**Supplemental Methods**

## SEQUENCING DATA AND ASSEMBLY

The genome of a female Abyssinian cat was sequenced using the whole-genome shotgun (WGS) approach by Agencourt Bioscience Corp. A total of ~8.02 million successful paired end reads were generated from two different clone types; 3.1 kb plasmids and 37 kb fosmids, both prepared from primary blood lymphocyte DNA. The sequencing was carried out according to standard methods using Solid Phase Reversible Immobilization (SPRI) based DNA purification and sequencing with Big Dye terminator reagents on ABI3730xl instruments. The genome size was initially estimated to be ~2.5 Gb, by placing paired end fosmid reads on the human genome and calculating the size ratio using the well defined size distribution of fosmid libraries. Subsequent mapping of the assembled contigs onto the human and dog genomes yielded a more accurate estimate of ~2.7 Gb (apparently due to a higher number segmental duplications in the cat genome). The final sequence coverage generated, based on the 2.7 Gb size estimate, is 1.9x. (see Table S5).

A two-step approach for genome assembly was used for assembly at the Broad Institute. First, a de novo assembly was generated using Arachne v2 (Jaffe et al. 2003). Second, an assisted assembly algorithm was used to iteratively improve the assembly by exploiting synteny between cat and two reference genomes: human and dog (the human genome build NCBI 35 (UCSC hg 17), and the canine genome NCBI 2 (CanFam2), respectively).

Assisting an assembly against a reference genome consists of aligning the whole genome shotgun sequencing (WGS) reads of the genome against the reference genome, and using the resulting alignments to validate both read-read alignments and linkage information in the initial de novo assembly. In particular, with low coverage genomes, multiple single links exist that are not sufficient to form scaffolds within the de novo assembly, but the supporting information from the assisted stage will validate some of these.

Read alignments against the reference genomes were generated using BLASTZ, followed by the S1-S2 test to filter the alignments (Margulies et al. 2005). The fraction of uniquely placing reads (74.5% placed on the dog genome, 61.5% placed on the human genome and 80.1% of reads placed on at least one of the genomes) performed as expected based on the methodology used and the evolutionary distance from cat to dog and human.

The resulting alignments were used in three different ways:

1. to confirm read-read alignments, both to extend existing contigs and to create new contigs. This substantially improved the total contig length, and the N50 contig length of the resulting assembly by increasing the number of reads used in the assembly.

2. to confirm single links, by verifying that the two end reads from the same insert mapped consistently and uniquely onto the reference genome. This allowed us to merge scaffolds joined by validated single links, resulting in improved contiguity.
3. to detect and break weak points in the assembly by syntenically mapping whole scaffolds to the reference genome, using the assembled reads of inserts that had been aligned consistently and uniquely onto the reference genome ("staples"). Illogical mapping staples onto the reference genome highlighted potential sites of assembly error, allowing for identification of weak joins which were accordingly broken.

This resulted in a more accurate final assembly. The final assisted Felis catus genome assembly consists of 217,790 scaffolds with N50 contig size of 2.4 kb and N50 scaffold size of 117 kb (Table S6). The sum of the spanning intervals of all scaffolds gives a total of 4.04 Gb of sequence. The excess sequence relative to the estimated 2.5 Gb genome size is due to the scaffold structure whereby many scaffolds interleaf with other scaffolds. The interleaving scaffolds are flattened onto a common reference genome at the mapping stage described in the following section.

## MAPPING THE ASSSEMBLED CONTIGS TO THE CAT GENOME

Using the scaffold structure defined by the assisted Arachne cat assembly, scaffold positions (x coordinate) for each contig were paired to a dog chromosome and position (y coordinate). For each scaffold's (x,y) coordinate lists, a linear regression was fit to the data, ignoring outliers greater than 40 kb outside the regression line. For scaffolds that mapped to more than one chromosome, a linear regression was calculated for each chromosome. For scaffolds that mapped to multiple chromosomes, scaffolds were broken if the number of contigs mapping to second best chromosome total more than two contigs. For all contigs within a scaffold that were unmapped (those that do not have a position on dog), or fall outside of 40 kb from their interpolated position, the interpolated position based on the linear regression was used as the position on dog. After these steps, contigs have a position on dog unless none of the contigs within a scaffold were mapped to a dog position, in which case, the scaffolds were assigned to chrUn.

From the RH map (Murphy et al. 2007) we deduced which segments of the dog chromosomes mapped to cat chromosomes and in what orientation. For each mapped segment, contigs were placed in the same order and orientation and relative position as found along the dog chromosome, with proper accounting for cases in which the dog segment is reverse complemented relative to the cat chromosome. For scaffolds that extended beyond the end of dog segment, these scaffolds were allowed to extend the reach of the mapped positions along a scaffold. Interleaved scaffolds from an anchored scaffold also extended the reach of the mappings. Contigs were not placed fewer than 10 bases apart. The number of nucleotides (Ns) between mapped segments was determined by where the last contig from the previous segment ended and the scaled map position of the start of the next segment. The scaled map position used the estimated cat genome size of 2.7Gb. Since RH map positions were in places distorted relative to the chromosomal base coordinate, some regions may be excessively expanded or contracted leaving gaps between mapped segments too large, or in other places, too small such that the estimated chromosomal base position of a segment would start before the end of the previous segment (a negative gap). If the initial estimate of the gap size between two segments was negative, then the gap size was set to 10,000 Ns. Examples of these extremes are chromosome E1 which is exaggerated in length and chromosome B2's centromere which was compressed to a 40 kb gap. For each segment of dog that contains unplaced cat sequence, an unmapped chromosome called chrUnN, where N is the dog chromosome identifier, was filled with these remaining pieces. Analysis of this chromosomally mapped cat genome shows that of the assembled bases, 83% are place on mapped chromosomes, although each chromosome consists on average of about 45% undetermined euchromatic nucleotides.

## MAMMALIAN GENOMES AND GENES

The gene annotation part of this study made use of MYSQL, perl modules from cpan (www.cpan.org) and the Eutil function from NCBI written by Oleg Khovayko (www.olegh.spedia.net) Extensive computing was done at the Advanced Biomedical Computing Center at NCI Frederick.
The mammalian genomes compared to the cat assembly were those from NCBI: Homo sapiens (Build 35, reference assembly), Pan troglodytes (Build 2, Arachne assembly), Mus musculus (Build 35, C57BL/6J assembly), Rattus norvegicus (Build 3, RGSC_v3.4 assembly), Canis familiaris (Build 2, Dog2.0 assembly) and Bos taurus (Build 2, Btau_2.0 assembly). The gene and homology annotation of each of these genomes were taken from the Genes and HomoloGene databases ( Build 48.1, ftp://ftp.ncbi.nih.gov/pub/HomoloGene/) at NCBI (Wheeler et al. 2005).

## MASKING OF REPETITIVE ELEMENTS

RepeatMasker (www.repeatmasker.org) , version open-3.1.0 , sensitive mode run with Cross_match version 0.990329, RepBase Update 10.04, RM database version 20050523) was used to soft-mask repetitive sequences of the cat WGS sequences, converting the masked base to lower case, as opposed to replacing the base with 'N'.

## ALIGNMENT ALGORITHM

The MegaBLAST alignment algorithm (Zhang et al. 2000) was used to align the cat and human genome to the other mammalian genomes. MegaBLAST arguments used (-D 3 -m 8 -s 100 -r 1 -q -1 -X 40 -W 16 -U T -F "m D" ) require an exact match between the two genomes of at least 16 bp in the unmasked portion of the genome, and allow extension of the alignment through masked regions. Subsequent filtering excluded alignments with a bitscore less than 200.

## RECIPROCAL BEST MATCHES

Using the alignment algorithm described, each of the mammalian genomes was aligned to the cat contigs and unplaced reads. Similarly, each genome was aligned to human. For each genome pair, reciprocal best match alignments (RBM) represent the best alignment of the aligned region in the first genome to the second genome, as well as the best alignment of the region of the second genome to the first genome. In this study, a list of reciprocal best matches was generated, for each genome pair, by first sorting all alignments by their quality (the MegaBLAST bitscore). Starting with highest quality alignments, each was sequentially retained if, for both genomes, the alignment represented regions that had no higher scoring alignment. A reciprocal best match was allowed to overlap with a previously defined one only if both pairs of alignments involved the same contig pairs and if the length of the overlap between the new alignment and the previous one accounted for less than half of length of the new alignment.

## FEATURE COVERAGE

The coverage of each of NCBI's annotated features was calculated as a percent of its nucleotides that were included in the reciprocal best matches to the cat sequences. The annotated features used here were the gene (5' end to 3' including introns), coding sequence, 5'UTR, 3'UTR, 5 kb upstream and downstream from the gene, and intergenic regions that are more than 5 kb from any annotated gene. For genes with multiple isoforms, the coverage of the coding sequences and UTRs from the isoform with the longest transcript was reported.

## ASSIGNING GENES TO THE CAT

Gene annotation was done using four steps:

1. Align the cat contigs to previously annotated mammalian genomes as provided by NCBI (human, chimp, mouse, rat, cow and dog, see Table 1) and define orthologous regions between cat and each of the other species using Reciprocal Best Matches (RBM, see previous section).
2. For each annotated gene of each annotated genome, use the RBM alignments to map the gene annotations to their corresponding regions on the cat contigs.
3. For each gene of each annotated genome, determine if the transferrred annotations on the cat contigs are consistent with their chromosomal assignments and can be combined to generate a single orthologous region representing the gene (exons and introns). This step results in a set of orthologous regions on the cat assembly that correspond to the genes annotated in each of the six mammalian genomes (details below).
4. Merge the six sets of orthologous regions on the cat assembly (representing the genes of the six other genomes) into a nonredundant set of putative genes on the cat genome.

The resulting set of putative cat genes has not been verified to have a transcript in cat, something which must await the availability of a cDNA library for cat. However, unlike ab initio gene prediction methods, this procedure allows the resulting putative cat genes to be easily cross referenced with descriptive information, including gene symbols and keywords describing gene function as assigned by the Gene Ontology and OMIM database.

To define the region in the cat assembly that corresponded to the annotated gene (step 3 above), a moving window of two times the length of the gene was used to scan each strand of each cat chromosome. For each window, the MegaBLAST bitscores of any RBM alignments between the cat contigs in the window and the region of the mammalian genome that included the annotated gene (introns or exons) were summed. The chromosome window position resulting in the highest summed bitscore was considered to span the orthologous region for the gene, and the alignments within the window were used to define a start and stop of the orthologue on the cat chromosome.

For each annotated genome, these steps resulted in one set of regions of the cat assembly that were orthologous to the annotated genes. A region was dropped if it had poor representation of the originally annotated gene, that is, if the orthologous region had a length that was less than 5% of the length of the originally annotated gene. When exon annotation was available, the mammalian exons were then aligned to the orthologous region using MegaBLAST, and the alignments were used to assign regions of the cat assembly that corresponded to exons.

Preliminary visual inspection of the placement of the putative orthologues on the cat assembly revealed inconsistencies in mammalian gene annotations, specifically, cases where regions of the cat genome that corresponded to one large gene model in one species corresponded to two or more neighboring genes in the other species. For example, the exons of the chimp gene model LOC612452 aligned to the cat genome at the same regions as the exons for both ALDH41 and TAS1R2 from mouse, human and rat. As such, the chimp gene formed a chimera of the two genes represented in human and rodent. In all, we found 1575 of these cases. These chimeric orthologues were excluded the subsequent steps of merging the orthologous regions into a non-redundant set of putative cat genes.

The six sets of orthologous regions included redundant representations on cat, in that the six genomes included different genes that have the same orthologue in cat. An effort was then made to merge these redundant representations to a single putative cat gene. The merging procedure (step 4 above) entailed starting with a core set of distinct putative cat genes, and then sequentially either merging the other potential orthologues with a member of the core group or using them to define new putative cat genes. The merging procedure used heuristics to decide when two regions represented the same putative cat gene and the procedure was only applied to orthologous representations that overlapped on the cat genome. The highest priority for merging was for annotated genes that had exons with overlapping mapping positions on the cat genome. Priority was also given to regions that corresponded to two genes that had the same gene symbol, or that belonged to the same entry of NCBI's HomoloGene database.

The steps used to merge redundant orthologous regions to one putative cat gene were as follows:

1. The preliminary core set of putative cat genes were the regions of the cat assembly that corresponded to human genes as described above. The regions in the core set were assumed to represent distinct putative genes on cat and were never merged with one another.
2. The regions of the cat genome that corresponded to exons of the remaining non-human orthologues were identified, and, if they overlapped with exons from a gene in the core set, the region representing the non-human gene on cat was merged with the putative gene in the core set. This merging entailed expanding the length of the putative cat gene, so that its start and stop would span both orthologous regions. However, this step results in the core set having the same total number of putative cat genes as in Step 1.
3. Those mouse genes that had exons that were assigned to the cat genome, but that were not yet merged in the core group in step 2 were used to define new putative genes on the cat genome, with the orthologous region of each unassigned mouse gene representing one new gene in the core group. This step resulted in an expansion of the core set, so that each entry included the orthologous region of at least one human or mouse gene.
4. Unmerged orthologous regions were then re-tested as in step 2, using the expanded core set from step 3.
5. Those genes with annotated exons that were not yet assigned to the core group were merged with one another based on the corresponding overlap of their exons on the cat genome, and were used to define a new putative gene in cat, thus expanding the number of putative cat genes in the core set.
6. The remaining orthologous regions on the cat genome (those which were not yet pulled into the core group), represented genes that either did not have annotated exons or that had an alignment to cat that did not span its exons. Merging of these regions with those in the core group was done when their orthologous region on the cat genome overlapped with that of another gene that shared the same gene symbol.
7. Step 6 was repeated for remaining genes, but merging regions corresponding to genes that shared the same entry of NCBI's HomoloGene database.
8. Those genes that did not participate in any of the previous steps yet overlapped with an entry in the core group were merged with that entry.
9. The last stage used the remaining genes to define a new putative cat gene if they overlapped with one another. This step increased the number of entries in the core set.
10. An entry in the core set was dropped if all of the annotated genes used to define the entry had recently been discontinued in NCBI's Gene's database (last modified: Sept 18, 2006).
11. An entry in the core set was dropped if the support for the putative cat gene was not based on at least two current annotated genes or one annotated human or mouse gene.

These steps resulted in a set of 20,285 putative genes annotated on the cat assembly. The span of each putative cat gene was defined by the total range of the member orthologous regions.

## SYNTENY

For each of the originally annotated species (human, chimpanzee, mouse, rat, dog, cow), the gene order of the putative orthologues on cat was compared with the order of the original genes on the annotated genome. To this end, the midpoint of the gene orthologues on cat and their counterparts on the annotated genomes were taken to represent their positions. These gene positions were then sorted for each chromosome of each genome (unplaced contigs in either genomes were not included), and the neighbors of each gene were determined. If the gene neighbors for the annotated gene were consistent with the order of the corresponding genes on the indexed genome, the gene position was flagged as having valid synteny relative to its neighboring genes. This analysis excluded those genes which were not flanked by neighboring genes in the cat genome.

Instances where a gene fails this syntenic test include micro-rearrangements, cases of assembly errors and cases where the mid-point position estimate for the gene may be an incomplete representation of the position of the gene. Further, as unplaced contigs and genes missing in the cat were not included in the analysis, passing of the test does not explicitly suggest that the genes form syntenic triplets on the biological chromosome of both genomes. For example, the cow and chimpanzee genomes include very short, unplaced contigs, some with annotated genes, which could eventually be placed within a proposed chromosome. These latter cases, of unplaced contigs in cat and in the other genomes, were considered as being beyond the scope of the investigation and were the basis of the exclusion of unplaced contigs from this analysis.

This analysis was also done relaxing the neighbor requirements, allowing gene triplets to be formed using neighbors two genes away in the annotated genome.

## CONSERVED SEQUENCE BLOCKS AND HOMOLOGOUS SYNTENY BLOCKS

The regions of the cat contigs that were longer than 50nt and which were consistently represented in reciprocal best match alignments with the other taxa (Rodent, Primate and Cetartiodactyla) were considered Conserved Sequence Blocks (CSB). These sequences were subsequences of the reciprocal best matches, and as such, may include the 16nt exact match used to seed the original reciprocal best match, but they may also represent the regions of the reciprocal best matches that are less well aligned.

Three levels of stringency were used to define the conserved sequence blocks. The least stringent method required that the region in cat be represented by reciprocal best alignments represented by human, mouse and cow. The second method required that the cat region be represented in reciprocal best matches to all six of the other mammalian genomes used here (human, chimpanzee, mouse, rat, dog, cow). The final and most stringent method further required that for each CSB, the orthologous region in human also form a reciprocal best match with each of the orthologous regions from the other genomes, thus forming a “three way best match”. For the calculation of homologous synteny blocks, the last, most stringent criterion for CSBs was used, but using only those 98,313 CSBs that had chromosomal assignments in the cat and the 5 other genomes used in the HSB analysis (human, chimp, mouse, rat, dog).

## GENERIC GENOME BROWSER

A summary of the result of this collaboration is a set of Gene Annotation Resource Field (GARFIELD) which can be displayed using The Laboratory of Genomic Diversity’s version of the Generic Genome Browser (lgd.abcc.ncifcrf.gov). GARFIELD is an interactive online web resource. The browser provides views for the cat chromosomes, as well as placeholder chromosomes for unplaced contigs. The Generic Genome Browser allows for groups of data to be displayed on individual ‘tracks’ either on a per-chromosome as well as more detailed views.

The overview tracks of GARFIELD include density of repetitive elements as measured using RepeatMasker, as well as GC content, and heterozygosity. The density tracks were constructed by dividing the chromosomes into non-overlapping windows of 100 kb and these windows were sorted with respect to the percent of bases of the various annotated regions. Each window was then assigned a percentile rank based on its density of the repetitive element relative to all the other windows.

The more detailed tracks include a track for individual contigs that make up the chromosome, as well as Gene regions as described in the text. Tracks are also available for fosmid reads and their partners, SNPS, STRs, miRNAs, regions that align to mitochondrial DNA, and regions that align to retroviruses. GC content is provided as a histogram and, for windows widths less than 200nt, the exact nucleotide sequence is  displayed.

## FELINE ENDOGENOUS RETROVIRUS-LIKE ELEMENTS (FERVS)

### *Screening of traces for matches to enFeLV, RD-114 and other FERVs.*

The masked (RepeatMasker) cat traces were compared to GenBank entries for a full-length endogenous FeLV with an intact genomic sequence (accession number AY364318); for the partial (largely env region) RD-114 sequence (accession number X87829); or for contig or BAC sequences of novel FERVs. Matching traces were identified using MegaBLAST (Zhang et al. 2000) including arguments:

```
-W 24 -U T -F "m D" -D 3 -m 8 - s 100 -r 1 -q -1 -X 40.
```

### *Editing and phylogenetic analyses of traces*

Traces were edited to remove end sequences with quality scores below 20; some traces were removed from the analyses due to factors such as low sequence quality or lack of overlap with the DNA region being analyzed. Sequences were aligned using the software CLUSTAL_X (Thompson et al. 1997) and (Rambaut 1996); alignment output was visually inspected; amino acid alignments were edited using MacClade 4.08 (Maddison and Maddison 2005). Phylogenetic analyses of the data sets were performed using maximum parsimony (MP), neighbor joining (NJ) or minimum evolution (ME), and maximum likelihood (ML) methods implemented in PAUP*4.0b10 (Swofford et al. 1996), and employed heuristic searches with 50 replicates of random taxon-addition and TBR branch swapping for MP, ME and ML. MP was used for both nucleotide and amino acid data, while nucleotide data was analyzed using the other methods also. The software Modeltest 3.7 (Posada and Crandall 1998) was used to determine the model of DNA sequence evolution that best fit nucleotide data. For each DNA segment, the model selected was implemented in PAUP*4.0b10 using Modeltest generated likelihood settings. The Modeltest ML settings were used for NJ or ME analyses and for ML analyses. Bootstrap resampling support was based on at least 100 replicates, with TBR branch swapping of starting trees obtained by stepwise addition.

*Search for retroviral elements in the genomic sequence of Felis catus*

Computational scanning of the cat genome for putative retroviral sequences was implemented in two main steps: identification of regions of the cat genome that are homologous to known retroviruses; and estimation of the putative borders of retroviral elements. Genomic contigs were scanned with BLAT (Kent 2002), (parameters "-q=rnax –t=dnax") using 703 retroviral sequences available from GenBank. We found 2260 genomic loci with segments at least 100 bp long with at least 70% similarity to a retroviral sequence. These homologous segments were extracted from the contigs with a maximum of 10 kb flanking regions on both sides. To identify the putative borders of the retroviral elements we analyzed the repetitive structure of these genomic fragments using REPuter (Kurtz et al. 2001; Kurtz and Schleiermacher 1999) with a minimal repeat length of 30 bp. Those fragments that included repeat pairs encompassing the homologous segment and that were 3 kb to 10 kb apart were reported as including a putative retroviral element with the boundaries defined by the repeat coordinates. There were 379 sequences found with these properties; 295 of them were similar to LINE1, while 84 had similarity to known retroviral sequences.

## INTERSPERSED REPEAT ELEMENTS:

Our goal was to characterize the occurrence of SINE and LINE elements in the 1.9x sequences and scaffold construction of the cat WGS. RepeatMasker was used to identify and quantify the occurrence of different classes of known elements, and to compare the representation of these elements in the traces and the scaffolds.

A preliminary comparison of the number and percentage of the repeats that were masked by RepeatMasker in the unplaced contigs (those which were not positioned by the assisted assembly) versus the scaffolds revealed that 60.4 % of the individual SINEs were placed on scaffolds compared with only 48.0% of the LINEs. The longest LINE segments were only around 1500 bp (around 50% of their expected length) and thus, the 1.9x WGS was unable to assemble full-length LINE elements. However, a larger percentage of LINE2 elements were placed on scaffolds relative to LINE1 probably because of their relatively smaller size. Similarly, the smaller size of SINEs (approximately 250-450 bp) is likely the reason for the larger number of full-length SINE elements that could be sequenced completely in the 1.9x cat genome.

*LINEs*

Relative to the dog (7x coverage), the percentage of the domestic cat genome (based on repeat masking of the traces) composed of repeat elements is substantially less (35.2 vs. 22.3 %), due primarily to a difference in the relative percentage of LINEs (18.7 vs. 7.9 %) (Table 2). The most common carnivore specific LINEs were Canis1, Carn1, Carn2, Carn3, Carn4, Carn5, Carn7, and Fc (Felid). Complete open reading frames (without stop codons) in the reverse transcriptase (RT) portions of LINE1 were only found in CANID and Fc LINEs, implying that these elements may remain active in the cat genome, and that the CANID LINE may be active in other carnivore species.

Phylogenetic analyses indicate that mean percent JTT amino acid divergence among Fc LINEs was 11.8% compared with 17.3% for CANID, which is consistent with a longer evolutionary existence of CANID elements. No strongly supported groupings of either CANID or Fc sequences were apparent. LINE2, which was spread initially prior to the mammalian radiation, made up only 1.5 % of the cat genome, compared with 2.8 % for the dog.

*SINEs*

Full-length SINE elements were obtained using RepeatMasker criteria. Each SINE element consists of three regions: a tRNA-related region containing motifs which presumably are RNA polymerase III promoters, a polypyrimidine repeat region, and A/T rich tail. Full-length SINEs (N=337) were subdivided into four different classes by RepeatMasker. These classes are B2, carnivore SINEs B1 and B2; canid SINEs of C1, C2, and felid specific SINEs FC 1 and 2 previously identified (Batzer et al. 1996; Schmid 1996; Smit 1993; Smit 1995; Smit 1996; Smit and Riggs 1995; Smit and Riggs 1996; Smit et al. 1995; Vassetzky and Kramerov 2002; Wilkerson et al. 1994; www.repeatmasker.org). RepeatMasker criteria classified the majority of SINEs as a tRNA-like origin (MIR) rather than B1 or B2-like (ALU).

Phylogenetic analyses were used to determine the origin and diversification of the cat SINEs. Due to extensive variation within the polypyrimidine repeat region and the A/T rich tail, these two regions were omitted from the phylogenetic analyses of 337 full-length SINEs. As presented in representative Figure S7, each SINE element is unique, with virtually no identical SINEs present within the cat 1.9x reads. The SINEs had been assigned by RepeatMasker as belonging to the SINE-Lys class of Carnivore Felis catus class 1 and 2. As shown in the phylogeny, the sequences appear to cluster in accordance with that classification.

This analysis was extended to all of the RepeatMasker defined groups (data not shown). In each case, the majority of the sequences were misidentified by this method used by RepeatMasker. Rather, phylogenetic analyses recapitulate the trends indicated in Figure S7 that the majority of cat SINEs are all more closely related to the Felis catus groups Fc1 and Fc2.

In addition, the phylogenetic trees indicate there are additional SINEs that are not clearly affiliated with Fc1 and 2, but appear to be more divergent and ancestral (Figure S7). These ancestral SINEs are not clearly associated with any of the reference sequences from other carnivore taxa.

*Distribution of LINEs and SINEs on cat chromosomes*.

Upon assembly into chromosomes, RepeatMasker indicated that 30-37% of the cat genome is composed of SINEs and LINEs. The chromosomes with the highest repeat content were X and E3 (Figures S6). In the "unknown" chromosome category, 68.8% of the sequence was repeat elements amounting to one element per every 600 bp.  The frequency distributions of LINEs and SINEs show a consistent pattern with LINE1 family of LINEs representing the greatest percentage of repeat nucleotides ranging from 8.4-13.8% per chromosome. Cat SINEs (MIR) ranged from 7.6-9.5% per chromosome nucleotides.

**miRNA**

In order to identify potential micro-RNA sequences in the cat genome we mapped all vertebrate specific miRNA precursors contained in the Micro-RNA Registry Version 8.0 (MR8.0 http://microrna.sanger.ac.uk/sequences/) to the cat scaffolds using BLAST ( E<0.01, alignment length >50bp). The overlapped hits were clustered together and were assigned the corresponding micro-RNA family name. These sequences were extracted from the cat genome and tested for stem-loop secondary structure by the RNAfold program (Hofacker 2003) . The verified sequences were selected as cat representatives of the micro-RNA family.

## REFERENCES

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Ziętkiewicz, E., and Zuckerkandl, E. 1996. Standardized nomenclature for Alu repeats. *J Mol Evol* **42:** 3-6.

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., Buzzel, B.L., Drummond, M.C., Wirth, B. and O'Brien, S.J. 2006. An ~140-kb deletion associated with feline spinal muscular atrophy implies an essential LIX1 function for motor neuron survival. *Genome Research* **16**:1084-1090.

Ishida, Y., David, V.A., Eizirik, E., Schäffer, 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 pheotype 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.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29:** 4633-4642.

Kurtz, S. and Schleiermacher, C. 1999. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15:** 426-427.

Maddison, D.R. and Maddison, W.P. 2005. *MacClade 4: Analysis of phylogeny and character evolution. Version 4.08.* Sinauer, Sunderland, MA.

Margulies, E.H., Vinson, J.P., NISCComparativeSequenceProgram, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C. et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A* **102:** 4795-4800.

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.

Posada, D. and Crandall, K.A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Rambaut, A. 1996. Se-Al: Sequence Alignment Editor. Available at http://evolve.zoo.ox.ac.uk/.

Schmid, C.W. 1996. Alu: structure, origin, evolution, significance and fucntion of one-tenth of human DNA. *Prog Nucleic Acids Res Mol Biol* **53:** 283-319.

Smit, A.F.A. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res* **21:** 1863-1872.

Smit, A.F.A. 1995. Ph.D. Thesis: Structure and evolution of mammalian interspersed repeats. University of Southern California.

Smit, A.F.A. 1996. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6:** 743-748.

Smit, A.F.A. and Riggs, A.D. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res* **23:** 98-102.

Smit, A.F.A. and Riggs, A.D. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93:** 1443-1448.

Smit, A.F.A., Tóth, G., Riggs, A.D., and Jurka, J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246:** 401-417.

Swofford, D.L., Olsen, G.J., Waddel, P.J., and Hillis, D.M. 1996. *Molecular Systematics*. Sinauer Associates, Sunderland, MA.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

Vassetzky, N.S. and Kramerov, D.A. 2002. CAN--a pan-carnivore SINE family. *Mamm Genome* **13:** 50-57.

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:** D39-45.

Wilkerson, D.A., Mager, D.L., and Leong, J.C. 1994. The Retroviridae (ed. J.A. Levy), pp. 465-535. Plenum Press, New York.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7:** 203-214.

ftp://ftp.ncbi.nih.gov/pub/HomoloGene/.

lgd.abcc.ncifcrf.gov

www.cpan.org.

www.olegh.spedia.net.

www.repeatmasker.org.

## Supplemental Tables

Table S1. Hereditary diseases in domestic cats

Albinism, oculocutaneous
Achondroplasia
Alpha fucosidosis
Alpha mannosidosis
Amyloidosis, renal
Amyloidosis, hepatic
Anencephaly
Anophthalmia congenital
Aortic stenois
Asthenia cutis
Atresia ani
Atresia coli
Atrial septal defect
Brachydactyly
Calcium oxalate urinary calculi
Cataract, congenital
Cataract, nuclear
Cataract, juvenile
Chondrodysplasia
Cerebellar abiothrophy
Cerebellar hypoplasia
Ceroid lipofucinosis
Chediak-Higashi syndrome
Ciliary dyskinesia
Cleft palate, severe
Cleft palate, mild
Cholesterolester storage disease
Chondrodysplasia
Chylothorax
Cobalamin malabsorption
Coloboma
Corneal dystrophy, stromal
Corneal dystrophy, endothelial
Corneal sequestration
Congenital goiterous hypothyroidism
Congenital hypothyroidism, aplastic
Craniofacial defect
Cryptorchism
Curl tail
Cystinuria type I
Cystinuria, non-type I
Deafness, white
Deafness, other
Dermoid
Dermatosparaxia
Diabetes mellitus
Diaphragmatic hernia
Dwarfism
Dystrophin deficiency
Ectrodactyly
Ehrler-Danlos syndrome type I
Endocardofibroelastosis
Epibulbar dermoid
Erythrocyte osmotic fragility
Esophageal achalasia
Exocrine pancreatic insufficiency
Factor X deficiency
Factor II deficiency
Factor XII deficiency
Femoral head epiphysal slip
Fanconi syndrome
Fold ear with osteodystrophy
Gangliosidosis GM1
Gangliosidosis GM2
Globoid cell leukodystrophy
Glaucoma
Glycogenosis type IV
Hairlessness
Hemivertebrae
Hemolytic transfusion reactions
Hemophilia A
Hemophilia B
Heterochromia
Hiatal hernia
Hip dysplasia
Hip (coxofemoral) luxation
Hydrocephalus, internal
Hydrocephalus, external
Hyperchylomicronemia
Hyperkalemic paralysis
Hypertrophic cardiomyopathy
Hypotrichosis
Hypotrichosis with thymic aplasia
Inguinal hernia
Inflammatory bowel disease
Intersex
Isovaleric aciduria
Kinked tail
Lymphoma
Lactic aciduria
Lipoprotein lipase deficiency
Lissencephaly
Methylmalonic aciduria
Megacolon, denervation
Megaesophagus
Meningiencaphalocele
Methemoglobin reductase deficiency
Microbrachia
Microophthalmia
Milliary dermatitis
Mitral valve dysplasia
Mucolipidosis type II
Mucopolysaccharidosis Type I
Mucopolysaccharidosis Type VI

Mucopolysaccharidosis Type VII
Myasthenia gravis
Myotonia congenital
Neonatal isoerythrolysis
Neuroaxonal dystrophy
Neutrophil granulation
Nieman-Pick disease type C
Osteogenesis imperfecta, dominant
Osteogenesis imperfecta, recessive
Odontoclastic resorptive lesions
Ornithin aminotransferase deficiency
Open central fontanel
Open lateral fontanel
Patellar luxation
Pectus excavatum
Pelger Huet anomaly
Perirenal pseudocysts
Persistent ductus arteriosus
Persistent hepatic ductus venosus
Persistent Muellerian duct syndrome
Persistent papillary membranes
Persistent right aortic arch
Persistent truncus arteriosus
Polycystic kidney disease
Polycythemia (erythrocytosis)
Polydactyly
Porphyria, dominant
Porphyria with anemia, recessive
Portocaval shunt
Predisp. to feline infectious peritonitis
Predisp. to ginigivitis
Primary hyperoxaluria type I
Primary hyperoxaluria type II
Progressive retinal atrophy, Siamese
Progressive retinal atrophy, Persian
Progressive retinal atrophy, other
Pulmonary stenosis
Pyloric stenosis
Pyruvate kinase deficiency
Radial agenesis
Renal dysplasia
Renal tubular acidosis
Retinal dystrophy
Restrictive cardiomyopathy
Sacrococcidial agenesis
Situs inversus
Spastic syndrome
Spheroid lysosomal storage disease
Sphingolipidosis C
Spina bifida
Spinal muscular atrophy
Spondylosis deformans
Spongiform encephalopathy
Strabismus
Syndactyly
Syringomelia
Taillessness
Taurin deficiency/malabsorption
Tendency for cotton chewing
Testicular feminization
Tetralogy of Fallot
Thrombopathia
Thyroid peroxidase deficiency
Tracheal hypoplasia
Tricuspid valve stenosis
Umbilical hernia
Urical diverticle
Ventricular septal defect
Vestibular defect
Vitamin K dependent coagulopathy
Von Willebrand disease type I
XXY karyotype

Table S2. Reciprocal Best Match (RBMs) alignments to F*elis catus* and six mammalian genomes.

| Genome | Avg.<br>Percent ID | Avg.<br>Length | Rel Length:Cat | # RBM | % of Cat<br>Assembly |
|---|---|---|---|---|---|
| *Homo sapiens* | 73.0 +/-- 5.0 | 1000 +/- 736 | 1.005 +/- 0.038 | 792,706 | 31.74 |
| *Pan troglodytes* | 72.9+/-5.1 | 976+/-710 | 1.005+/-0.038 | 776,865 | 30.0 |
| *Mus musculus* | 69.1+/-5.7 | 973+/-663 | 0.980+/-0.039 | 283,426 | 11.0 |
| *Rattus norvegicus* | 69.1+/-5.8 | 968+/-659 | 0.978+/-0.039 | 267,764 | 10.4 |
| *Canis familiaris* | 78.8+/-5.2 | 964+/-795 | 0.999+/-0.035 | 1,235,641 | 47.6 |
| *Bos taurus* | 73.4+/-5.1 | 927+/-667 | 0.996+/-0.038 | 807,061 | 29.9 |

Summary of the reciprocal best alignments between cat WGS sequences and the other mammalian genomes. Provided is the mean and standard deviation of the percent ID of the reciprocal best match alignments, the mean and standard deviation of the length of the alignment. The Relative Length is a ratio of the length of the aligned region of the second genome relative to that of the cat sequence. Thus, primate orthologous sequences are on the average 0.5% larger than cat, while rodent orthologous sequences are 2% smaller.

Table S3. Deriving Conserved Sequence Blocks (CSBs) based upon multi-species RBMs

| Method | | Ave. Length | Number of CSB |
|---|---|---|---|
| A. | Four taxa: Cat human, mouse and cow | 919.2+/-623.8 | 208,048 |
| B. | Cat and six index mammals | 830.3+/-565.6 | 166,843 |
| C. | Seven genomes and three way best matches | 882.2+/-584.8 | 133,499 |

A. Regions in cat contigs that include a RBM match in alignments to human, mouse and cow genome sequences.
B. Regions in cat contigs that include a RBM match in alignments to all six genomes (human, chimpanzee, mouse, rat, cow and dog).
C. RBM regions in the cat and six mammal genomes (as for B) which also were confirmed as being three-way best matches between human, cat and each of the other four species. Drop-out RBMs here include ambiguous alignments between paralogous sequences in the compared species genomes.

Table S4. Detailed description of the rearrangements on each edge of the tree shown in Figure 2

a) broken down by intrachromosomal events (INVersions) and interchromosomal events (TRAnslocations, FISsions, and FUSions).

| | Intra | Interchromosomal | | | |
|---|---|---|---|---|---|
| | INV | TRA | FIS | FUS | TOT |
| EA → BA → CDA | 22 | 9 | 0 | 0 | 31 |
| CDA → Cat | 37 | 7 | 0 | 3 | 47 |
| CDA → Dog | 5 | 31 | 17 | 0 | 53 |
| EA → MRA | 68 | 69 | 0 | 5 | 142 |
| MRA → Mouse | 19 | 14 | 3 | 0 | 36 |
| MRA → Rat | 25 | 9 | 4 | 0 | 38 |
| EA → PHA | 42 | 8 | 1 | 0 | 51 |
| PHA → Chimpanzee | 6 | 0 | 1 | 0 | 7 |
| PHA → Human | 5 | 1 | 0 | 0 | 6 |
| Total | 229 | 148 | 26 | 8 | 411 |

b) Rates per Myr of the events on the tree in Figure 2

| | Intra | Interchromosomal | | | |
|---|---|---|---|---|---|
| | INV | TRA | FIS | FUS | TOT |
| EA → BA → CDA | 0.48 | 0.20 | 0.00 | 0.00 | 0.67 |
| CDA → Cat | 0.67 | 0.13 | 0.00 | 0.05 | 0.85 |
| CDA → Dog | 0.09 | 0.56 | 0.31 | 0.00 | 0.96 |
| EA → MRA | 0.96 | 0.97 | 0.00 | 0.07 | 2.00 |
| MRA → Mouse | 1.19 | 0.88 | 0.19 | 0.00 | 2.25 |
| MRA → Rat | 1.56 | 0.56 | 0.25 | 0.00 | 2.38 |
| EA → PHA | 0.52 | 0.10 | 0.01 | 0.00 | 0.63 |
| PHA → Chimpanzee | 1.09 | 0.00 | 0.18 | 0.00 | 1.27 |
| PHA → Human | 0.91 | 0.18 | 0.00 | 0.00 | 1.09 |

Table S5. The 1.9x WGS cat genome statistics

| Input data* | Number of reads | trimmed read length | total bases (millions) | sequence coverage | Q20 coverage | % paired | Fraction assembled |
|---|---|---|---|---|---|---|---|
| Fosmid | 1,292,111 | 591 | 763.6 | 0.28 | 0.25 | 90.7% | 64.9% |
| Plasmid | 6,735,561 | 648 | 4,364.6 | 1.62 | 1.48 | 90.0% | 81.6% |
| Total | 8,027,672 | 639 | 5,129.7 | 1.90 | 1.73 | 90.1% | 78.9% |

* Approximately 159,000 reads were excluded from this data set due to low quality or vector content. The final sequence data provided 1.9x overall sequence redundancy, and 1.73x Q20 coverage, based on an estimated genome size of 2.7 Gb.

Table S6. Assembly of Cat Genome

| | |
|---|---|
| Estimated genome size | 2.7 Gb |
| Coverage | 1.9x |
| Total reads | 8,186,934 |
| - in assembly | 6,334,156 |
| - unassembled | 1,852,778 |
| | |
| No contigs | 817,956 |
| Total bases in contigs | 1,642,698,337 bp |
| N50 (contig size) | 2,378 bp |
| No. of scaffolds (supercontigs) | 217,790 |
| N50 (scaffold size) gapped | 117,081 bp |
| Total length scaffolds gapped | 3,937,914,851 bp |
| N50 (scaffold size) ungapped | 45,200 bp |
| No. of RH markers | 1,680 |

*Four genome assemblies were generated using PHUSION or Arachne. The best, used here, is the BROAD FelCat3 which produced slightly larger scaffolds. All assemblies produced comparable genome coverage.

Table S7. Linkage analysis between microsatellites about two genes that recapitulate exactly the marker order determined in the cat assembly( Fyfe etal 2006;Ishida et al, 2006)

| | Cat locus | | | | Human locus | | Dog locus | |
|---|---|---|---|---|---|---|---|---|
| Marker | Chr | Start[a] | LOD | θ [b] | Chr | Start[c] | Chr | Start[c] |
| *MLPH* and *dilute* | | | | | | | | |
| FCA664 | C1 | 216854694 | 7.48 | 0.20 | 2 | 229918573 | 25 | 40658700 |
| FCA890 | C1 | 221490981 | 10.81 | 0.08 | 2 | 237564169 | 25 | 51421590 |
| FCA_HSA2;237.0 | C1 | 223504481 | 14.11 | 0.06 | 2 | 237001129 | 25 | 50135482 |
| FCA_HSA2;237.6 | C1 | 224124879 | 19.74 | 0.01 | 2 | 237563928 | 25 | 50626236 |
| MLPH | C1 | 224776040 | | | 2 | 238177930 | 25 | 51174485 |
| FCA_CFA25;51.2 | C1 | 224818983 | 29.50 | 0 | 2 | 238201464 | 25 | 51191310 |
| FCA_HSA2;240.1 | C1 | 226597264 | 8.98 | 0.05 | 2 | 240076739 | 25 | 52650295 |
| FCA_HSA2;241.7 | Un25 | 1597038 | 15.45 | 0.04 | 2 | 241655221 | 25 | 53844529 |
| FCA_HSA2;241.8 | Un25 | 1738973 | 14.35 | 0.04 | 2 | 241842037 | 25 | 53994911 |
| *SMA* | | | | | | | | |
| FCA765 | A1 | 160,033,754 | 0.5 | 0.28 | 5 | 55,604,292 | 2 | 46,301,953 |
| FCA767 | A1 | 172,738,995 | 1.61 | 0.18 | 5 | 67,472,758 | 2 | 56,273,001 |
| SMN | A1 | 174,171,565 | 3.7 | 0.14 | 5 | 70,273,558 | 2 | 57,494,052 |
| FCA689 | A1 | 180,430,782 | 7.23 | 0.11 | 5 | 79,353,399 | 3 | 29,879,091 |
| FCA225d | A1 | 182,582,854 | 2.52 | 0.08 | 5 | 81,770,517 | 3 | 27,934,003 |
| FCA768 | A1 | 182,658,524 | 11.5 | 0.07 | 5 | 81,870,949 | 3 | 27,862,439 |
| FCA071 | A1 | 188,085,036 | 17.5 | 0.02 | 5 | 88,098,710 | 3 | 22,783,845 |
| RHOBTB3 | A1 | 194,644,302 | 4.84 | 0.00 | 5 | 95,156,641 | 3 | 16,704,737 |
| LIX1 | A1 | 195,931,144 | 18.92 | 0.00 | 5 | 96,456,140 | 3 | 5,512,185 |
| EFNA5 | A1 | 196,828,696 | 15.35 | 0.01 | 5 | 106,743,742 | 3 | 7,425,325 |
| PAM | A1 | 200,603,302 | 12.64 | 0.01 | 5 | 102,392,381 | 3 | 11,023,200 |
| FCA771 | A1 | 212,179,441 | 5.4 | 0.16 | 5 | 171,056,257 | 4 | 43,700,484 |

a: The position of each locus is based on the cat GARFIELD Browser
b: The θ column shows the optimal recombination fraction between that marker and the gene to within 0.01; the LOD column shows the lod score at the optimal θ.
c: The position of each locus is based on The UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway)

Supplemental Figures

Supplemental Figure Legends

Figure S1. Phylogenetic tree of mammalian species nominated for whole genome sequencing with colors representing different stages of sequencing priority. Black indicates 7 fold coverage complete; other colors represent earlier stages of sequence determination.

Figure S2. *Numt* fragments assigned to the domestic cat chromosomes.

Figure S3 MegaBLAST alignment of highest scoring traces to match endogenous retroviruses:
a) the coding regions of an intact proviral enFeLV genome (accession number AY364318, excluding LTRs and flanks)
b) RD114 sequence. Both intact and truncated genomic RD114 sequences are evident.

Figure S4. Coverage and average contig size for each of the available ENCODE assemblies generated for the cat genome.

Figure S5. Six example regions showing order and orientation of the position of the contigs in the WGS assembly relative to their corresponding multi-BAC assembly positions from the ENCODE project.

Figure S6. Percentage of sequence of LINE1, LINE2, and SINE elements on cat chromosomes based on final assembly.

Figure S7. Maximum likelihood phylogeny of SINEs defined by RepeatMasker as Carnivore class Fc 1, 2, and 3. Analyses were based only on the tRNA-like region of each SINE element with polyA/T tail and di-nucleotide repeat region removed (145 bp). Shown is the ML tree (-Ln likelihood score = 1309.9; 31675 rearrangements tried) derived by PAUP (Swofford et al. 1996) using the GTR+G model with parameters estimated from ModelTest (Posada and Crandall 1998) of 1) rate matrix AC=1.166100, AG = 6.408500, AT = 2.096200, GC = 0.480800, CT = 6.408500, and GT = 1; 2) estimated nucleotide frequencies of A = 0.17270 C = 0.24010 G = 0.35700 T=0.23020; and 3) gamma = 1.7295. Specific search conditions for the ML analyses used starting trees obtained by step-wise addition and branch-swapping using the tree-bisection-reconnection (TBR) algorithm. Numbers at nodes of tree represent bootstrap proportion based on 100 iterations. Each SINE identified by cat chromosome and the GenBank accession number of mapped position.

Figure S8. 35 SNPs were selected across each of ten ~600 kb regions and genotyped in multiple individuals from each of 24 certified cat breeds. Homozygosity within the first 10kb was assessed at 53%. Conditional on homozygosity within the first 10 kb the fraction of observations remaining homozygous at different distances was plotted on a log distance scale. The distance at which 50% of observations remained homozygous was estimated at ~150 kb, rough 1/3 of the distance seen in the dog population.

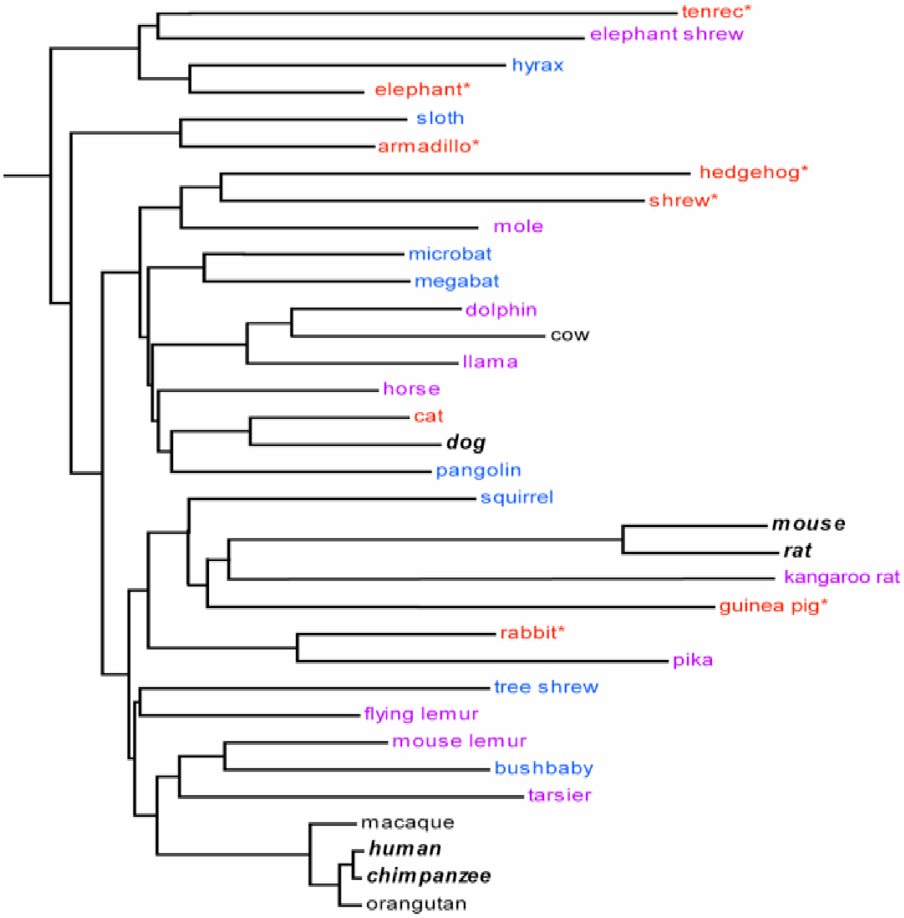

tenrec*
elephant shrew
hyrax
elephant*
sloth
armadillo*
hedgehog*
shrew*
mole
microbat
megabat
dolphin
cow
llama
horse
cat
dog
pangolin
squirrel
mouse
rat
kangaroo rat
guinea pig*
rabbit*
pika
tree shrew
flying lemur
mouse lemur
bushbaby
tarsier
macaque
human
chimpanzee
orangutan

Figure S1

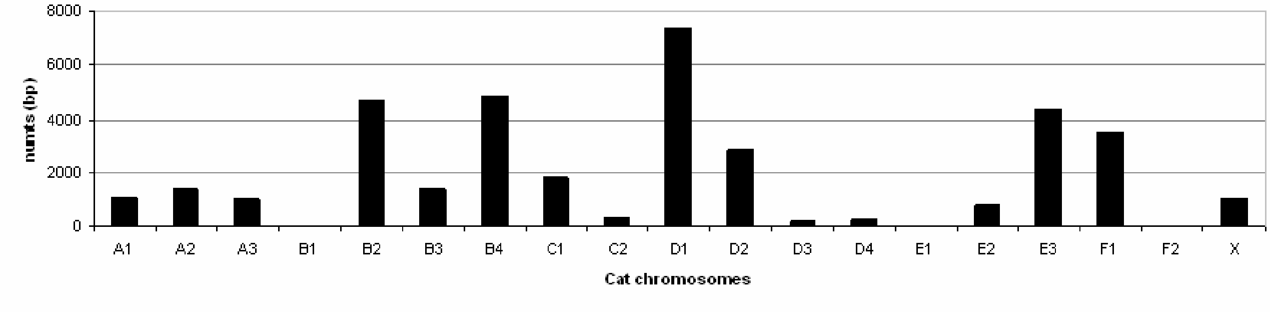
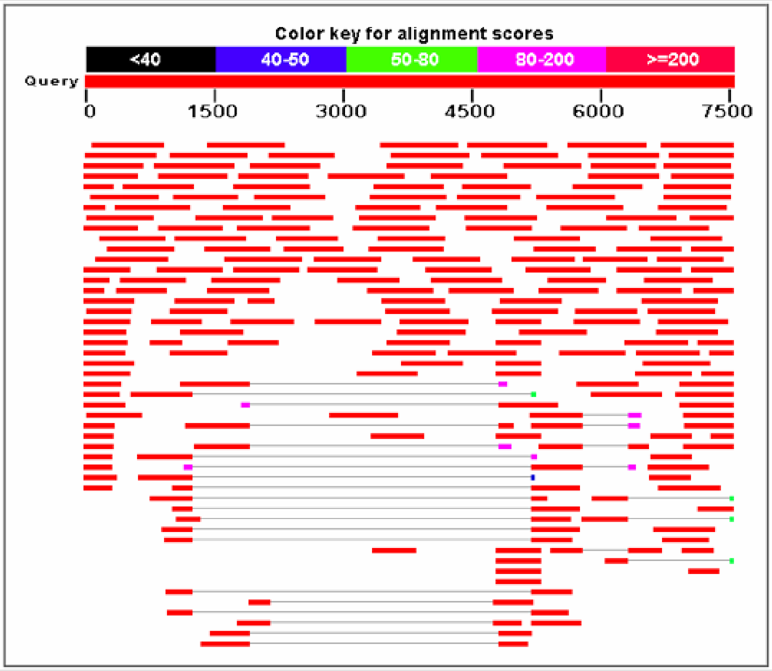
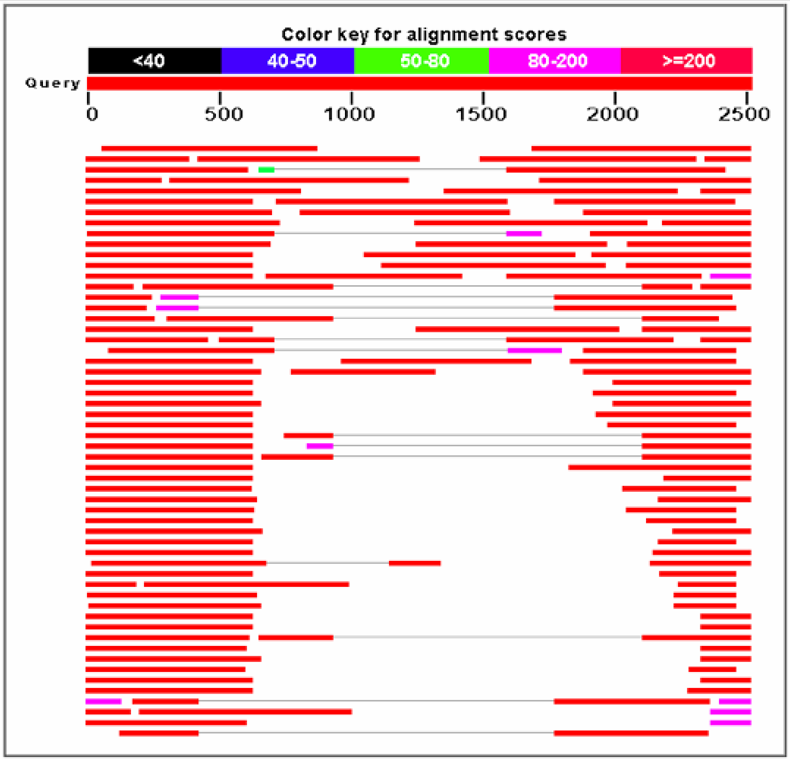


Figure S2. *Numt* fragments assigned to domestic cat chromosomes

Figure S3. Retrovirus coverage by sequence reads
a) enFeLV coverage by sequence reads





FigureS3 b) Coverage of RD114 by sequence reads

Figure S4

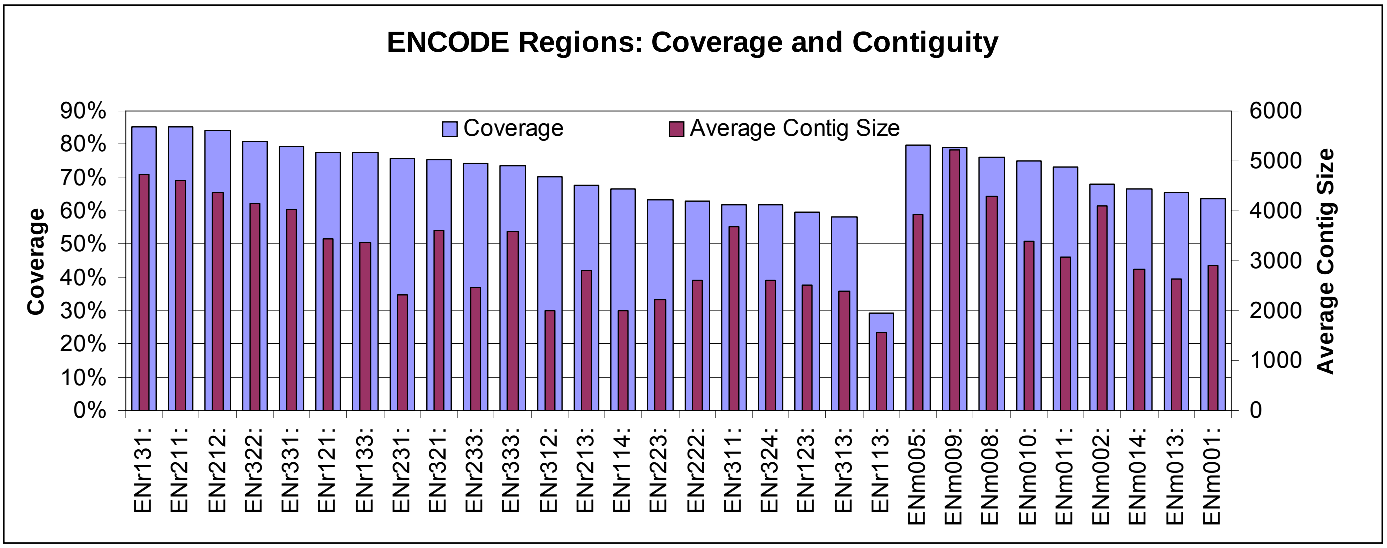
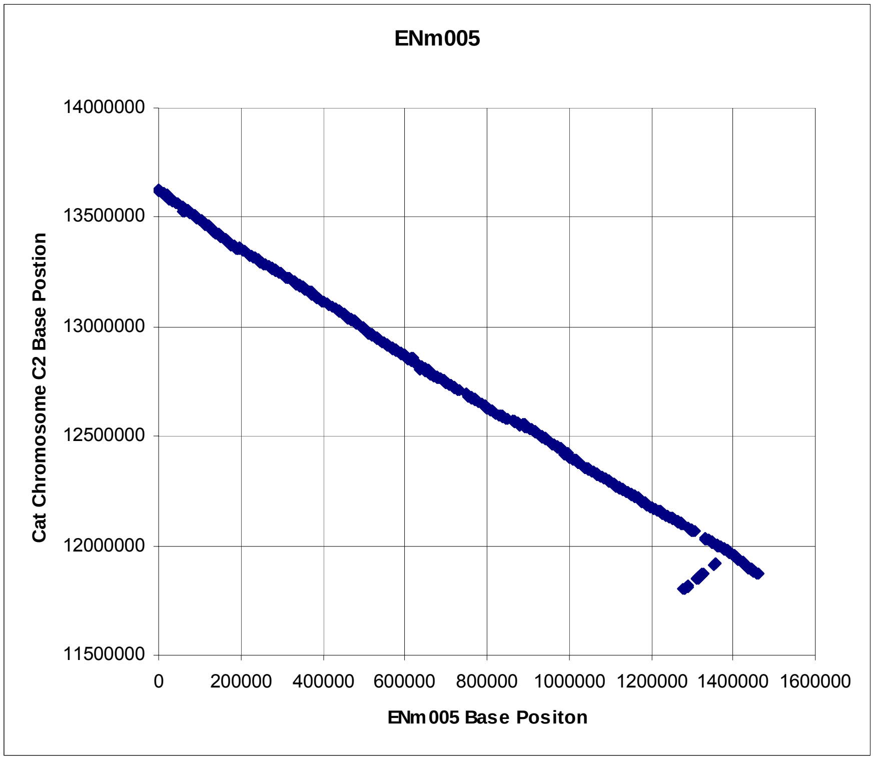
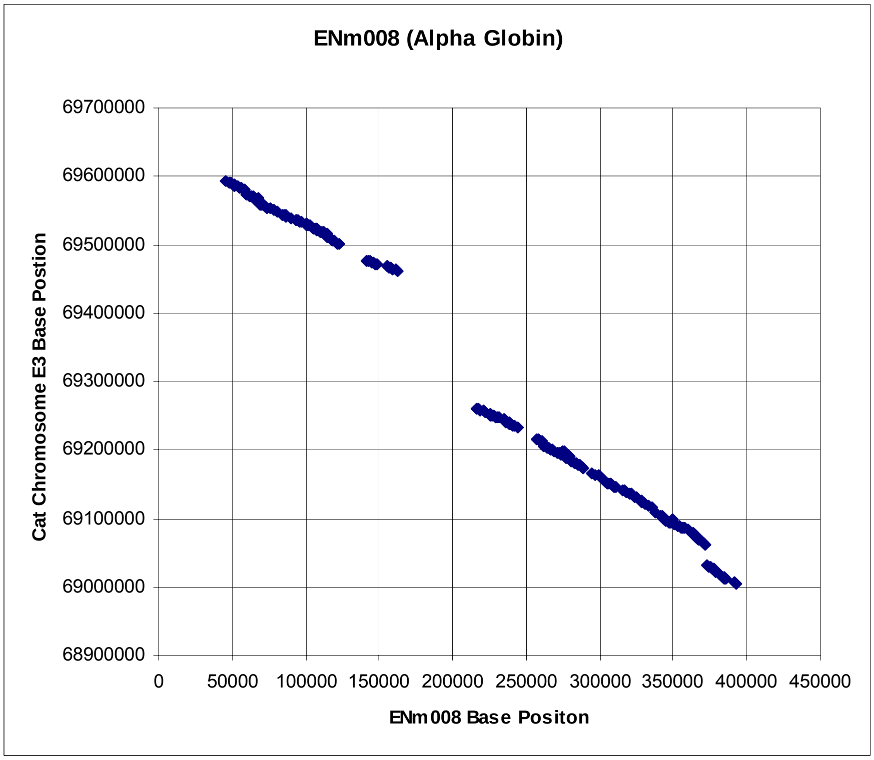
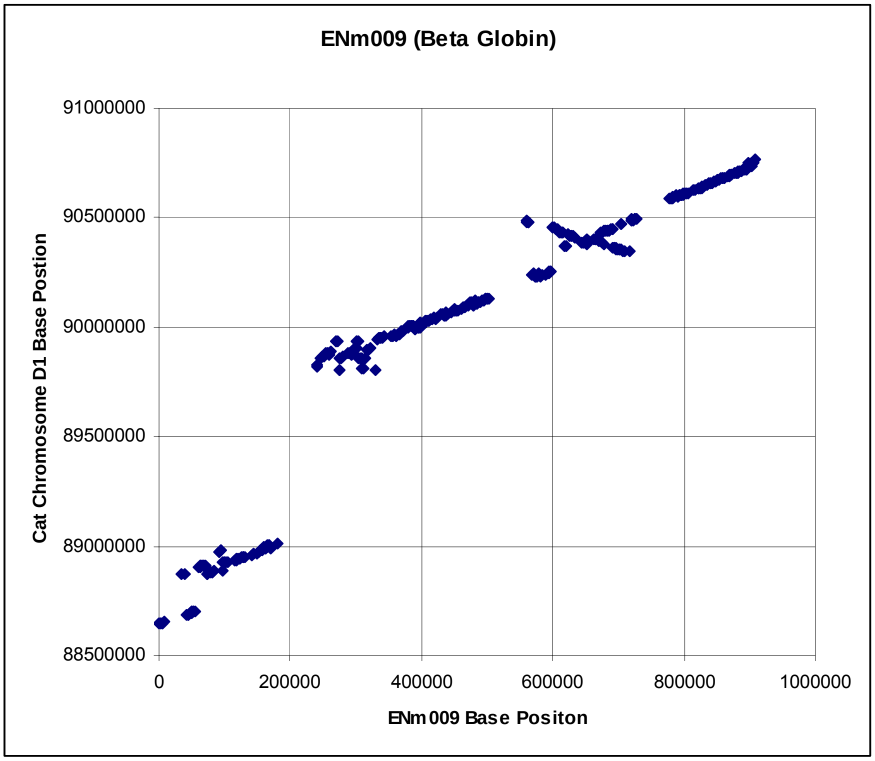
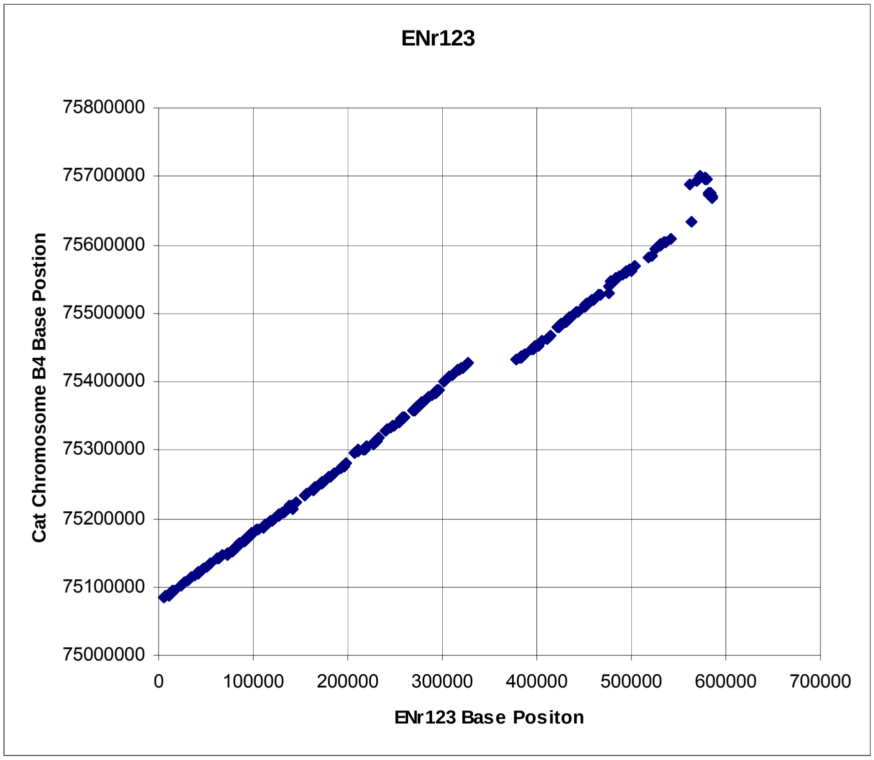


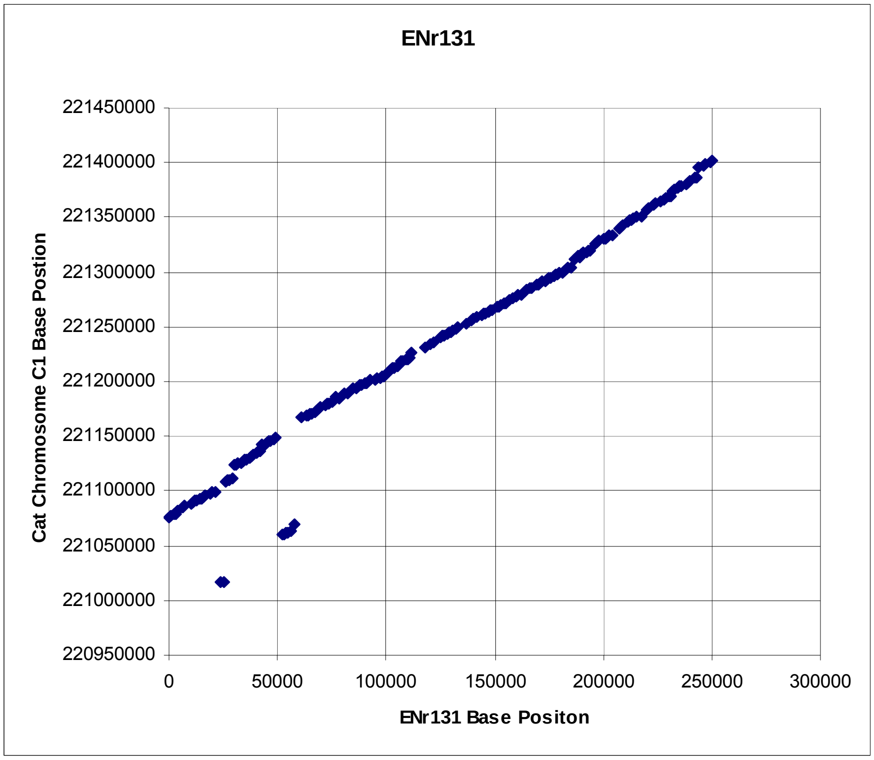
Figure S4. Coverage and average contig size of the 1.9x sequence, corresponding to the ENCODE assemblies.


ENm005
14000000
13500000
13000000
12500000
12000000
11500000
Cat Chromosome C2 Base Postion
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
ENm005 Base Positon


ENm008 (Alpha Globin)
69700000
69600000
69500000
69400000
69300000
69200000
69100000
69000000
68900000
Cat Chromosome E3 Base Postion
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
ENm008 Base Positon


ENm009 (Beta Globin)
91000000
90500000
90000000
89500000
89000000
88500000
0
200000
400000
600000
800000
1000000
Cat Chromosome D1 Base Postion
ENm009 Base Positon


ENr123
75800000
75700000
75600000
75500000
75400000
75300000
75200000
75100000
75000000
0
100000
200000
300000
400000
500000
600000
700000
Cat Chromosome B4 Base Postion
ENr123 Base Positon



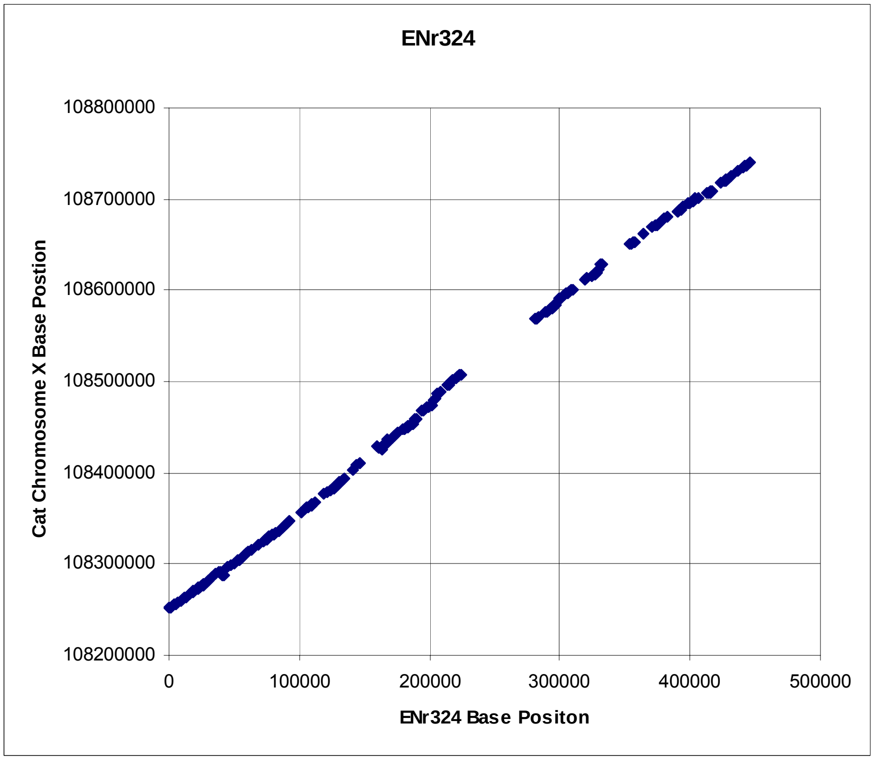


Figure S5 . Example regions showing order and orientation of the position of the contigs in the WGS assembly relative to their corresponding multi-BAC assembly positions There are NISC clone sequence gaps in ENm008, ENm009 (3 gaps), ENr123, and ENr324, which shows up as missing alignments on the respective graphs above.

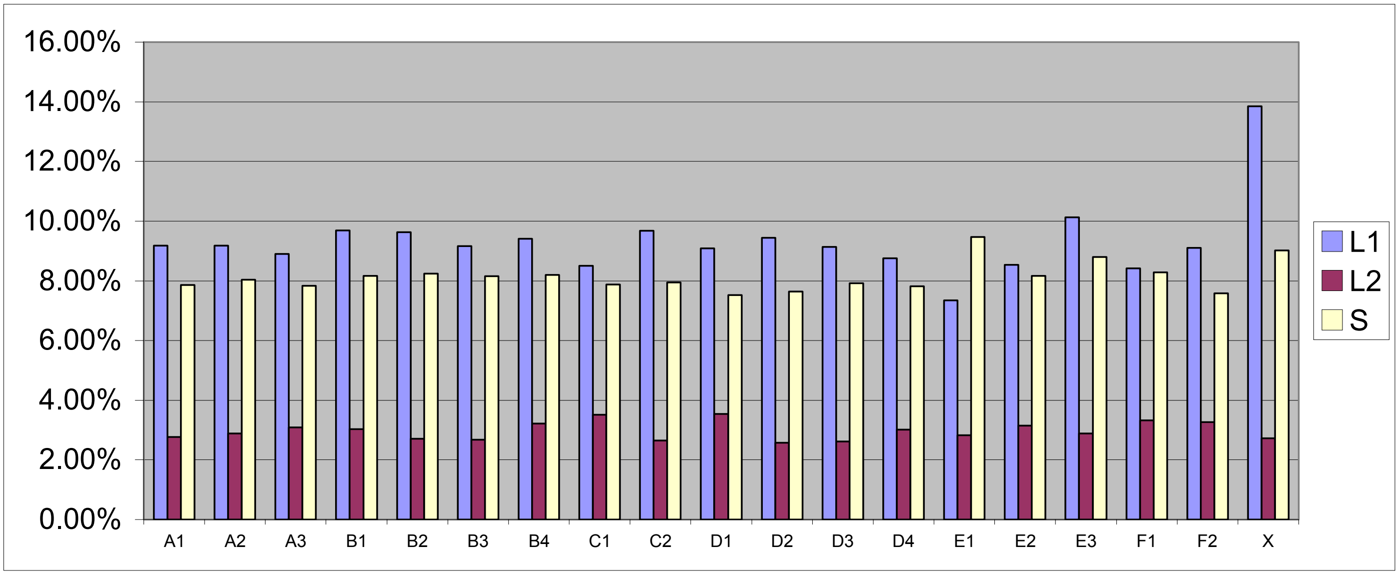
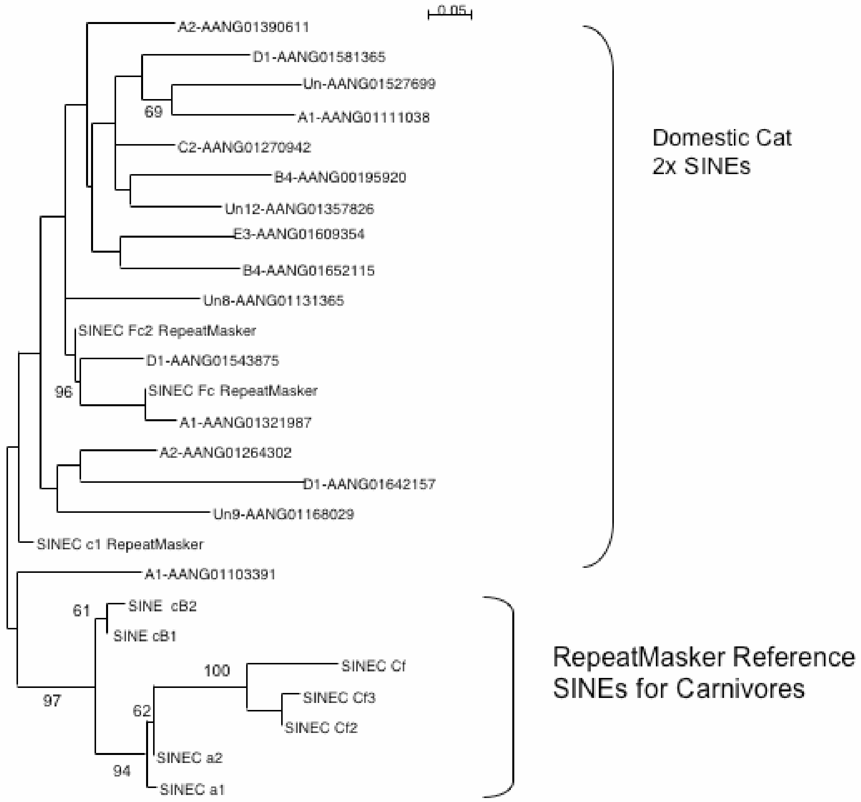


Figure S6 Percentage of sequence of LINE1, LINE2 and SINE elements on cat chromosomes based on final assembly

Figure S7


0.05
A2-AANG01390611
D1-AANG01581365
Un-AANG01527699
69
A1-AANG01111038
C2-AANG01270942
B4-AANG00195920
Un12-AANG01357826
E3-AANG01609354
B4-AANG01652115
Un8-AANG01131365
SINEC Fc2 RepeatMasker
D1-AANG01543875
96
SINEC Fc RepeatMasker
A1-AANG01321987
A2-AANG01264302
D1-AANG01642157
Un9-AANG01168029
SINEC c1 RepeatMasker
A1-AANG01103391
61
SINE cB2
SINE cB1
100
SINEC Cf
97
SINEC Cf3
62
SINEC Cf2
SINEC a2
94
SINEC a1
Domestic Cat
2x SINEs
RepeatMasker Reference
SINEs for Carnivores

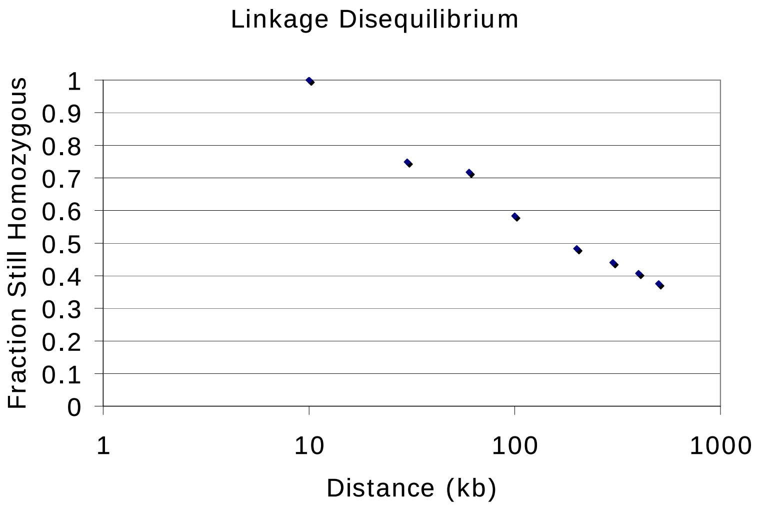


Figure S8. Extent of homozygosity conditional on homozygosity within a 10 kb region.