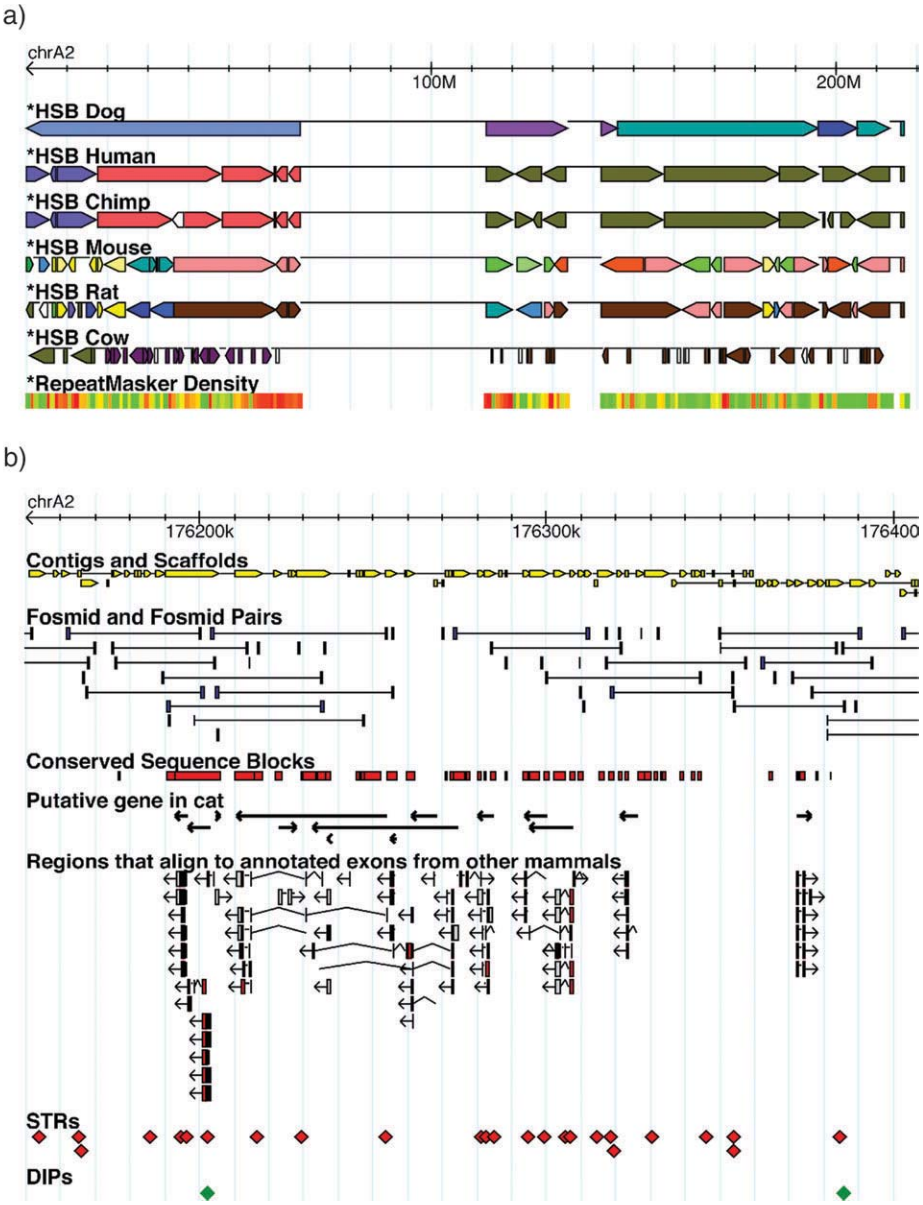

Figure 1. Example display of GARFIELD genome browser for cat. (**a**) Chromosome A2 showing the gaps in the assembly, HSBs corresponding to 6 other mammalian genomes, and the density of repetitive elements represented using red–green box (red being high density, green being low). (**b**) HOX gene cluster of chromosome A2, showing the sequenced contigs of the region, fosmids with their end reads, Conserved Sequence Blocks (CSBs) shared by 6 other mammalian genomes, putative cat genes, the alignments to the transcripts of the six other mammalian genomes, Short Tandem Repeats (STRs), and Deletion Insertion Polymorphisms (DIPs).

it and the second genome as measured by the MEGABLAST "bitscore." These alignments were used to define conserved sequence blocks (CSBs), sequences common to all the mammalian genomes analyzed here. For each genome pair, the CSBs that fell in consistent order on both the chromosomes of cat and the second genome were merged to form homologous synteny blocks (HSBs), defining large-scale orthologous regions on the chromosomes. The termini of the HSBs represent chromosomal breakpoints that have resulted in evolutionary reorganization of the genome segments among different mammals.

The RBM alignments were also used to assign putative genes in cat. Mammalian gene annotations that spanned the RBM aligned regions were assigned to their corresponding regions on the cat genome. This resulted in annotations from more than 19 000 genes each from the chimp, human, dog, and cow genomes and more than 17 000 genes each from the mouse and rat genomes being assigned to

Table 1. GARFIELD features

Map coordinates and GC content histogram	2.7 Gb
Contigs and their scaffolds	817 956 contigs
Fosmid end reads and fosmid pairs	656 655 fosmid reads (192 006 pairs)
CSBs	133 499
HSBs	2590 HSBs among 6 indexed mammals
Putative cat genes	20 285
Micro RNA loci	201
SNPs	327 037
Deletion–insertion polymorphisms	34 850
Repetitive elements	904 414
STRs	208 177
Integrated elements	Numts and endogenous retroviruses
Features with tracks to represent density per 100 kb	Repetitive elements heterozygosity (SNP density)

orthologous regions on the cat genome. These 6 sets of gene orthologs discerned by the 6 indexed mammalian genomes were then reviewed and merged to generate a nonredundant set of 20 285 putative cat genes. This merging of mammalian gene orthologs took into account the extent of the orthologs' representation and overlap on the cat chromosomes (Table 2), as well as their orthology as reported by NCBI's HomoloGene's database. On GARFIELD, the region of the assembly that spans each gene is shown in the Genes track. In the mRNA track, GARFIELD shows regions that align to the longest transcript of the annotated mammalian gene.

Currently, the majority of GARFIELD annotations are not available from the genome browsers at NCBI (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=9685), University of California Santa Cruz (UCSC) (http://genome.ucsc.edu/cgi-bin/hgGateway?org=Cat), and ENSEMBL (http://www.ensembl.org/Felis_catus/index.html). The genome browser at NCBI currently consists of a minimal set of genetic and radiation hybrid markers, whereas the browsers at UCSC and ENSEMBL include the cat contigs and scaffolds, as well as alignments to other genomes. Feature annotations that are unique to GARFIELD include chromosomal assignments of the contigs, SNPs, deletion/insertion polymorphisms, and regions representing potential nuclear mitochondrial DNA (*numts*) and endogenous retroviruses. Several resources provided by Advanced Biomedical Computing Center increase the functionality of GARFIELD. These include suggested primer pairs for the amplification of STRs, as well as the ability to query the cat assembly using a DNA sequence.

GARFIELD has proved useful in detecting not only biologically relevant aspects of the cat genome but also in revealing assembly and annotation artifacts that should be considered in the interpretation of genomic data. For example, at first, the putative cat gene *DDX25* presents an unusual rearrangement in cat, with the 3′ untranslated region being placed between 2 coding exons. However, the disposition of this unusual arrangement is the consequence of the positioning of a single contig, suggesting that the arrangement could stem from a single contig being misplaced.

Another interesting result from the genome annotation includes a list of 1586 cases of genes from the annotated mammalian gene sets that were flagged as chimeric representations of two genes. For example, the exons of a gene annotated in chimp as being from the gene *CCR5* align to the cat genome at precisely the same loci of the exons of what is annotated as being *CCR2* and *CCR5* in the other mammalian genomes. These cases were striking using the representations of data on the GARFIELD browser and would have been missed without a visual representation of the data. We suggest that the genome annotation of these chimeric cases be reviewed and the exons perhaps reassigned to 2 separate genes.

In the future, we hope to incorporate additional functionality into GARFIELD. One challenge in the display of a genome is the representation of large-scale insertions and deletions compared with other genomes. For example, because of the low coverage of the cat genome, it is not immediately obvious that the 1.9x WGS assembly includes the appropriate deletion in the gene *TAS1R2*, which is

Table 2. Genes screened for use in the modeling of 20 285 putative cat genes

	Human	Chimp	Mouse	Rat	Cow	Dog
NCBI genome release build number	35	1	35	3	2	2
No. protein-coding genes annotated by NCBI	22 073	21 465	31 097	22 573	22 783	19 756
No. genes withdrawn by NCBI (September 2006)	1197	5627	1466	1254	10	0
No coverage on the cat genome	2628	2441	12 244	6078	3279	2093
Low coverage on the cat genome (<5%)	362	368	1056	1019	172	65
Chimeric	NA	1019	52	189	153	173
Singletons	NA	387	NA	200	643	830
Total genes used in modeling cat genes	17 886	11 623	16 279	13 833	18 525	16 595

NA, not applicable.

Protein-coding genes from NCBI were excluded from use in modeling a cat gene if they were withdrawn from an updated release of NCBI's gene's database, had lower than 5% of their length represented on the cat genome, were chimeras of what was represented as 2 separate genes in other genomes, or, lastly, in the case of all genomes except human and mouse, if their aligned region on the cat genome was not confirmed by a gene from at least one other genome.

responsible for cats' inability to taste sweet foods (Li et al. 2005).

Acknowledgments/Funding

This project has been funded in whole or in part with federal funds from the National Cancer Institute. National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-Co-12400. This Research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

Li X, Li W, Wang H, Cao J, Maehashi K, Huang L, Bachmanov AA, Reed DR, Legrand-Defretin V, Beauchamp GK, et al. 2005. Pseudogenization of a sweet-receptor gene accounts for cats' indifference toward sugar. PLoS Genet. 1(1):27–35.

Murphy WJ, Davis B, David VA, Agarwala R, Schäffer AA, Pearks-Wilkerson AJ, Neelam B, O'Brien SJ, Menotti-Raymond M. 2007. A 1.5-Mb-resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human genomes. Genomics. 89:189–196.

Pontius JU, Mullikin JC, Smith D, Agencourt Sequencing Team, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B, et al. Forthcoming. Initial Sequence and Comparative Analysis of the Cat Genome. Genome Research.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The generic genome browser: a building block for a model organism system database. Genome Res. 12(10):1599–1610.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, et al. 2005. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33(Database issue):D39–D45.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7. (1–2)203–214.

This paper was delivered at the 3rd International Conference on the Advances in Canine and Feline Genomics, School of Veterinary Medicine, University of California, Davis, CA, August 3–5, 2006.

Corresponding Editor: Urs Giger