Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

GENOME RESEARCH

Initial sequence and comparative analysis of the cat genome

Joan U. Pontius, James C. Mullikin, Douglas R. Smith, Agencourt Sequencing Team, Kerstin Lindblad-Toh, Sante Gnerre, Michele Clamp, Jean Chang, Robert Stephens, Beena Neelam, Natalia Volfovsky, Alejandro A. Schäffer, Richa Agarwala, Kristina Narfström, William J. Murphy, Urs Giger, Alfred L. Roca, Agostinho Antunes, Marilyn Menotti-Raymond, Naoya Yuhki, Jill Pecon-Slattery, Warren E. Johnson, Guillaume Bourque, Glenn Tesler, NISC Comparative Sequencing Program and Stephen J. O'Brien

Genome Res. 2007 17: 1675-1689
Access the most recent version at doi:10.1101/gr.6380007

Supplementary data *"Supplemental Research Data"*
http://www.genome.org/cgi/content/full/17/11/1675/DC1

References This article cites 97 articles, 41 of which can be accessed free at:
http://www.genome.org/cgi/content/full/17/11/1675#References

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Notes

To subscribe to *Genome Research* go to:
http://www.genome.org/subscriptions/

© 2007 Cold Spring Harbor Laboratory Press

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Initial sequence and comparative analysis of the cat genome

Joan U. Pontius,[1,17] James C. Mullikin,[2] Douglas R. Smith,[3] Agencourt Sequencing Team,[3,16] Kerstin Lindblad-Toh,[4] Sante Gnerre,[4] Michele Clamp,[4] Jean Chang,[4] Robert Stephens,[5] Beena Neelam,[5] Natalia Volfovsky,[5] Alejandro A. Schäffer,[6] Richa Agarwala,[6] Kristina Narfström,[7] William J. Murphy,[8] Urs Giger,[9] Alfred L. Roca,[1] Agostinho Antunes,[10,11,12] Marilyn Menotti-Raymond,[10] Naoya Yuhki,[10] Jill Pecon-Slattery,[10] Warren E. Johnson,[10] Guillaume Bourque,[13] Glenn Tesler,[14] NISC Comparative Sequencing Program,[15] and Stephen J. O'Brien[10,17]

[1]Laboratory of Genomic Diversity, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA; [2]Comparative Genomics Unit, National Human Genome Research Institute, Rockville, Maryland 20892, USA; [3]Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA; [4]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02141, USA; [5]Advanced Biomedical Computing Center, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA; [6]National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, Maryland 20894, USA; [7]Department of Ophthalmology (Mason Eye Institute), Department of Veterinary Medicine & Surgery, University of Missouri–Columbia, Columbia, Missouri 65211, USA; [8]Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843, USA; [9]Section of Medical Genetics, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [10]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702, USA; [11]REQUIMTE, Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal; [12]CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal; [13]Genome Institute of Singapore, Singapore 138672, Republic of Singapore; [14]Department of Mathematics, University of California, San Diego, California 92093-0112, USA; [15]NISC, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

The genome sequence (1.9-fold coverage) of an inbred Abyssinian domestic cat was assembled, mapped, and annotated with a comparative approach that involved cross-reference to annotated genome assemblies of six mammals (human, chimpanzee, mouse, rat, dog, and cow). The results resolved chromosomal positions for 663,480 contigs, 20,285 putative feline gene orthologs, and 133,499 conserved sequence blocks (CSBs). Additional annotated features include repetitive elements, endogenous retroviral sequences, nuclear mitochondrial (*numt*) sequences, micro-RNAs, and evolutionary breakpoints that suggest historic balancing of translocation and inversion incidences in distinct mammalian lineages. Large numbers of single nucleotide polymorphisms (SNPs), deletion insertion polymorphisms (DIPs), and short tandem repeats (STRs), suitable for linkage or association studies were characterized in the context of long stretches of chromosome homozygosity. In spite of the light coverage capturing ~65% of euchromatin sequence from the cat genome, these comparative insights shed new light on the tempo and mode of gene/genome evolution in mammals, promise several research applications for the cat, and also illustrate that a comparative approach using more deeply covered mammals provides an informative, preliminary annotation of a light (1.9-fold) coverage mammal genome sequence.

[Supplemental material is available online at www.genome.org.]

During 2005, the National Human Genome Research Institute (NHGRI) endorsed a "light" coverage (2×) whole-genome sequencing strategy for 26 mammals, including *Felis catus*, the domestic cat (Supplemental Fig. S1). The domestic cat was included in this mammalian genome set mainly to stimulate genome research on a species that provides a large number of important human medical models (Supplemental Table S1). Cats, like dogs, enjoy an extensive veterinary medical surveillance that has described ~200 genetic diseases analogous to human disorders (Griffin and Baker 2000; O'Brien et al. 2002; O'Brien 2004). Feline infectious agents offer powerful natural models of deadly human diseases including feline immunodeficiency virus (FIV)-AIDS, feline coronavirus (FeCoV)-SARS and avian influenza, canine distemper virus (CDV)-neurotropic viruses, and feline leukemia and

[16]Agencourt Sequencing Team: Adrianne Brand, Heather Ebling, David J. Saranga, Marc Rubenfield, Matthew J. Parisi, Wei Tao, Nadeem Tusneem, Robert David, Erick Gustafson, Jason Tsolas, and Keven McKernan.
[17]Corresponding authors.
E-mail Pontiusj@ncifcrf.gov; fax (301) 846-1686.
E-mail Obrien@ncifcrf.gov; fax (301) 846-1686.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.6380007.

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

sarcoma virus (FeLV, FeSV)-leukemia and sarcoma (O'Brien et al. 2002; Kuiken et al. 2004; O'Brien 2004). Cats are a domesticated representative of a family, Felidae, that includes some of the most successful, but now the most threatened, predator species to walk the earth (Nowell and Jackson 1996; O'Brien and Johnson 2005). The rich literature of feline disease pathogenesis, human fascination for cats in art and history, plus the demonstration of a highly conserved ancestral genome organization make the cat genome annotation a highly informative advance that complements other research endeavors (O'Brien et al. 1999; Lyons et al. 2004; Fyfe et al. 2006; Johnson et al. 2006).

Although 2× genome coverage would provide limited (<80%) sequence representation of a species, it supports the primary goal of cost-effective identification of highly conserved sequence elements revealing patterns of conservation and divergence across the mammalian radiation. Moreover, the availability of annotation of six mammal genomes sequenced at high coverage (Wheeler et al. 2005; http://www.ncbi.nlm.nih.gov), a radiation hybrid (RH) physical map of cat including some 1680 markers (Murphy et al. 2007), plus new algorithms and bioinformatics tools described here have facilitated a depth of genome annotation not possible for earlier light-coverage genomes (Kirkness et al. 2003; O'Brien and Murphy 2003).

Recognition of expressed genes in a genome entails deciding which regions correspond to potential transcripts. Common strategies to this end include aligning the genome to cDNA sequences, such as the Riken mouse cDNA library (The FANTOM Consortium and the RIKEN Exploration Research Group Phase I and II 2002); the detection of similarity between imputed translations of genomic DNA and sequences of known proteins; and the use of de novo predictions of genes based on open reading frames using such programs as GENSCAN (Burge and Karlin 1997). The availability of six deep-coverage (3.6- to 7-fold) mammalian genomes enables the annotation of additional mammalian genome sequences by a combination of sequence alignments, map locations, and gene identification described here for the cat genome. The comparative strategy has the added advantage of sequence homology alignments between annotated genes, introns, and intergenic regions, as well as gene order synteny as a criterion for gene homology.

Here we describe an annotation of the cat genome based on 1.9× genome sequence. In spite of light sequence coverage, the analysis revealed numerous genome sequence features; a view of the rearrangements that have occurred since the primate, rodent, and carnivore orders diverged; and interesting aspects that stand out for cats among mammals. Our key findings include:

1. Definition, assembly, and chromosome mapping of >1 million reciprocal best match alignments (RBMs) between cat and six other mammalian genomes, revealing a set of 133,499 conserved sequence blocks (CSBs) present in all seven genomes (Supplemental Tables S2 and S3).
2. Assigning the chromosomal position of CSBs in six mammals enabling the construction of a set of homologous synteny blocks (HSBs) and use of these blocks to discern chromosomal exchanges between cat and other mammalian genomes. Deduced patterns of chromosomal exchanges that punctuate mammalian genome evolution (Figs. 1, 2; Supplemental Table S4), revealing a balance between inter- and intrachromosomal rearrangements in primate, rodent, and carnivore lineages.
3. Discovery and mapping of some 20,285 regions orthologous to genes annotated in other mammalian genomes (Table 1).

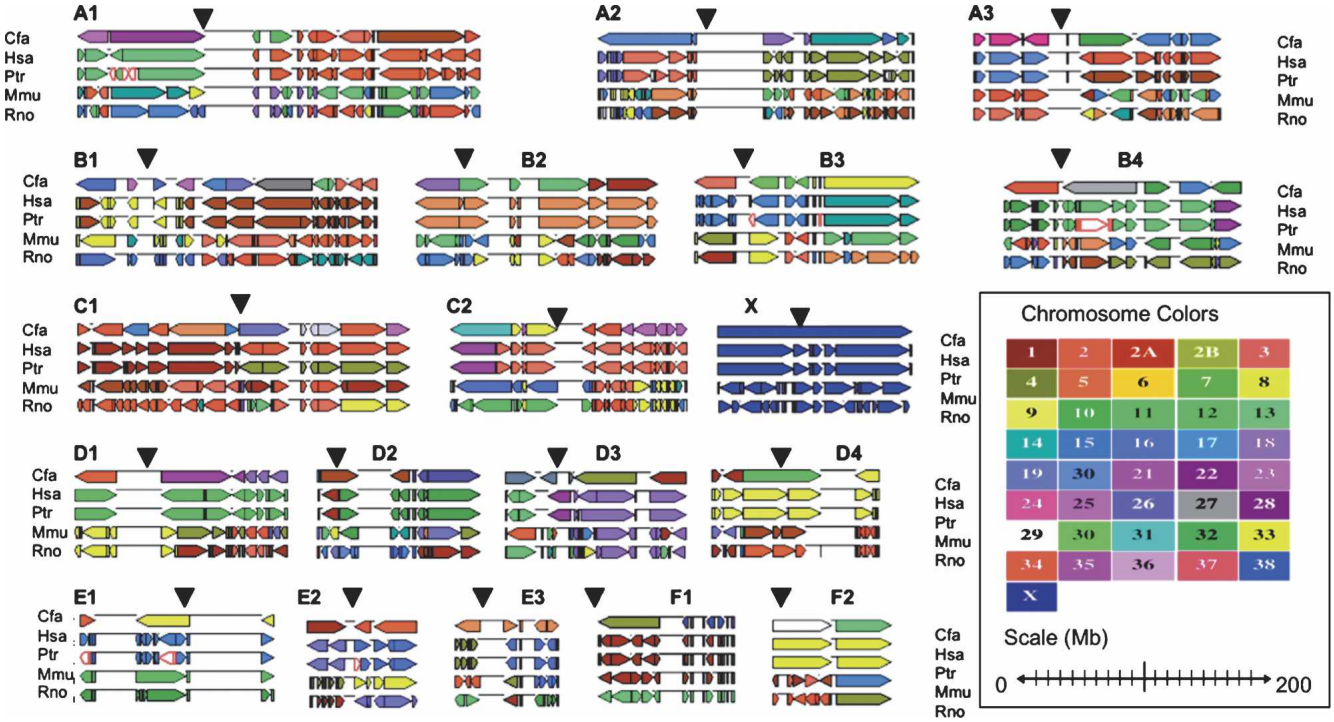

Figure 1. Homologous synteny blocks (HSBs) of the cat genome as compared to corresponding syntenic blocks in five mammalian species: (Cfa) *Canis familiaris*, (Hsa) *Homo sapiens*, (Ptr) *Pan troglodytes*, (Mmu) *Mus musculus*, and (Rno) *Rattus norvegicus*. The empty line between blocks indicates ambiguous regions, which may arise, for example from gaps in the genome assemblies. White regions with red borders represent alignments to unplaced contigs. Black triangle represents approximate centromere position.

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

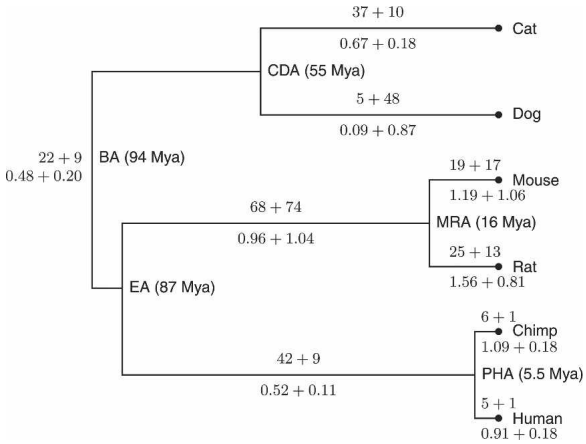

Figure 2. Counts and rates of chromosome rearrangements detected among human, chimpanzee, cat, dog, mouse, and rat genomes, based on HSBs of at least 500 kb. Branches are labeled on the *top* with the estimated minimum number of intrachromosomal rearrangements (inversions) plus interchromosomal rearrangements (translocations, fissions, and fusions). Branches are labeled on the *bottom* with the estimated minimum rates of intrachromosomal plus interchromosomal rearrangements per million yr. The ancestors are BA (Boreoeutherian), EA (Euarchontoglires), CDA (cat/dog), MRA (mouse/rat), and PHA (chimpanzee/human).

4. Details of interspersed repeat families relative to the human and dog genomes (Table 2).
5. Detection of previously undiscovered lineages of nuclear mitochondrial (*numt*) gene sequences distributed across all cat chromosomes (Fig. 3; Supplemental Fig. S2).
6. Description of previously undiscovered retroviral elements that are 10 times more abundant in the cat genome than FeLV or RD114 sequences (Fig. 4; Supplemental Fig. S3; Table 3).
7. Identification of 327,000 new SNPs, 208,177 new STR loci, and a mosaic genome pattern of homozygosity useful for linkage disequilibrium mapping of complex traits in the cat (Fig. 5).
8. A dynamic online genome browser of Genome Annotation Resource Field (GARFIELD) (Pontius and O'Brien 2007) for the display and download of the assembly annotations, with hyperlinks to and related bioinformatics resources (http://lgd.abcc.ncifcrf.gov; Fig. 6).

The feline genome sequence, assembly, and map

We sequenced the genome of a female Abyssinian cat (Cinnamon, who resides at the University of Missouri, Columbia, MO) at Agencourt Bioscience Corp. (Beverly, MA), using a whole-genome shotgun (WGS) approach (Supplemental Table S5). Sequence data were assembled using both the PHUSION (Mullikin and Ning 2003) and Arachne (Batzoglou et al. 2002; Jaffe et al. 2003) programs, with the latter incorporating a novel assisted assembly algorithm (Supplemental Methods). This algorithm uses the read placement on two reference genomes (human and dog) to confirm the linking information from their independent assembly. A total of 8,027,672 sequence reads (84% from plasmids and 16% from fosmid paired ends) were assembled to 817,956 contigs, covering 1.642 Gb of sequence with an N50 contig length (i.e., half of the sequenced base pairs reside in contigs >N50) of 2378 bp (Supplemental Table S6). Contigs were then assembled into scaffolds (N = 217,790, N50 length of 117 kb) assisted by a close evolutionary relationship of cat and dog, both members of the mammalian order Carnivora. The cat genome size estimate of 2.7 Gb (2.5 Gb euchromatin) was imputed by extrapolating the average length of the cat sequence within fosmids as compared to the homologous sequence stretches in the dog and human genomes, presuming a genome size of 2.8 Gb for human and 2.4 Gb for dog. The cat genome coverage (~1.9×) is lower than the finished or "deep" coverage mammalian genome sequence recently annotated (finished human, mouse and rat 7×, dog 7.5×, and cow 7.1×), but between the chimpanzee (6×) and the earlier dog (1.5×) (International Human Genome Sequencing Consortium 2001; Venter et al. 2001; Mouse Genome Sequencing Consortium 2002; Kirkness et al. 2003; Rat Genome Sequencing Project Consortium 2004; Chimpanzee Sequencing and Analysis Consortium 2005; Lindblad-Toh et al. 2005; http://www.hgsc.bcm.tmc.edu/projects/bovine/). The relatively low coverage is reflected in more contigs, lower contig N50, and lower overall euchromatin genome coverage (60% of the 2.7-Gb genome, or 65% of the 2.5-Gb euchromatin genome) than for the higher coverage genomes.

To order and orient the assembly onto the feline chromosomes, we constructed a sequence map using comparative data available from high-coverage dog and human genomes. The map was built by first aligning the feline sequence to the CanFam2 version of the domestic dog genome (Lindblad-Toh et al. 2005) using BLASTZ (see Supplemental Methods). Initially, scaffolds were placed in the same order and orientation as their homologs in dog. When scaffolds defined by the Arachne assembly mapped to more than a single dog chromosome (425 scaffolds represented by at least two included contigs), the scaffolds were broken and placed on separate dog chromosomes. A total of 1680 ordered markers from the radiation hybrid map (Murphy et al. 2007) were then used to place scaffolds on cat chromosomes. The mapped feline contigs include 1.36 Gb from 663,480 contigs, or 54% of the 2.5-Gb euchromatin genome. The cat whole-genome sequence (WGS) contigs are available at the Broad Institute, the National Center for Biotechnology Information (NCBI), Ensembl, UCSC (http://www.broad.mit.edu; http://www.ncbi.nlm.nih.gov; http://www.ensembl.org; www.genome.ucsc.edu), and the assembly ordered on the cat chromosomes can be found at the NCI-GARFIELD browser (http://lgd.abcc.ncifcrf.gov).

To evaluate the assembly and mapping achieved for the cat genome sequence, we compared it to two "finished" cat genome subset regions, each derived from BAC-based (bacterial artificial chromosome library) sequence assemblies. First, we evaluated 18 Mb of the 30 Mb of the ENCODE project sequenced for cat by the NIH Intramural Sequencing Center (ENCODE Project: ENCyclopedia Of DNA Elements; The ENCODE Project Consortium 2004; Guigó et al. 2006). All but four of 30 ENCODE regions (Table 4) map to cat chromosomes, and two of these are split into segments that are not assigned to cat chromosomes. To gauge the level of orientation problems, the alignment of each ENCODE region was divided into 1-kb segments, and these segments as mapped onto the cat assembly were evaluated for correct orientation. The most challenging region was the beta-globin region, ENm009, with the fraction of correctly oriented 1-kb segments at 85%. This region is also difficult for the NISC BAC assembly as there were three BAC clone gaps for this region. Supplemental Figure S4 shows the coverage and average contig size for each of the regions calculated by aligning the multi-BAC assemblies for each region to the WGS assembly and then taking all ordered and oriented contigs within that segment of the WGS assembly and realigning the two versions of each ENCODE region to each other

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 1. Features of the cat genome annotation and of other mammalian genomes

	Human	Chimpanzee	Mouse	Rat	Cow	Dog
NCBI Build Release no.	35	1	35	3	2	2
No. of contigs (cat = 817,956)	377	37,922	5041	740	102,339	3314
Gene no. (cat = 20,285)						
NCBI annotated as protein coding[a]	22,073	21,465	31,093	22,573	22,782	19,756
NCBI with annotated CDS	22,039	21,452	31,030	22,539	22,771	19,747
Genes with homologs in cat	17,882	11,342	16,153	13,359	18,163	16,177
Percent of genes represented[b]	90.2	78.8	67.4	69.3	79.8	81.9
Average percent coverage (introns + exons)	39.7	36.9	27.4	26.9	44.6	55.7
Genes with CDS represented in cat[c]	17,101	11,065	15,415	13,305	17,746	15,948
Percent of genes represented[b]	86.2	76.9	64.3	68.1	77.9	80.7
Average percent coverage of CDS only	69.4	66.3	62.2	59.9	69.1	71.7
No. of cat genes analyzed for syntenic order[d]	16,190	15,352	14,546	12,678	11,972	14,934
Triplet exact no.	14,217	10,680	11,863	10,617	6256	13,793
Triplet exact percent	87.8	69.6	81.6	83.7	52.3	92.4
Triplet inexact no.	14,567	10,986	12,272	10,876	6464	14,044
Triplet inexact percent	90.0	71.6	84.4	85.8	54.0	94.0
Percent of genome represented by cat RBMs	27.0	26.0	9.6	9.0	26.1	41.0
STR counts (cat = 208 K)	542 K	478 K	1.35 M	1.224 M	NA	955 K
SNP count (cat = 327 K)	1.4 M	1.7 M	80 K	—	—	2.5 M

[a]Pseudogenes, RNA genes, and genes restricted to the Y chromosome were excluded from this analysis.
[b]The percent of NCBI annotated genes represented in cat was calculated excluding any genes that had been withdrawn from NCBI's Genes database as of September 18, 2006.
[c]CDS coding sequence is lower than genes in these counts because CDS are exons while genes include exons, introns, and 5′- and 3′-UTRs.
[d]Genes were considered to have syntenic orthology when the three same genes occur in the same order in the cat genome as in the indexed genome, or when at least one of the genes two neighbors away is also flanking the index species gene (see text). Species with large numbers of unmapped contigs (cow and chimp) show lower syntenic orthology as a consequence, and these estimates should not be considered robust.

using Cross_match software (P. Green, unpubl.). From this, we derived coverage and average contig size for each region. We also mapped the position of each contig in the assembly to its corresponding region in the ENCODE multi-BAC assembly, producing the plots shown in Supplemental Figure S5 for six example regions. The long-range order and orientation (length of chromosomes) were confirmed in these analyses across the sampled ENCODE regions.

The second region was the 3.3-Mb feline major histocompatibility complex *FLA*, which has been sequenced and ordered from a BAC library (O'Brien et al. 1999; Yuhki et al. 2003; Beck et al. 2005) (Fig. 7). *FLA* is broken into two distinct regions of the cat's genome, the first containing class II, class III, and 12 class I-like genes (2,975,515 bp) adjacent to the chromosome B2q centromere, and the other (361,545 bp) on the B2p telomere containing six genes of the distal class I region. The high density of genes, the existence of multiple gene paralogs, plus the rapid evolution of class I and class II genes render this region among the most challenging for gene annotation. In Figure 7, we present a map comparing the sequence coverage of the feline whole-genome sequence aligned to the *FLA* BAC, to the dog major histocompatibility complex, *DLA*, represented in the dog WGS assembly, and to the human major histocompatibility complex, *HLA*. From this comparison, the feline WGS includes a detectable sequence from 191 of 202 *FLA* gene sequences (95%) present in the *FLA* BACS (58 class II, 40 class III, and 93 class I region genes).

Table 2. Interspersed repeat elements of cat and other mammals

	Percent of genome				Percent of MHC		ENCODE region[a]	
	Human	Mouse	Dog	Cat	Cat	Human	Cinnamon	ENCODE
SINEs	13.63	7.96	10.57	11.20	8.53	17.59	11.3	12.6
MIRs	2.91	0.58	2.70	3.10	1.05	16.06	2.6	2.9
LINEs	21.05	19.54	18.74	14.26	21.31	16.59	11.5	16.0
LINE1	17.43	19.10	15.57	10.79	18.63	13.35	8.9	12.2
LINE2	3.25	0.38	2.84	2.82	2.54	3.09	2.3	3.2
L3/CR1	0.37	0.06	0.33	0.36	0.14	0.16	0.3	0.4
LTR elements	8.62	10.39	3.68	4.44	2.69	10.55	4.1	4.0
MaLRs	3.79	4.05	2.05	2.14	1.04	2.61	2.1	2.0
ERVL	1.61	1.08	1.19	1.21	0.81	2.11	1.1	1.1
ERV class I	2.93	0.76	0.61	1.05	0.81	4.25	0.9	0.9
ERV class II	0.01	0.00	0.01	0.04	0.00	1.57	0.0	0.0
DNA elements	3.01	0.88	1.98	2.19	1.62	2.64	2.0	2.3
MER1_type	1.39	0.62	1.31	1.26	1.31	1.52	1.2	1.3
MER2_type	1.06	0.17	0.19	0.39	0.14	0.88	0.4	0.4
Total interspersed	46.46	39.10	35.15	32.10	34.14	48.14	28.9	34.8

Cat data are based on sequences assembled into contigs with dog, human, and mouse data from (Lindblad-Toh et al. 2005).
[a]Repeat estimate in megabase region of cats sequenced by ENCODE project (The ENCODE Project Consortium 2004; Guigó et al. 2006) as well as repeat incidence in the ENCODE homologous segments of Cinnamon's genome.

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

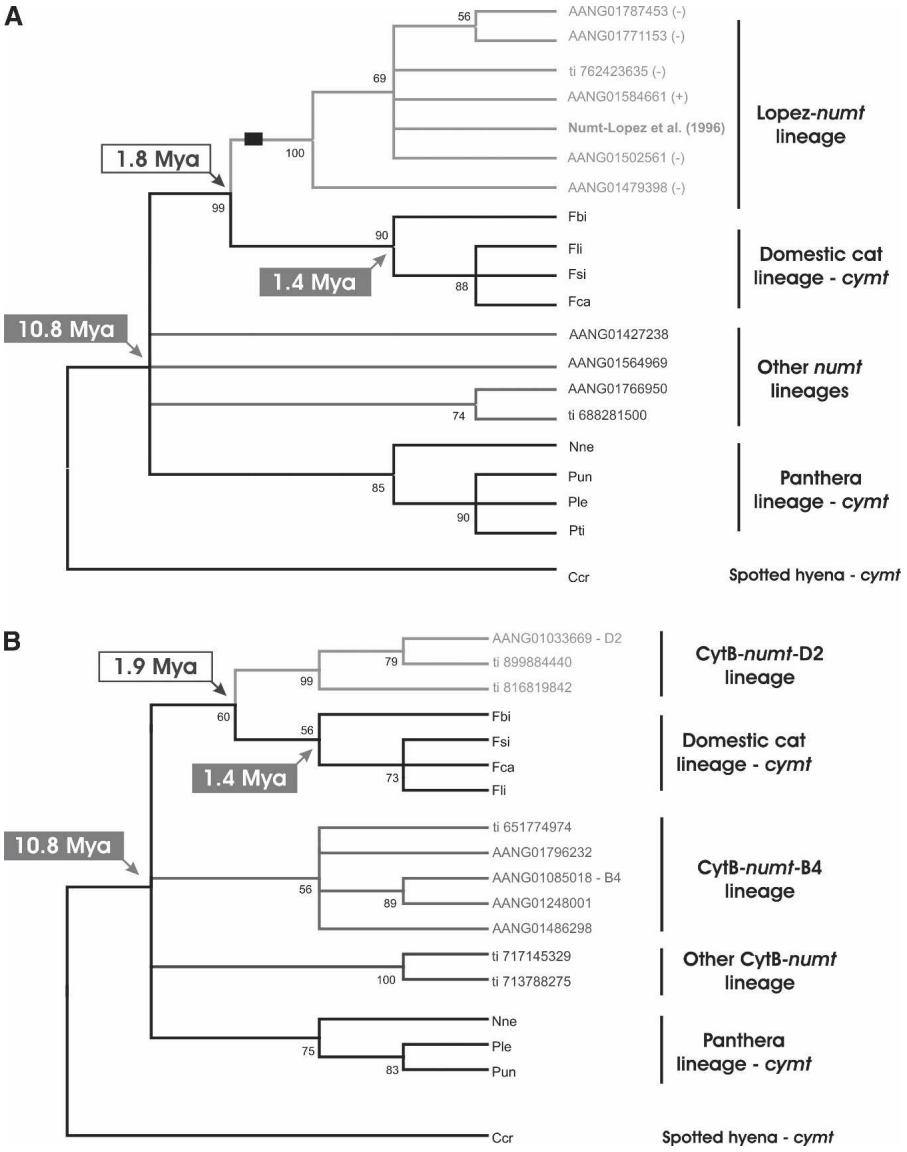

Figure 3. Phylogenetic analyses of felid mitochondrial (*cymt*) and homologous domestic cat *numt* sequences. Divergence dates of *numt* lineages-defining nodes were estimated following Lopez et al. (1994). Divergence dates (in gray boxes) assumed for the Felidae radiation and the represented species of the domestic cat lineage were retrieved from Johnson et al. (2006). Bootstrap values, calculated from 1000 replications, are placed at each branchpoint. (*A*) Minimum-evolution condensed tree (50% cutoff value) of a 570-bp segment of the mtDNA NADH dehydrogenase subunit I (*NDI* is a gene segment represented in the Lopez-*numt*) and homologous *numt* detected by BLAST searches in *F. catus*. (Black rectangle) The phylogenetically informative 10-bp deletion shared by all representatives of the Lopez-*numt* lineage. The symbols (+) and (−) represent the orientation of the genomic DNA sequence. The *numt* sequences are labeled using trace (ti), scaffold, GenBank accession, and chromosome. (*B*) Minimum-evolution condensed tree (50% cutoff value) of a 426-bp segment of the mtDNA cytochrome *b* (*CytB* is a gene segment not represented in the Lopez-*numt*) and several homologous *numt*s detected by BLAST searches in *F. catus*. The *cymt* sequences are labeled with a three-letter code: (Fca) *F. catus*, domestic cat; (Fsi) *Felis silvestris*, European wild cat; (Fli) *Felis libyca*, African wild cat; (Fbi) *Felis bieti*, Chinese desert cat; (Nne) *Neofelis nebulosa*, clouded leopard; (Pti) *Panthera tigris*, tiger; (Pun) *Panthera uncia*, snow leopard; (Ple) *Panthera leo*, lion; and outgroup (Ccr) *Crocuta crocuta*, spotted hyena.

Of these, 104 (54% of *FLA* genes) included >50% of the gene exons seen in *DLA* (CanFam 2). The large representation of MHC genes in cat WGS (95% of cat MHC genes), the recapitulation of gene syntenic orthology, the recovery of 54% of BAC-based *FLA* genes with >50% exonic sequences, plus the designation of 11,000 SNPs within *FLA* (see below) constitute an extensive annotation considering the light sequence coverage and the refractory nature of this region.

Over the course of the sequencing of the cat genome and assembly of the sequence traces, the data were used in the mapping and characterization of several cat genes (footnote a in Table 5). For two phenotypes, spinal muscular atrophy SMA-*LIX1* (Fyfe et al. 2006) and dilute *MLPH* (Ishida et al. 2006), linkage mapping was performed in pedigrees using additional STRs, tightly linked to candidate genes. The order of the STRs in these genomic regions, which cover tens of centimorgans, was exactly concordant with the orders of the markers imputed in the WGS assembly (Supplemental Table S7). In the case of both *MLPH* and SMA-*LIX1* regions, the cat marker order includes a rearrangement in the cat genome with respect to human and dog. Agreement between marker orders for SMA-*LIX1* and *MLPH* by linkage analyses versus the cat WGS assembly of the same regions plus the finished ENCODE and *FLA* regions' order agreement with the cat WGS assembly provide independent validation of cat WGS assembly for all these regions.

Genome landscape

Gene annotation

The cat genome contigs were aligned to NCBI annotated genome sequence of six index mammalian genomes (human, chimpanzee, mouse, rat, dog, and cow) (Wheeler et al. 2005; http://www.ncbi.nlm.nih.gov) using MegaBLAST (Zhang et al. 2000). These alignments include between 267,764 (cat vs. rat) and 1,235,641 (cat vs. dog) reciprocal best matches (RBMs) of average length ranging from 927 to 1000 nt (Supplemental Table S2). The mean percent identity of the alignments was highest for dog (79%), followed by cow (73.4%), primate (73.0%), and rodent (69%). Slight length discrepancies between the species imply that the primate-aligned regions are on average 0.5% longer than their cat counterpart, while those of rodent are 2.0% shorter, and cow and dog span regions of similar lengths.

Each RBM between cat and an NCBI-annotated mammal genome sequence was screened for matches with annotated genes and gene features (exons, introns, UTRs, upstream and downstream regions of protein coding genes) within that mammal. The average percent nucleotide of the NCBI annotated

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

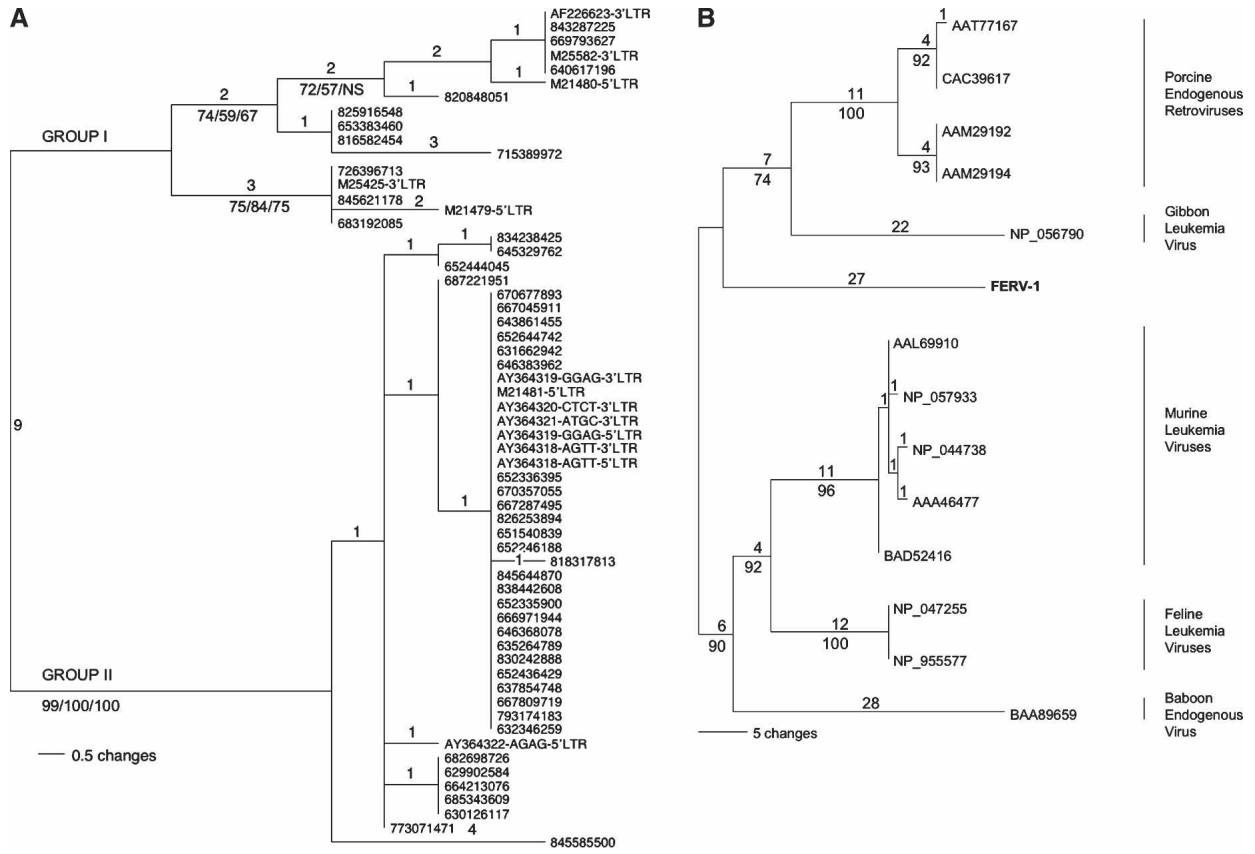

Figure 4. Phylogenetic relationship among endogenous retroviruses. (*A*) Endogenous FeLV present in the genome of the domestic cat. Analysis is based on a 274-bp alignment of 13 previously published proviral LTRs (GenBank accessions) and 47 cat genomic sequences (trace IDs). Strong bootstrap support suggests that the proviral sequences fall into two groups, as labeled. The maximum parsimony tree is shown (length = 39, CI = 0.923, RC = 0.911). Bootstrap support (>70%) is indicated for maximum parsimony, neighbor joining, and maximum likelihood methods. (*B*) FERV-1. Phylogenetic relationships among previously identified retroviruses and a novel retroelement sequence in the cat genome. One copy of the novel provirus (FERV-1) in an unpublished BAC sequence was used to generate the maximum parsimony (MP) tree (length = 142, CI = 0.915, RC = 0.851), based on the predicted sequence of 205 amino acids of Pol. The number of steps is listed *above* the branches, while MP bootstrap support is indicated *below* the branches of the major clades.

mammalian gene features that aligned to the cat sequence is summarized in Table 6. To support gene identification, syntenic orthology of a putative cat gene was assessed by determining whether its adjacent feline genes (one upstream and one downstream of a particular cat gene candidate) were homologous to adjacent genes of the gene ortholog in the index mammal genome. Cat genes with homologous neighbors on both sides in an index species were considered to be part of an exact syntenic gene triplet. For dog, 92% of the cat orthologs are exact syntenic triplets. Inexact triplet gene matches (when at least one of two upstream neighbors and one of two downstream neighbor genes matched) reach a high of 94% in dog–cat and 90% in human–cat comparisons (Table 1).

A summary of gene annotation statistics is presented in Table 1 and is available in our feline genome Web browser GARFIELD (Fig. 6). Feline gene orthologs from different mammalian species comparisons were then merged to define a set of nonredundant genes on the cat genome. The merging process involved 11 steps (detailed in Supplemental Methods) that are based on those RBM alignments that span the annotated mammalian genes, with priority given to those alignments that span exons. A list of 20,285 putative genes discovered includes homologs for 80% of annotated dog genes, 90% of human genes, and slightly fewer in the other mammals' gene lists (Table 1). The average percent coverage of these gene coding sequences ranged from 64% to 81% of the length of the gene in the NCBI annotated mammals (Table 1). Although there exist observed and anticipated weaknesses in the comparative approach (e.g., inconsistencies between different mammals, evolutionary distance, gene birth and death, gene annotation quality for different mammals, and absence of cDNA transcript sequences), the identification of >20,000 gene candidates represents a preliminary glimpse of the disposition of the cat's gene complement.

Comparative genome organization

A CSB is a sequence that is represented by an RBM in two or more genomes (Bejerano et al. 2004; Siepel et al. 2005). An invaluable application of a new mammalian genome assembly is comparative genomic inference derived from inspecting the linear position of conserved sequence (Bourque et al. 2004; Murphy et al. 2004; Everts-van der Wind et al. 2005). By applying the principles of rearrangement parsimony, one can reconstruct the extent and pattern of chromosome segment exchanges that occurred during lineage evolution among different mammalian orders. Such comparative genomic analyses enjoy much higher precision with the availability of whole-genome sequences of additional mammals.

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 3. Genomic coverage of feline retroelements

	Coverage in base pairs				
	Traces		Contigs		
Retroelement	No. of kb	Percent total	No. of kb	Percent total	Retroviral lineage
FeLV	177	2.8	85	4.8	FeLV
RD114 (partial)[a]	60	2.0	11	0.65	RD114
FERV-1	2461	39.0	612	34.7	Primate and porcine ERVs; Type C leukemia viruses
FERV-2	805	13.0	257	14.6	HERV HCML-ARV; HERV-R
FERV-3	275	4.3	69	3.9	Jaagsiekte sheep retrovirus; Ovine enzootic nasal tumor virus; HERV-K
FERV-4	1006	16.1	270	15.3	HERV-W (syncitin)
FERV-5	1478	23.6	454	25.8	Mouse mammary tumor virus; Python ERV
Total	6264		1762		

[a]RD114-matching sequence is likely an underestimate of cat RD114-related sequence by >50% since the entire RD114 genome has yet to be sequenced, limiting the regions of the RD114 genome screened (see Supplemental Fig. S3b).

To build comparative maps between cat and each of the other genomes, we used GRIMM-Synteny (Pevzner and Tesler 2003; Bourque et al. 2004) to construct 339 HSBs of size ≥500 kb (Fig. 1). These were based on the 98,313 six-way CSBs (i.e., shared between cat-dog-human-chimpanzee-mouse-rat) that were placed on chromosomes (unplaced contigs, as well as the entire cow genome, which includes large numbers of unplaced contigs, were not included in this analysis; see Supplemental Methods). For each species comparison, the imputed coordinates of the chromosome breakpoints were assembled and tabulated. These blocks represent large-scale orthologous regions that may include small-scale internal shuffling but reflect ancestral chromosomal rearrangements that occurred in different ordinal lineages.

We used the MGR and GRIMM algorithms to construct a rearrangement-based evolutionary scenario minimizing inversions, translocations, fusions, and fissions (Bourque and Pevzner 2002; Tesler 2002a,b). We estimated the number of rearrangement events among these species' genomes as well as between modern species and imputed genome arrangement of four putative ancestors: a cat/dog carnivore ancestor (CDA; 55 million years ago [Mya]), a mouse/rat ancestor (MRA; 16 Mya), a chimpanzee/human (Pan/Homo) ancestor (PHA; 5.5 Mya), and a Euarchontoglires ancestor (EA; 87 Mya) (Kumar and Hedges 1998; Springer et al. 2003).

In Figure 2, we present a phylogenetic tree of the six mammals with the number of imputed intrachromosomal and interchromosomal arrangements listed on each lineage. Three to four times more intrachromosomal rearrangements versus interchromosomal rearrangements were observed for cat and primate lineages. On the path from the Carnivore ancestor (CDA) to cat, we estimate at least 37 inversions and 10 interchromosomal rearrangements (seven translocations and three fusions), while from CDA to dog we estimate five inversions and 48 interchromosomal rearrangements (31 translocations and 17 fissions). For the primates, we estimate 42 inversions, eight translocations, and one fission on the path from the Euarchontoglires ancestor (EA) to the chimpanzee/human ancestor (PHA). For rodents, intrachromosomal and interchromosomal rearrangements are roughly balanced (Fig. 2; Supplemental Table S4).

These results reflect the well-known interchromosomal reshuffling of murid and canine genome organization relative to the more conserved disposition in the human (primate) or cat (Felidae) genomic history (O'Brien et al. 1999; Bourque et al. 2004; Murphy et al. 2004, 2005). The apparent dichotomy of the interchromosomal exchange rate (i.e., slow in most lineages, but accelerated in others) was apparent from early comparative gene mapping and chromosome painting studies, which were insensitive to intrachromosomal inversions. However, the number of intrachromosomal rearrangements for cat and primates is higher than the number of interchromosomal, while interchromosomal are predominant over intrachromosomal rearrangements in the dog and rodent lineages. Thus, in the carnivore lineage, the total rate of chromosomal exchanges (i.e., inter- and intrachromosomal) for cat (0.85/Myr) is roughly

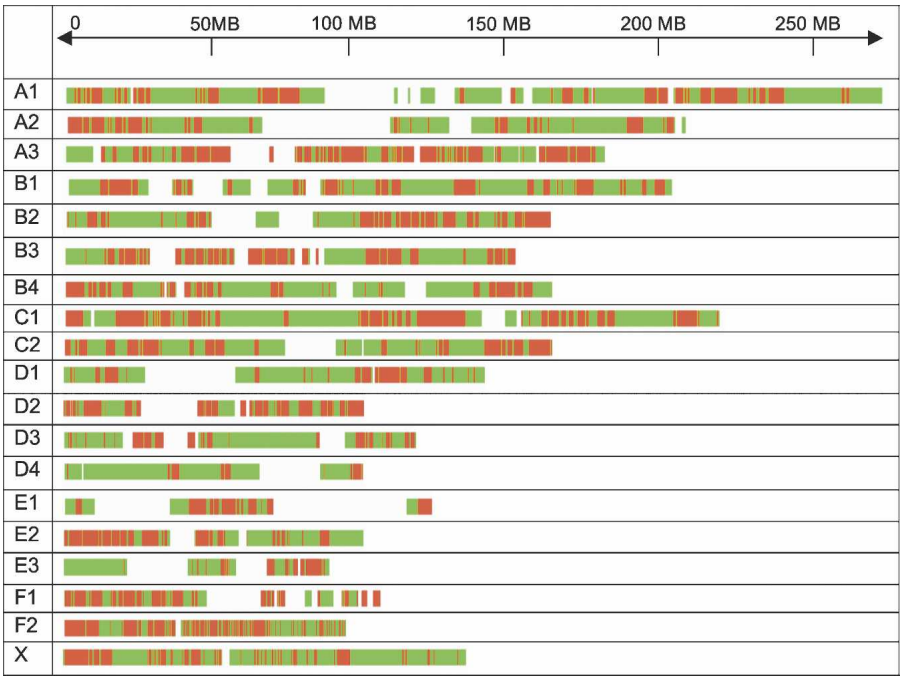

Figure 5. Homozygosity across Cinnamon's chromosomes represented in non-overlapping windows of 100 kb. (Red) Regions with more than two SNPs per 100 kb; (green) homozygous regions (<2 SNPs/100 kb); (white gaps) gaps in the chromosome assembly.

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

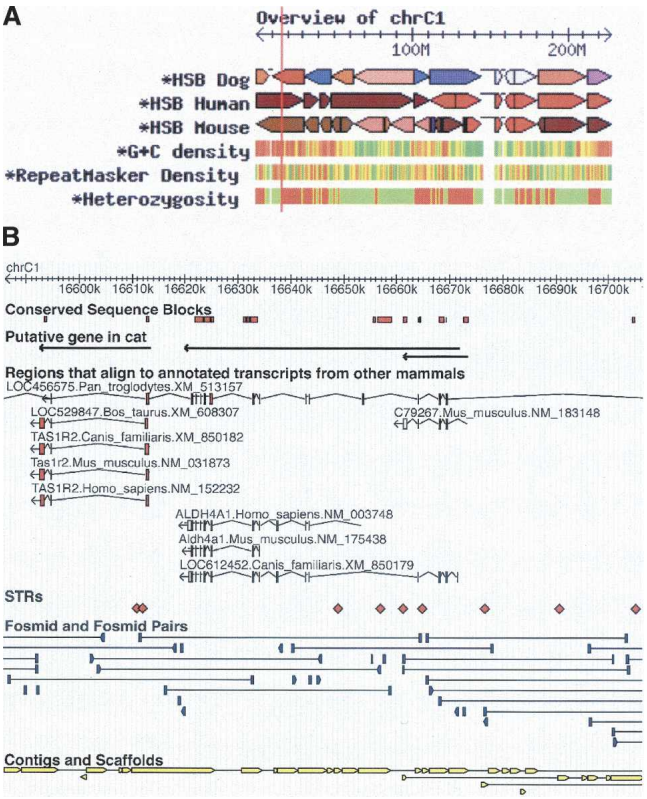

Figure 6. Gene Annotation Region Field (GARFIELD genome browser; http://lgd.abcc.ncifcrf.gov) showing the region of the cat genome corresponding to the taste receptor gene *TAS1R2* on chromosome C1 at two levels of resolution. (*A*) Chromosome view showing homologous synteny blocks (HSBs) for dog, human, and mouse; representation of G+C density; density of repetitive elements; and heterozygosity of chromosome C1. (*B*) A 110-kb view, CSBs, putative cat genes, regions that align to annotated genes in other mammalian genomes, STRs, fosmid reads with their partners, and contigs and scaffolds for the region. The GARFIELD name and image are used with permission from TM & Paws, Inc. All Rights Reserved; http://www.garfield.com.

equivalent to that of dog (0.96/Myr), and these rates are similar to the rates of rearrangements in the primate lineage for human (1.09/Myr) and chimpanzee (1.27/Myr), while being approximately half the rate in the rodent lineages, specifically, of mouse (2.25/Myr) and rat (2.38/Myr).

Together, the observations suggest a re-thinking of the previous paradigm of dichotomous modes of chromosome exchange, that is, rapid in genome-shuffled lineages such as dog, bears, murids, gibbons, and New World monkeys and a slower default mode in conserved lineages such as felids, cetaceans, most primates, and other mammals (O'Brien et al. 1999; Murphy et al. 2004). The heretofore conserved species' genomes (the slow default mode group) display a slow translocation rate that suggested a close homology with the imputed ancestral genome organization for placental mammals, estimated to have lived some 105 Mya (Murphy et al. 2001; Springer et al. 2003). However, these differing ratios of translocations to inversions, which were also suggested by the recent dog genome and cat RH map analyses (Lindblad-Toh et al. 2005; Murphy et al. 2007), show that both human and cat lineages display an apparent speed-up of intrachromosomal inversions relative to the rapid-translocation species, suggesting an overall range of breakpoint occurrence rates among the mammals studied that is narrower than previously supposed.

Repetitive elements

Mammalian genome sequences carry a sizable proportion of interspersed repeat sequences (Table 2), including vestigial mobile elements that invaded the ancestral genomes of modern species. Nearly half (46.5%) of the human genome is flagged by RepeatMasker (A.F.A. Smit, R. Hubley, and P. Green, 1996–2004. *RepeatMasker Open-3.0*; http://www.repeatmasker.org) as being repetitive, with slightly lower representation (39.1%) in the mouse. The cat sequence also shows a relatively low prevalence of interspersed repeats (32.12% of contigs). Low levels of repeats were also detected in the "finished" *FLA* region (3.8 Mb from cat BAC sequence and assembly) (Yuhki et al. 2003; Beck et al. 2005) and also in the cat's 18-Mb ENCODE region (The ENCODE Project Consortium 2004; Guigó et al. 2006), suggesting that the reduction is not an artifact of low (1.9×) coverage. The observed level does raise the possibility that the cat genome includes repetitive elements that are not included in or detected by the RepeatMasker libraries.

A more detailed analysis was done of specific repeat classes, including LINES, SINES, and Satellite DNA (Table 2; Supplemental Fig. S6; see Supplemental Materials). A feline-specific Satellite DNA (FA-SAT) reported as representing 1%–2% of the cat genome (Fanning 1987) comprised 2.1% of the 1.9× cat contigs. In spite of assigning a high percent of the LINE sequences to cat chromosomes, no full-length LINE elements were assembled. Sequence analyses of SINE elements revealed that the majority likely originated from $tRNA^{Lys}$ as observed in previous felid studies (Pecon-Slattery et al. 2000, 2004). Additional phylogenetic analyses (Supplemental Fig. S7) affirmed that the majority are members of the FC1 and FC2 SINE group that is exclusively found in cats (Fanning et al. 1988; Smit 1996). The remainder were SINEs that were related to canid and carnivore SINES (Fanning et al. 1988; Smit 1996; Vassetzky and Kramerov 2002) and appeared more ancestral.

Short tandem repeats (STRs)

An extremely useful category of polymorphic short tandem repeats (also termed microsatellites or simple sequence length polymorphism, SSLPs) is abundant across all cat chromosomes in the cat. These hypervariable STR loci have been applied in linkage mapping (Lyons et al. 2004; Fyfe et al. 2006), in forensic individualization (Menotti-Raymond et al. 1997a,b, 2005), and in assessment of historic demographic events that have molded domestic cats and wild felid species (O'Brien and Johnson 2005). We annotated 208,177 STR loci, fewer than were described for the human (542,183), dog (955,555), or mouse (1,346,134) genomes (Table 1). These STR loci have been placed on the feline map, annotated with specific position and suggested primer pairs in GARFIELD (Fig. 6).

Micro-RNAS (miRNAs)

Micro-RNAs (miRNAs) are a family of highly conserved short RNA transcripts that regulate translation of gene products (Lee et al. 2002; Lagos-Quintana et al. 2003; Lecellier et al. 2005; Hertel et al. 2006). The regulatory effects of these molecules are mediated by an interaction between a short processed section of the miRNA and the 3′-UTR of the target mRNA. Based on homology with sequences from the Micro-RNA Registry (Griffith-Jones 2004) and potential stem–loop secondary structure, we identified 179 potential miRNA feline sequences (Table 7; Supplemental Methods). These include 177 sequences that are found across most

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 4. Comparison of ENCODE regions in the multi-BAC cat assemblies to same regions in the cat genome assembly

ENCODE region	Cat chromosome	Start position	End position	Description	Size on cat assembly	Size on human hg17 assembly	Fraction of correctly oriented 1-kb segments[a]
ENm001	ChrA2	192811862	194663327	CFTR	1851465	1877426	100%
ENm002	ChrA1	140705763	141871323	Interleukin	1165560	1000000	100%
ENm005	ChrC2	11803676	13617097	Chr21 Pick	1813421	1695985	97%
ENm008	ChrE3	69002436	69597747	Alpha-globin	595311	500000	100%
ENm009	ChrD1	88638587	90768584	Beta-globin	2129997	1001592	85%
ENm010	ChrA2	176151152	176592472	HOXA	441320	500000	100%
ENm011	ChrD1	144373409	144722509	IGF2/H19	349100	606048	100%
ENm013	ChrA2	149866489	150999774	Chr7 Pick	1133285	1114424	98%
ENm014	ChrA2	211942166	213177052	Chr7 Pick	1234886	1163197	100%
ENr113	ChrB1	123220673	123341135	Random	120462	500000	100%
ENr114	ChrUn26	5255272	5573376	Random	318104	500000	100%
ENr121	ChrC1	123584532	124124315	Random	539783	500000	100%
ENr123	ChrB4	75087784	75705088	Random	617304	500000	100%
ENr131	ChrC1	221083110	221405671	Random	322561	500064	100%
ENr133	ChrC2	5672242	6243888	Random	571646	500000	100%
ENr211	ChrE3	36550752	37167922	Random	617170	500001	100%
ENr212	ChrA1	152143207	152718137	Random	574930	500000	100%
ENr213	ChrD3	69275011	69835090	Random	560079	500000	100%
ENr222	ChrUn1	1	62194	Random	62193		
ENr222	ChrUn1	2191602	2280652	Random	89050		
ENr222	ChrB2	130978637	131343199	Random	364562	500000	97%
ENr223	ChrUn12	6975089	7201574	Random	226485		
ENr223	ChrB2	84494147	84637047	Random	142900	500000	100%
ENr231	ChrUn17	2738775	3158127	Random	419352	500000	100%
ENr233	ChrUn30	1720948	2055625	Random	334677	500000	99%
ENr311	ChrB3	103205560	103651920	Random	446360	500000	100%
ENr312	ChrUn5	2268912	2720625	Random	451713	500000	100%
ENr313	ChrE2	48380625	48804428	Random	423803	500000	100%
ENr321	ChrF2	60858485	61441113	Random	582628	500000	100%
ENr322	ChrB3	146890663	147512708	Random	622045	500000	99%
ENr324	ChrX	108254412	108742688	Random	488276	500000	98%
ENr331	ChrC1	208289082	208897593	Random	608511	500000	100%
ENr333	ChrA3	5135391	5798783	Random	663392	500000	100%

There are NISC clone sequence gaps in ENm008, ENm009 (three gaps), ENr123, and ENr324, which show up as missing alignments on the respective graphs above.
[a]Alignment of each ENCODE region was divided into 1-kb segments, and the orientation of these segments as mapped onto the cat assembly was evaluated for correct orientation.

mammals and two that are annotated as being specific to rodents, and several that have duplicated copies. A total of 201 feline homologs of human miRNAs are distributed across the assembly; 93 loci correspond to a single miRNA locus, while the others belong to 37 clusters (<10 kb apart) of multiple miRNA sequences. Seventeen of the 37 locus clusters have the same number of sequences as their human counterpart and are located in the homologous syntenic region, annotated in GARFIELD. Twenty locus clusters have a different number of included miRNA copies from that found in the human homologous locus. Table 7 summarizes 10 locus clusters that have three or more miRNAs in the human and cat genome, along with the disposition of the homologous locus cluster in the cat.

Nuclear mitochondrial (*numt*) sequences in cat

Eukaryotic genomes retain relict sequences of mitochondrial genes that were transposed to nuclear chromosomes in their ancestry (Richly and Leister 2004). The cat family has two well-characterized *numt* loci: (1) Lopez-*numt* in domestic cats, which comprises a 7.9-kb segment spanning the *<CR-12S-16S-ND1-ND2-CO1-CO2>* gene segments of mtDNA repeated in 38–76 tandem copies on chromosome D2 distinguished by a 10-bp deletion (Lopez et al. 1994, 1996); and (2) a recent 12.5-kb mtDNA transposition (~3.5 Mya) to chromosome F2 in the common ancestors of the great cats, genus *Panthera* (Kim et al. 2006).

A BLAST search comparing cat genome sequence to full-length cytoplasmic mtDNA (*cymt*) sequences (Lopez et al. 1996) yielded 489 sequence matches of 334 kb total, of which 36 kb (10.8%) was identical to *cymt*, leaving 298 kb of potential *numt* (covering 99% of the *cymt* sequence span). One-third of the cat's *numt* sequences (96 kb) corresponded to Lopez-*numt*. One large 78-kb scaffold (scaffold ID 112,167) showed 12 Lopez-*numt*s and likely represents the chromosome D2 *numt* tandem repeat. Twelve percent of the mtDNA sequence matches that were neither *cymt* nor Lopez-*numt* had nuclear DNA flanking regions (5′ or 3′) that allowed them to be mapped to specific chromosomes (Supplemental Fig. S2 shows the chromosome distribution of *numt*s). Phylogenetic analyses of homologous *numt*s suggest multiple *numt* historic insertions over time in the cat genome (Fig. 3). When *numt* segments that included the *ND1* gene (~570 bp) present in Lopez-*numt* were examined, a minimum of four distinct *numt* lineages were apparent, three of which predated the species divergence of the genus *Felis* (Fig. 3A). In addition, novel *numt* sequences that included mitochondrial genes not present in Lopez-*numt* (e.g., *CytB*; Fig. 3B) suggested three additional *numt* insertions that originated since the onset of the Felidae family radiations (Johnson et al. 2006).

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

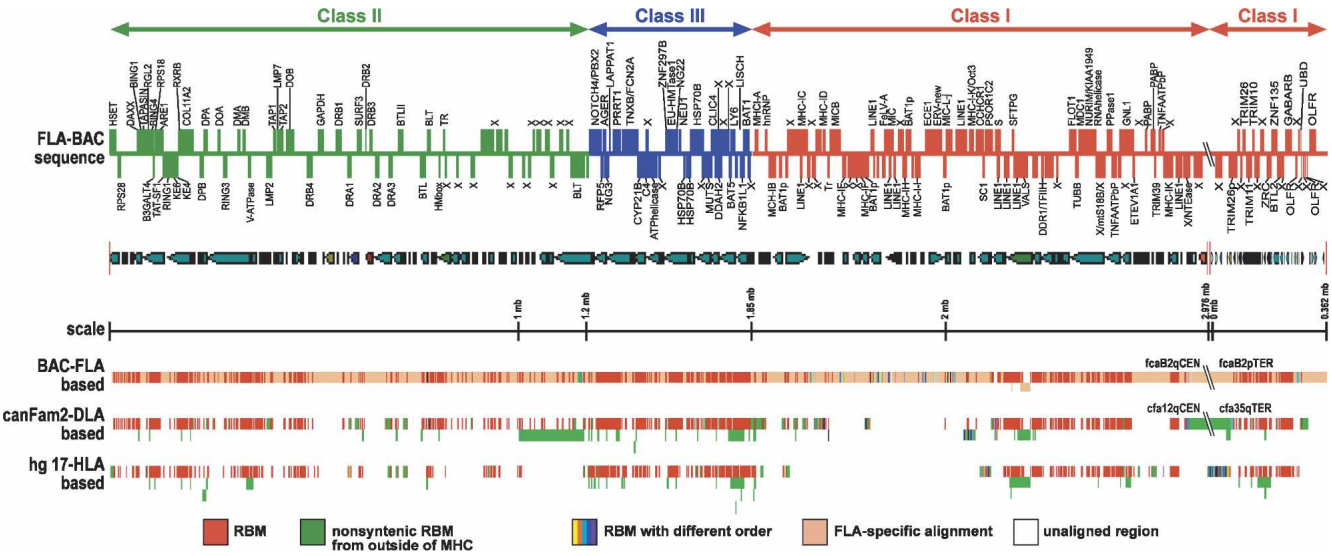

Figure 7. *FLA* annotation from finished BAC sequence (RPCI 86 BAC library constructed from the DNA of a domestic cat, Gus) and 1.9× WGS contigs alignment with three MHC models (Yuhki et al. 2003; Beck et al. 2005). Two segments of cat MHC (*FLA*) sequences—2.976-Mb B2cen MHC sequence and 0.0362-Mb B2ter MHC sequence—were assembled based on the human *HLA* sequence (The MHC Sequencing Consortium 1999). Gene annotation of this *FLA* sequence was performed using the GENSCAN (Burge and Karlin 1997) program. Genes with forward and reverse orientations were placed as solid blocks *above* and *below* the line, respectively. A gap position of these two *FLA* sequences is indicated as double slash lines. Contigs from the 1.9× WGS genome sequence were aligned using the Cross_match program with three models—*FLA*, *DLA* (canFam2), and *HLA* (hg17 assembly), respectively. These three sets of contig alignments were compared and highlighted with different colors with the following categories: (red) conserved sequence blocks with the same order between *FLA* and *DLA* shown on *FLA* and *DLA* lines or between *FLA* and *HLA* shown on the *HLA* line; (green) nonconserved sequences between *FLA/DLA* or *FLA/HLA* but a part of *FLA* sequences conserved in other genomic regions in dog and human and translocated to MHC regions in *DLA* or *HLA*; (multiple colors) conserved sequence blocks between *FLA/DLA* or *FLA/HLA* with different order; (pink) *FLA*-specific blocks; (white) no 2× WGS contigs were aligned in *DLA* or *HLA*.

Based on differences in mutational rates and differences in the genetic code between the mitochondrial and nuclear genomes, it would seem that the transpositions of *numt* sequence across the cat genome (and in other species) are likely vestigial and with no function. However, the origins of mitochondrial–nuclear complementation in energy metabolism (mitochondria contain 37 and nuclear chromosomes contain >1500 mitochondrial bound genes) clearly involved an early *numt* transposition, which by evolutionary processes would build the energetics pathway now operative in every eukaryotic species (Margulis 1970; Wallace 2005). The chromosomal disposition of feline *numt* and the multiple phylogenetic lineages reflect a more recent record of dynamic mitochondrial DNA transposition to disparate chromosomal positions in the ancestors of modern cats, providing yet another informative character for genomic inferences.

Feline endogenous retrovirus-like sequences (FERVs)

Domestic cats carry endogenous retroviral genomic sequences descended from ancestral infections and integrations into the germline. RD114 is an endogenous retrovirus related to baboon endogenous retrovirus with ~20 copies in cats (Benveniste et al. 1974; Reeves and O'Brien 1984; Coffin et al. 1997). Full-length RD114 virus is inducible with halogenated pyrimidines. Because RD114 is relatively innocuous but replicates well in human tissues, RD114-based vector constructs have been used extensively in human gene therapy applications (Relander et al. 2005; Ting-De Ravin et al. 2006).

Endogenous feline leukemia viruses (enFeLVs) are a second group with nine to 16 copies per cat genome, many of them truncated and insertionally polymorphic, with sequences related to exogenous feline leukemia virus (exFeLV) (Roca et al. 2005). By themselves, enFeLVs do not cause disease, but they recombine with exogenous viruses to create new virus subtypes that can augment the pathogenicity of exogenous FeLV (Roy-Burman 1995). In some cases, translation of partial enFeLV envelope protein may also protect against infection by some exFeLVs (McDougall et al. 1994). A phylogenetic analysis of endogenous FeLV LTR sequences from the cat genome sequence defined two distinct lineages, suggesting that at least two different germline FeLV infections occurred in the history of cats (Fig. 4A). Numerous RD114-related traces were also evident, consistent with the presence of both functional and truncated endogenous RD114 sequences (Supplemental Fig. S3b).

Approximately 4% of feline genome sequences are retrovirus-like sequences (Table 2). A homology search using 703 known retroviral sequences uncovered five new FERV lineages distinct from and more abundant than enFeLV and RD114 (Table 3). One major new group of retroviruses, FERV-1, included one locus embedded in the *FLA* BAC sequence, is related to porcine ERV (Fig. 4B), and is the most abundant of the novel retroviral elements. The other less abundant FERV sequences were related to human ERV lineages and to other mammalian retroviruses including mouse mammary tumor viruses. As with mouse and human, the cat genome is littered with relict FERV sequences descended from ancestral infections of virulent retroviruses (annotated in GARFIELD).

Single nucleotide polymorphisms (SNPs)

The contig sequence reads were examined to discover sites of nucleotide variation in cats or, more precisely, sites of heterozygous SNPs in Cinnamon. A total of 327,000 SNP variants were

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 5. Feline genetic diseases/phenotypes characterized at the molecular level

Symbol	Chromosome	Start	Phenotype	Mutation[a]	Reference
ARSB	A1	179,251,643	Mucopolysaccharoidosis Type VI (MPS VI)	L476P (severe phenotype) D520N (mild phenotype)	(Crawley et al. 1998)
HEXB	A1	177,250,358	Hexosaminidase B	39delC leads to premature stop or 1467_1491 inv; del exon12	(Martin et al. 2004; Muldoon et al. 1994)
GM2A	A1	231,863,263	Gangliosidosis GM2	Del4bp in 3′-region leads to frameshift	(Martin et al. 2005)
LIX1[b]	A1	195,928,460	Spinal muscular atrophy	~140-kb deletion	(Fyfe et al. 2006)
MAN2B1	A2	10,231,211	Alpha-mannosidosis	1749_1752delCCAG leads to premature stop	(Berg et al. 1997)
ASIP	A3	4,207,495	Melanism (domestic cat)	123_124 del CA leads to frameshift	(Eizirik et al. 2003)
LPL	B1	42,039,286	Hypertriglyceridemia (lipoprotein lipase deficiency)	G412R	(Ginzinger et al. 1996)
FGF5[b]	B1	158,450,684	Long hair	del474T, T159P, R136X, ins356T	(Drögemüller et al. 2007; Kehler et al. 2007)
CMAH[b]	B2	3,881,216	Blood group antigens	A217G, T371C	(Bighignoli et al. 2007)
CEP290[b]	B4	131,765,255	Retinitis pigmentosa (AR)	50 (IVS50 + 9T>G)	(Menotti-Raymond et al. 2007)
TAS1R2	C1	16,592,461	Sweet taste receptor	454_700del	(Li et al. 2005)
MLPH[b]	C1	224,776,040	Dilute	del83T leads to premature stop	(Ishida et al. 2006)
GLB1	C2	150,053,454	Gangliosidosis GM1 (Sandhoff disease)	R482P	(Baker et al. 2001)
GBE1	C2	36,408,947	Glycogenosis IV	Gene rearrangement with insertion and deletion in exon 12	(Fyfe et al. 2007)
TYR	D1	70,511,772	Albino	del975C leads to premature stop	(Imes et al. 2006)
			Siamese	G301R	(Schmidt-Küntzel et al. 2005)
			Oculocutaneous albinism (Type II)	R422Q	(Giebel et al. 1991)
			Burmese	G227W	(Lyons et al. 2005; Schmidt-Küntzel et al. 2005)
MYBPC3	D1	127,996,446	Hypertrophic cardiomyopathy	A31P	(Meurs et al. 2005)
NPC1	D3	64,204,013	Niemann-Pick disease, Type C	G2864C	(Somers et al. 2003)
TYRP1	D4	28,085,448	Brown	A3G and 421_422 ins 18AA/19AA	(Schmidt-Küntzel et al. 2005)
			Cinnamon	R100X	
MC1R	E2	74,966,297	Melanism (jaguar)	301_315del	(Eizirik et al. 2003)
			Melanism (jagaroundi)	283_306del	
GUSB	E3	16,412,143	Mucopolysaccharoidosis Type VII (MPS VII)	E351K	(Fyfe et al. 1999)
PKD1	E3	67,583,844	Polycystic kidney disease	C3284X	(Lyons et al. 2004)
PKLR	F1	88,296,247	Pyruvate kinase deficiency	Splicing defect leads to 13-bp deletion in exon 6	(Giger et al. 1997)
DMD	X	29,987,141	Muscular dystrophy, Duchenne type	Deletion in the dystrophin muscle promoter	(Winand et al. 1994)
F8	X	139,021,992	Hemophilia B	R338X, C82Y	(Goree et al. 2005)
IDUA	Un3	12,617,143	Mucopolysaccharidosis Type I (MPS I)	107_109 delCGA	(He et al. 1999)
GNPTAB	Un15	3,863,129	Mucolipidosis II (I- cell disease)	C2655T	(Giger et al. 2006)

[a]Mutation notation according to Den Dunnen and Antonarakis (2001).
[b]Mutation discovered using the 1.9× WGS sequence.

detected, submitted to dbSNP, and are annotated in GARFIELD. To verify SNP recognition, a random sampling of 200 SNPs were selected for re-sequencing. Ninety percent (180 SNPs) sequenced well, and of these 91% were validated SNP heterozygotes. The genome of Cinnamon was separated into segments covering 43% of the genome that contains multiple heterozygous SNP loci, while the remaining 57% contains long homozygous segments (Fig. 5). Within the heterozygous segments, the SNP loci incidence was 1/600 (0.00167), while the size of homozygous segments in Cinnamon varied from <10 kb to >4.0 Mb with a median length of 170 kb (N50 ≥ 60 kb). The long stretches of alternating homozygous and heterozygous segments are likely a consequence of the domestication process, close inbreeding during Abyssinian breed development, and disease pedigree establishment. Homozygous segments also occur in dogs, possibly for the same reasons (Lindblad-Toh et al. 2005). The *FLA* region sequence (Fig. 7) derived from Cinnamon was largely homozygous for SNP variants. However, comparison of the Cinnamon sequence to the *FLA* BACs from a different cat revealed 11,654 SNPs (873 coding) (Yuhki et al. 2003; Beck et al. 2005). The SNP loci incidence of this region ($f = 0.00391$, or 1/256) is comparable to that of *HLA* ($f = 0.00349$, or 1/286).

To explore the breed-specific patterns of common segment homozygosity, as well as to estimate the size of linkage disequilibrium stretches in cat breeds, a group of 350 SNPs were genotyped in multiple individuals from each of 24 certified cat breeds. Briefly, 35 SNPs were selected across 10 ~600-kb highly heterozygous segments (in Cinnamon) from different cat autosomes. For each region, eight SNPs fell within 15 kb of each other, while an additional 27 SNPs were added at 20-kb intervals to fill out the

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

Table 6. Annotated features from mammalian genomes and their representation in the *Felis catus* assembly

Feature	Human	Chimpanzee	Mouse	Rat	Cow	Dog
Gene (exons + introns)	34.5 ± 25.4 22073	31.7 ± 25.3 21465	15.5 ± 22.7 31093	18.1 ± 23.9 22573	38.0 ± 28.2 22782	49.4 ± 27.2 19756
Coding sequence (CDS)	58.6 ± 33.9 26365	54.0 ± 33.9 21452	34.4 ± 36.1 37205	39.6 ± 34.8 22565	61.9 ± 32.6 35868	64.8 ± 30.7 733624
UTR5	44.6 ± 45.6 18120	48.5 ± 46.3 4963	28.6 ± 41.7 19496	32.0 ± 43.8 7738	57.1 ± 45.7 14173	56.0 ± 46.0 12944
UTR3	56.7 ± 42.5 23082	52.2 ± 46.9 6464	35.6 ± 42.0 25278	43.0 ± 45.3 11892	66.5 ± 41.2 23117	69.0 ± 41.7 20112
CDS + UTR	55.7 ± 31.6 26382	53.4 ± 32.8 21452	30.9 ± 33.0 37294	38.3 ± 33.6 22581	61.1 ± 31.5 35889	64.6 ± 30.0 33642
Intergenic	16.6 ± 16.7 15594	17.3 ± 17.0 13450	4.5 ± 8.3 18228	4.1 ± 8.1 16040	19.6 ± 18.8 11763	37.4 ± 21.6 11962
Intron	32.3 ± 24.0 24841	29.1 ± 24.0 20366	13.0 ± 19.1 31736	14.1 ± 19.5 19870	39.6 ± 26.4 33449	50.8 ± 24.3 31725
Downstream	29.8 ± 27.5 22066	31.1 ± 27.6 21306	10.7 ± 18.8 31043	13.8 ± 20.8 22561	32.8 ± 28.7 22483	50.3 ± 28.8 19749
Upstream	26.4 ± 25.1 22066	25.5 ± 25.1 21310	9.0 ± 16.5 31016	10.0 ± 17.3 22559	27.8 ± 26.4 22507	44.1 ± 27.3 19750

For each annotated feature from each genome, the percent of its nucleotides that are included in reciprocal best match alignments to the cat WGS sequences was calculated. Provided in the table is, for each genome and feature type, the mean and standard deviation of the percent coverage of the feature type, with the total number of features.

600 kb. The average homozygosity for SNP loci across the 10 10-kb regions was ~53%. Conditional on homozygosity within the first 10-kb window, the extent of homozygosity was recorded, and the fraction of loci that remained homozygous at different distances was plotted (Supplemental Fig. S8). The fraction of homozygous loci decays as a function of physical distance roughly threefold faster in cats than in dogs (Lindblad-Toh et al. 2005). This may reflect more recent inbreeding and/or restricted gene flow between dog breeds than for cat breeds resulting in shorter haplotypes and linkage disequilibrium in cats. A rough estimate (based on genotyping two to three individuals per breed) would suggest approximately three to five haplotypes per breed within 10-kb and 100-kb windows, very similar to that seen for dog breeds. The extent of homozygosity together with haplotype diversity can be used to infer the number of equally spaced SNPs required for genomewide association mapping within a specific breed. Since ~15,000 SNPs are required for mapping in dogs, we estimate that ~45,000 equivalently spaced SNPs (three times the number for dogs) would be appropriate in cat breeds.

Conclusions and applications of the feline genome sequence

The feline genome sequence, here annotated, has immediate value in many aspects of biology, particularly in the discovery of the genetic basis of hereditary and infectious diseases (Table 5; Supplemental Table S1). Other areas to benefit include comparative genomics for which mammalian CSBs, representing both long genes and short conserved elements, provide the means of reconstructing chromosome exchanges that punctuate mammal evolution. Feline models of hereditary disease/phenotypes are

Table 7. MiRNA clusters with three or more members shared by human and cat

MiRNA cluster	No. of members in the human genome	Chromosomal location in the human genome	Chromosomal location in the cat assembly	Absent in cat assembly	Absent in human assembly[a]
MIRN25	3	7:99335834–99336348	E3:6238059–6238571		
MIRN17	6	13:90800860–90801646	A1:65582233–65583016		
MIRN23B	3	9:94927045–94927925	Un1:6010466–6010021		
MIRN127	7	14:100405150–100420873	B3:148975105–148992826	*MIRN431*	
MIRN424	5	X:133399891–133406261	X:119173069–119174202	*MIRN503* *MIRN424*	*MIRN450-1*
MIRN188	5	X:49471145–49482327	X:48518861–48524556	*MIRN188*	
MIRN18b	6	X:133028928–133029828	X:118848137–118849041		
MIRN134	28	14:100558156–100602081	B3:149152984–149202131	*MIRN329-1* *MIRN409* *MIRN412* *MIRN369* *MIRN410*	*MIRN411* *MIRN300*
MIRNLET7A1	3	9:94017794–94020757	Un1:13102544–13104533	*MIRNLET7A1*	
MIRNLET7E	3	19:56887677–56888404	E2:6534658–6534747	*MIRN99B* *MIRN125a*	

Each cluster is identified by one of its members.
[a]Human assembly annotation taken from Micro-RNA Registry Version 8.0. These sequences have subsequently been annotated as being in human (Micro-RNA Registry Version 10.0).

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

already being uncovered using the genome sequence annotated here. Forensic evidence from cats, already established in legal precedent (Menotti-Raymond et al. 1997a,b, 2005), can now be further characterized in terms of the additional SNP and STR variants made available here. Cat models of emerging infectious agents can now be approached in the context of host genetic variation in immune response, such as the cat *FLA* complex studied here. Finally, the genome sequence and variation shared with other felid species can increase natural history studies of free-ranging cat species for conservation and management purposes (O'Brien and Johnson 2005).

In spite of the benefits derived from the comparative genomics-based genome annotation presented here, there are some notable weaknesses due to a light coverage. Among them are the following: (1) The assembled cat genome retains only 65% of the euchromatin genome sequence, leaving some 660,000 gaps between the contigs; (2) fewer than 58% of the genes have >50% of their gene feature sequence captured (based on cat–dog gene homologs); and (3) estimating the number, extent, and location of segmental duplications (which comprise 5% of the human genome) is difficult with low coverage since segmental duplication discovery depends on highly redundant genome coverage for accuracy (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002).

These limitations notwithstanding, our analysis of the cat genome sequence in a comparative context has allowed an examination of genome structure and features, genome evolution, and useful applications for comparative genomics and cat biology. The cat genome annotation has increased the depth of evolutionary perspective required for comparative inference. We anticipate that genome annotation of additional species will reveal the cryptic process of species differentiation, development and adaptation. The approach used here could hopefully be applied to the other mammalian species scheduled for 2× sequence coverage (Supplemental Fig. S1); however, the availability of the 1680-marker cat RH map, the BAC and fosmid libraries, breed populations, linkage map, and gene discovery analyses has aptly complemented the cat genome annotation exercise.

Acknowledgments

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 U.S. Government. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. This research was supported in part by the Intramural Research Program of the National Institutes of Health, NCI, NHGRI, and NLM. G.T. was funded by a Sloan Research Fellowship in Molecular Biology. G.B. was supported by the Agency for Science Technology and Research (A*STAR), Singapore. W.J.M. was funded by the Winn Feline Foundation. U.G. was funded by NIH Grant NIH RR 02512.

References

Baker, H., Smith, B.R., Martin, D.R., and Foureman, P. 2001. *Consultations in feline internal medicine*. W.B. Saunders & Co., Philadelphia.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12:** 177–189.

Beck, T.W., Menninger, J., Murphy, W.J., Nash, W.G., O'Brien, S.J., and Yuhki, N. 2005. The feline major histocompatibility complex is rearranged by an inversion with a breakpoint in the distal class I region. *Immunogenetics* **56:** 702–709.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Benveniste, R.E., Lieber, M.M., Livingston, D.M., Sherr, C.J., Todaro, G.J., and Kalter, S.S. 1974. Infectious C-type virus isolated from a baboon placenta. *Nature* **248:** 17–20.

Berg, T., Tollersrud, O.K., Walkley, S.U., Siegel, D., and Nilssen, Ø. 1997. Purification of feline lysosomal α-mannosidase, determination of its cDNA sequence and identification of a mutation causing α-mannosidosis in Persian cats. *Biochem. J.* **328:** 863–870.

Bighignoli, B., Niini, T., Grahn, R.A., Pedersen, N.C., Millon, L.V., Polli, M., Longeri, M., and Lyons, L.A. 2007. Cytidine monophospho-N-acetylneuraminic acid hydroxylase (*CMAH*) mutations associated with the domestic cat AB blood group. *BMC Genet.* **8:** 27. doi: 10.1186/1471-2156-8-27.

Bourque, G. and Pevzner, P.A. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.* **12:** 26–36.

Bourque, G., Pevzner, P.A., and Tesler, G. 2004. Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Res.* **14:** 507–516.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Coffin, J., Hughes, S., and Varmus, H. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Crawley, A.C., Yogalingam, G., Muller, V.J., and Hopwood, J.J. 1998. Two mutations within a feline mucopolysaccharidosis type VI colony cause three different clinical phenotypes. *J. Clin. Invest.* **101:** 109–119.

den Dunnen, J.T. and Antonarakis, S.E. 2001. Nomenclature for the description of human sequence variations. *Hum. Genet.* **109:** 121–124.

Drögemüller, C., Rüfenacht, S., Wichert, B., and Leeb, T. 2007. Mutations within the *FGF5* gene are associated with hair length in cats. *Anim. Genet.* **38:** 218–221.

Eizirik, E., Yuhki, N., Johnson, W.E., Menotti-Raymond, M., Hannah, S.S., and O'Brien, S.J. 2003. Molecular genetics and evolution of melanism in the cat family. *Curr. Biol.* **13:** 448–453.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

Everts-van der Wind, A., Larkin, D.M., Green, C.A., Elliott, J.S., Olmstead, C.A., Chiu, R., Schein, J.E., Marra, M.A., Womack, J.E., and Lewin, H.A. 2005. A high-resolution whole-genome cattle–human comparative map reveals details of mammalian chromosome evolution. *Proc. Natl. Acad. Sci.* **102:** 18526–18531.

Fanning, T.G. 1987. Origin and evolution of a major feline satellite DNA. *J. Mol. Biol.* **197:** 627–634.

Fanning, T.G., Modi, W.S., Wayne, R.K., and O'Brien, S.J. 1988. Evolution of heterochromatin-associated satellite DNA loci in felids and canids (Carnivora). *Cytogenet. Cell Genet.* **48:** 214–219.

The FANTOM Consortium and the RIKEN Exploration Research Group Phase I and II. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Fyfe, J.C., Giger, U., Van Winkle, T.J., Haskins, M.E., Steinberg, S.A., Wang, P., and Patterson, D.F. 1992. Glycogen storage disease type IV: Inherited deficiency of branching enzyme activity in cats. *Pediatr. Res.* **32:** 719–725.

Fyfe, J.C., Kurzhals, R.L., Lassaline, M.E., Henthorn, P.S., Alur, P.R.K., Wang, P., Wolfe, J.H., Giger, U., Haskins, M.E., Patterson, D.F., et al. 1999. Molecular basis of feline β-glucuronidase deficiency: An animal model of mucopolysaccharidosis VII. *Genomics* **58:** 121–128.

Fyfe, J.C., Menotti-Raymond, M., David, V.A., Brichta, L., Schäffer, A.A., Agarwala, R., Murphy, W.J., Wedemeyer, W.J., Gregory, B.L., Buzzell, B.G., et al. 2006. An ~140-kb deletion associated with feline spinal muscular atrophy implies an essential *LIX1* function for motor neuron survival. *Genome Res.* **16:** 1084–1090.

Fyfe, J.C., Kurzhals, R.L., Hawkins, M.G., Wang, P., Yuhki, N., Giger, U., Van Winkle, T.J., Haskins, M.E., Patterson, D.F., and Henthorn, P.S. 2007. A complex rearrangement in *GBE1* causes both perinatal hypoglycemic collapse and late-juvenile-onset neuromuscular degeneration in glycogen storage disease type IV of Norwegian forest

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

cats. *Mol. Genet. Metab.* **90:** 383–392.
Giebel, L.B., Tripathi, R.K., King, R.A., and Spritz, R.A. 1991. A tyrosinase gene missense mutation in temperature-sensitive type I oculocutaneous albinism. A human homologue to the Siamese cat and the Himalayan mouse. *J. Clin. Invest.* **87:** 1119–1122.
Giger, U., Rajpurohit, Y., Wang, P., Ford, S., Kohn, B., Patterson, D.F., Beutler, E., and Henthorn, P.S. 1997. Molecular basis of erythrocyte pyruvate kinase (R-PK) deficiency in cats. *Blood* **90S:** 5.
Giger, U., Tcherneva, E., Caverly, J., Seng, A., Huff, A.M., Cullen, K., Van Hoeven, M., Mazrier, H., and Haskins, M.E. 2006. A missense point mutation in N-acetylglucosamine-1-phosphotransferase causes mucolipidosis II in domestic shorthair cats. *J. Vet. Intern. Med.* **20:** 781.
Ginzinger, D.G., Lewis, M.E.S., Ma, Y., Jones, B.R., Liu, G., and Jones, S.D. 1996. A mutation in the lipoprotein lipase gene is the molecular basis of chylomicronemia in a colony of domestic cats. *J. Clin. Invest.* **97:** 1257–1266.
Goree, M., Catalfamo, J.L., Aber, S., and Boudreaux, M.K. 2005. Characterization of the mutations causing hemophilia B in 2 domestic cats. *J. Vet. Intern. Med.* **19:** 200–204.
Griffin, B. and Baker, H.J. 2000. *Domestic cats as laboratory animals*. Academic Press, San Diego.
Griffith-Jones, S. 2004. The micro-RNA Registry. *Nucleic Acids Res.* **32:** D109–D111. doi: 10.1093/nar/gkh023.
Guigó, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., et al. 2006. EGASP: The human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7:** S2.1–S2.31. doi: 10.1186/gb-2006-7-s1-s2.
He, X., Li, C.-M., Simonaro, C.M., Wan, Q., Haskins, M.E., Desnick, R.J., and Schuchman, E.H. 1999. Identification and characterization of the molecular lesion causing mucopolysaccharidosis type I in cats. *Mol. Genet. Metab.* **67:** 106–112.
Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I.L., and Stadler, P.F. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7:** 25. doi: 10.1186/1471-2164-7-25.
Imes, D.L., Geary, L.A., Grahn, R.A., and Lyons, L.A. 2006. Albinism in the domestic cat (*Felis catus*) is associated with a tyrosinase (*TYR*) mutation. *Anim. Genet.* **37:** 175–178.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Ishida, Y., David, V.A., Eizirik, E., Schäffer, A.A., Neelam, B.A., Roelke, M.E., Hannah, S.S., O'Brien, S.J., and Menotti-Raymond, M. 2006. A homozygous single-base deletion in *MLPH* causes the *dilute* coat color phenotype in the domestic cat. *Genomics* **88:** 698–705.
Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13:** 91–96.
Johnson, W.E., Eizirik, E., Pecon-Slattery, J., Murphy, W.J., Antunes, A., Teeling, E., and O'Brien, S.J. 2006. The late Miocene radiation of modern Felidae: A genetic assessment. *Science* **311:** 73–77.
Kehler, J.S., David, V.A., Schäffer, A.A., Bajema, K., Eizirik, E., Ryugo, D.K., Hannah, S.S., O'Brien, S.J., and Menotti-Raymond, M. 2007. Four independent mutations in the *Feline Fibroblast Growth Factor 5* gene determine the long-haired phenotype in domestic cats. *J. Hered.* **98:** 555–566.
Kim, J.-H., Antunes, A., Luo, S.-J., Menninger, J., Nash, W.G., O'Brien, S.J., and Johnson, W.E. 2006. Evolutionary analysis of a large mtDNA translocation (*numt*) into the nuclear genome of the *Panthera* genus species. *Gene* **366:** 292–302.
Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., et al. 2003. The dog genome: Survey sequencing and comparative analysis. *Science* **301:** 1898–1903.
Kuiken, T., Rimmelzwaan, G., van Riel, D., van Amerongen, G., Baars, M., Fouchier, R., and Osterhaus, A. 2004. Avian H5N1 influenza in cats. *Science* **306:** 241.
Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.
Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. 2003. New microRNAs from mouse and human. *RNA* **9:** 175–179.
Lecellier, C.-H., Dunoyer, P., Arar, K., Lehmann-Che, J., Eyquem, S., Himber, C., Saib, A., and Voinnet, O. 2005. A cellular microRNA mediates antiviral defense in human cells. *Science* **308:** 557–560.
Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21:** 4663–4670.
Li, X., Li, W., Wang, H., Cao, J., Maehashi, K., Huang, L., Bachmanov, A.A., Reed, D.R., Lengrand-Defretin, V., Beauchamp, G.K., et al. 2005. Pseudogenization of a sweet-receptor gene accounts for cats' indifference toward sugar. *PLoS Genet.* **1:** e3. doi: 10.1371/journal.pgen.0010003.
Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas III, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.
Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S.J. 1994. *Numt,* a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* **39:** 174–190.
Lopez, J.V., Cevario, S., and O'Brien, S.J. 1996. Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and a transposed mtDNA tandem repeat (*Numt*) in the nuclear genome. *Genomics* **33:** 229–246.
Lyons, L.A., Biller, D.S., Erdman, C.A., Lipinski, M.J., Young, A.E., Roe, B.A., Qin, B., and Grahn, R.A. 2004. Feline polycystic kidney disease mutation identified in PKD1. *J. Am. Soc. Nephrol.* **15:** 2548–2555.
Lyons, L.A., Imes, D.L., Rah, H.C., and Grahn, R.A. 2005. Tyrosinase mutations associated with Siamese and Burmese patterns in the domestic cat (*Felis catus*). *Anim. Genet.* **36:** 119–126.
Margulis, L. 1970. *Origin of eukaryotic cells*. Yale University Press, New Haven CT.
Martin, D.R., Krum, B.K., Varadarajan, G.S., Hathcock, T.L., Smith, B.F., and Baker, H.J. 2004. An inversion of 25 base pairs causes feline G_{M2} gangliosidosis variant. *Exp. Neurol.* **187:** 30–37.
Martin, D.R., Cox, N.R., Morrison, N.E., Kennamer, D.M., Peck, S.L., Dodson, A.N., Gentry, A.S., Griffin, B., Rolsma, M.D., and Baker, H.J. 2005. Mutation of the G_{M2} activator protein in a feline model of G_{M2} gangliosidosis. *Acta Neuropathol. (Berl.)* **110:** 443–450.
McDougall, A.S., Terry, A., Tzavaras, T., Cheney, C., Rojko, J., and Neil, J.C. 1994. Defective endogenous proviruses are expressed in feline lymphoid cells: Evidence for a role in natural resistance to subgroup B feline leukemia viruses. *J. Virol.* **68:** 2151–2160.
Menotti-Raymond, M., David, V.A., Stephens, J.C., Lyons, L.A., and O'Brien, S.J. 1997a. Genetic individualization of domestic cats using feline STR loci for forensic applications. *J. Forensic Sci.* **42:** 1039–1051.
Menotti-Raymond, M.A., David, V.A., and O'Brien, S.J. 1997b. Pet cat hair implicates murder suspect. *Nature* **386:** 774.
Menotti-Raymond, M., David, V.A., Wachter, L.L., Butler, J.M., and O'Brien, S.J. 2005. A forensic typing system of STR loci for genetic individualizationi of domestic cat (*Felis catus*) samples. *J. Forensic Sci.* **50:** 1061–1070.
Menotti-Raymond, M., David, V.A., Schäffer, A.A., Stephens, R., Wells, D., Kumar-Singh, R., O'Brien, S.J., and Narfström, K. 2007. Mutation in *CEP290* discovered for cat model of human retinal degeneration. *J. Hered.* **98:** 211–220.
Meurs, K.M., Sanchez, X., David, R.M., Bowles, N.E., Towbin, J.A., Reiser, P.J., Kittleson, J.A., Munro, M.J., Dryburgh, K., MacDonald, K.A., et al. 2005. A cardiac myosin binding protein C mutation in the Maine Coon cat with familial hypertrophic cardiomyopathy. *Hum. Mol. Genet.* **14:** 3587–3593.
The MHC Sequencing Consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401:** 921–923.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Muldoon, L.L., Neuwelt, E.A., Pagel, M.A., and Weiss, D.L. 1994. Characterization of the molecular defect in a feline model for type II G_{M2}-gangliosidosis (Sandhoff disease). *Am. J. Pathol.* **144:** 1109–1118.
Mullikin, J.C. and Ning, Z. 2003. The Phusion assembler. *Genome Res.* **13:** 81–90.
Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., De Jong, P., et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294:** 2348–2351.
Murphy, W.J., Pevzner, P.A., and O'Brien, S.J. 2004. Mammalian phylogenomics comes of age. *Trends Genet.* **20:** 631–639.
Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Getzke, L., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309:** 613–617.
Murphy, W.J., Davis, B., David, V.A., Agarwala, R., Schäffer, A.A., Pearks Wilkerson, A.J., Neelam, B., O'Brien, S.J., and 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.
Nowell, K. and Jackson, P. 1996. *Status survey and conservation action*

Downloaded from www.genome.org on November 29, 2007 - Published by Cold Spring Harbor Laboratory Press

plan, wild cats. International Union for Conservation of Nature and Natural Resources, Gland, Switzerland.
O'Brien, S.J. 2004. Cats. *Curr. Biol.* **14:** R988–R989.
O'Brien, S.J. and Johnson, W.E. 2005. Big cat genomics. *Annu. Rev. Genomics Hum. Genet.* **6:** 407–429.
O'Brien, S.J. and Murphy, W.J. 2003. Genomics. A dog's breakfast? *Science* **301:** 1854–1855.
O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Marshall Graves, J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–481.
O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., and Yuhki, N. 2002. The Feline Genome Project. *Annu. Rev. Genet.* **36:** 657–686.
Pecon-Slattery, J., Sanner-Wachter, L., and O'Brien, S.J. 2000. Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). *Proc. Natl. Acad. Sci.* **97:** 5307–5312.
Pecon-Slattery, J., Pearks Wilkerson, A.J., Murphy, W.J., and O'Brien, S.J. 2004. Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family Felidae as a species tree. *Mol. Biol. Evol.* **21:** 2299–2309.
Pevzner, P.A. and Tesler, G. 2003. Genome rearrangements in mammalian evolution: Lessons from human and mouse genome sequences. *Genome Res.* **13:** 13–26.
Pontius, J.U. and O'Brien, S.J. 2007. Genome Annotation Resource Fields—GARFIELD: a genome browser for *Felis catus*. *J. Hered.* **98:** 386–389.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.
Reeves, R.H. and O'Brien, S.J. 1984. Molecular genetic characterization of the RD-114 gene family of endogenous feline retroviral sequences. *J. Virol.* **52:** 164–171.
Relander, T., Johansson, M., Olsson, K., Ikeda, Y., Takeuchi, Y., Collins, M., and Richter, J. 2005. Gene transfer to repopulating human CD34$^+$ cells using amphotropic-, GALV-, or RD114-pseudotyped HIV-1-based vectors from stable producer cells. *Mol. Ther.* **11:** 452–459.
Richly, E. and Leister, D. 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **21:** 1081–1084.
Roca, A.L., Nash, W.G., Menninger, J.C., Murphy, W.J., and O'Brien, S.J. 2005. Insertional polymorphisms of endogenous feline leukemia viruses. *J. Virol.* **79:** 3979–3986.
Roy-Burman, P. 1995. Endogenous *env* elements: Partners in generation of pathogenic feline leukemia viruses. *Virus Genes* **11:** 147–161.
Schmidt-Küntzel, A., Eizirik, E., O'Brien, S.J., and Menotti-Raymond, M. 2005. *Tyrosinase* and *tyrosinase related protein 1* alleles specify domestic cat coat color phenotypes of the *albino* and *brown* loci. *J. Hered.* **96:** 289–301.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Smit, A.F.A. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.
Somers, K.L., Royals, M.A., Carstea, E.D., Rafi, M.A., Wenger, D.A., and Thrall, M.A. 2003. Mutation analysis of feline Niemann-Pick C1 disease. *Mol. Genet. Metab.* **79:** 99–103.
Springer, M.S., Murphy, W.J., Eizirik, E., and O'Brien, S.J. 2003. Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl. Acad. Sci.* **100:** 1056–1061.
Tesler, G. 2002a. GRIMM: Genome rearrangements web server. *Bioinformatics* **18:** 492–493.
Tesler, G. 2002b. Efficient algorithms for multichromosomal genome rearrangements. *J. Comp. Sys. Sci.* **65:** 587–609.
Ting-De Ravin, S.S., Kennedy, D.R., Naumann, N., Kennedy, J.S., Choi, U., Hartnett, B.J., Linton, G.F., Whiting-Theobald, N.L., Moore, P.F., Vernau, W., et al. 2006. Correction of canine X-linked severe combined immunodeficiency by in vivo retroviral gene therapy. *Blood* **107:** 3091–3097.
Vassetzky, N.S. and Kramerov, D.A. 2002. CAN—A pan-carnivore SINE family. *Mamm. Genome* **13:** 50–57.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Wallace, D.C. 2005. The mitochondrial genome in human adaptive radiation and disease: On the road to therapeutics and performance enhancement. *Gene* **354:** 169–180.
Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., et al. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **33:** D39–D45. doi: 10.1093/nar/gki062.
Winand, N.J., Edwards, M., Pradhan, D., Berian, C.A., and Cooper, B.J. 1994. Deletion of the dystrophin muscle promoter in feline muscular dystrophy. *Neuromuscul. Disord.* **4:** 433–445.
Yuhki, N., Beck, T., Stephens, R.M., Nishigaki, Y., Newmann, K., and O'Brien, S.J. 2003. Comparative genome organization of human, murine, and feline MHC class II region. *Genome Res.* **13:** 1169–1179.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.

Received February 18, 2007; accepted in revised form June 25, 2007.