1 Introduction

The comparative felidae project aims to perform analysis of gene evolution, evolution of breakpoints and synteny, and also the analysis of selection signatures among the felines. Gene evolution analysis requires a good gene annotation including correct predictions of 5' and 3' UTR regions, exons and intron structure. We plan to use transMap from the domestic cat gene annotation to other felines. After an overview of the current available annotations of the domestic cat genome we came to a conclusion that we can improve those using Augustus prediction tool guided by the available cat mRNA data. The improvement of the cat's annotation would allow for more precise gene evolution analysis. In particular, inaccurate stop/start codon predictions as well as errors in splice cite positions and in particular lack of exons may cause artifacts when mapping genes through the species. Below we summarize the information about the available Felis catus gene annotation resources. We compare the results under consideration that the estimate of the feline gene number is 19000 - 20000, and the number of their alternative isoforms is ∼60000.

2 Available gene annotations

2.1 Annotation from *Pontius et al.*

It was based on the first draft of cat genome which consisted of the contigs that were aligned with the other available annotated mammalian genomes sequences. The gene predictions were based on pairwise RBM between the aligned sequences that were screened for annotated genes and gene features.

However this annotation was lost so that we were not able to evaluate its results. Anyway, this analysis incorporated data and methods that were available on 2007. Since that time the genome assemblies were improved, as well as mapped gene annotations were updated.

2.2 Annotation from *Tamazian et al.*

This annotation is based on the interspecies exon alignments. The genome features were derived from a comparative gene identification strategy using BLAST alignments between gene exons of the reference genome from eight reference mammalian gene maps (human, chimpanzee, macaque, dog, cow, horse, rat, and mouse) obtained from the Ensembl Gene 75 database. However the goal of this analysis was not the full gene annotation but the identification of the coding sequences. For this reason it lacks the prediction of exact gene structure that would be important for the gene evolution analysis. For this reason we dont included the numeric statistics obtained on this annotation as we do for the annotations discussed further.

2.3 Ensembl annotation

Ensembl gene annotation is based on protein, mRNA, EST, and RNAseq alignments and gene predictions by GeneWise and Exonerate. The cat mRNA and EST sequences were obtained from the GenBank database, the proteins were obtained from the UniProt database. There are about 200 cat proteins in the SwissProt manually curated compartment of the UniProt. However non-cat proteins were also used in these alignments. According to our estimation this is the best known gene annotation. However it has a significant number of errors among the predicted genes beginning with a few nucleotide imprecision (like a TT splice site 2 base pairs off from a regular AG splice site that was a conserved exon boundary in 15 of 16 other mammals) and ending with wrong prediction or loss of exons. The example of inaccurate gene prediction from this dataset is shown on Figures 1, 2, 3. We performed an automatic estimation of each reported transcript from this annotation for the following qualities: gaps in coding sequences, CDS and UTR showing unknown and non-canonical splice sites. The transcripts with the identified issues were called erroneous. Next we grouped the overlapping transcripts in order to identify the clusters corresponding to the gene locations. We distinguish the clusters for those that didnt contain errors in transcripts and those that contained only erroneous transcripts (1).

Montague et al. based their research on the ensembl annotation.

Genome browser examples (http://hgwdev.cse.ucsc.edu) of erroneous ensembl gene annotation instances for the FelCat 6.2 assembly.

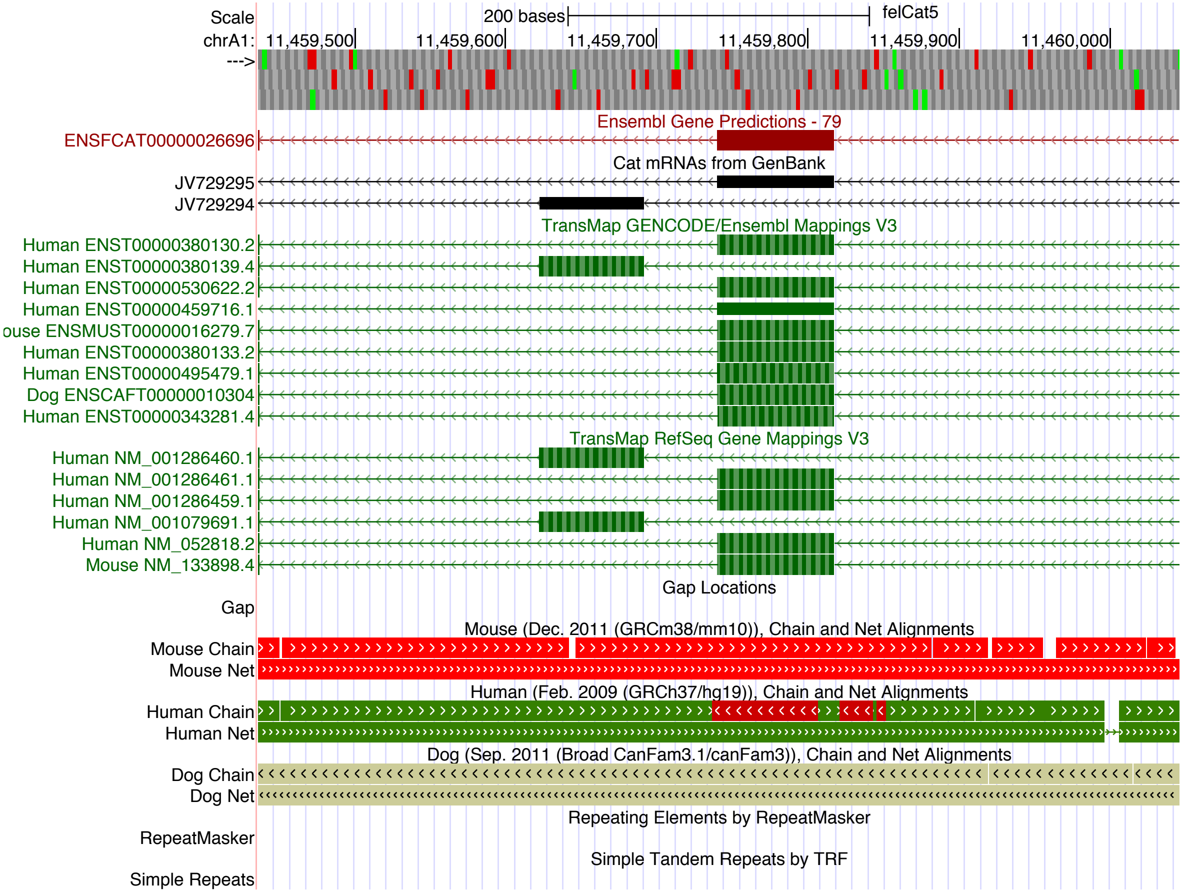

Figure 1: According to mRNA track there are two exons in the gene ENSFCAT00000026696, however the annotation lacks prediction of the first exon. TransMap from other species is able to predict that.

2.4 RefSeq annotation

refSeq gene prediction is based on the alignments of the NCBI RefSeq collection of the cat RNA to the reference cat genome.

3 Suggested approach

We suggest an incorporation of Augustus guided by mRNA data and TransMap approaches in order to create a newer annotation of the domestic cat genome. We will try to use mRNA obtained in RNA-Seq experiments for domestic cat, cheetah, and tiger in order to support Augustus prediction model with the hits. We will check any input data for quality before including it into the pipeline. Due to the modest size of the available mRNA data we are going also to include as evidence the TransMap of human gene annotation.

	Ensembl	RefSeq
Transcripts	22656	431
Gene clusters	19490	382
No-error clusters	40%	67.9%
All-error clusters	59%	30.8%
Other clusters	0.98%	1.0%

Table 1: Ensembl and RefSeq gene annotations quality.

We discussed this idea with Mario Stanke who agrees to support this work.

4 References

1. Pontius, Joan U., et al. ”Initial sequence and comparative analysis of the cat genome.” Genome research 17.11 (2007): 1675-1689.
2. Tamazian, Gaik, et al. ”Annotated features of domestic catFelis catus genome.” GigaScience 3.1 (2014): 13.
3. Montague, Michael J., et al. ”Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication.” Proceedings of the National Academy of Sciences 111.48 (2014): 17230-17235.)

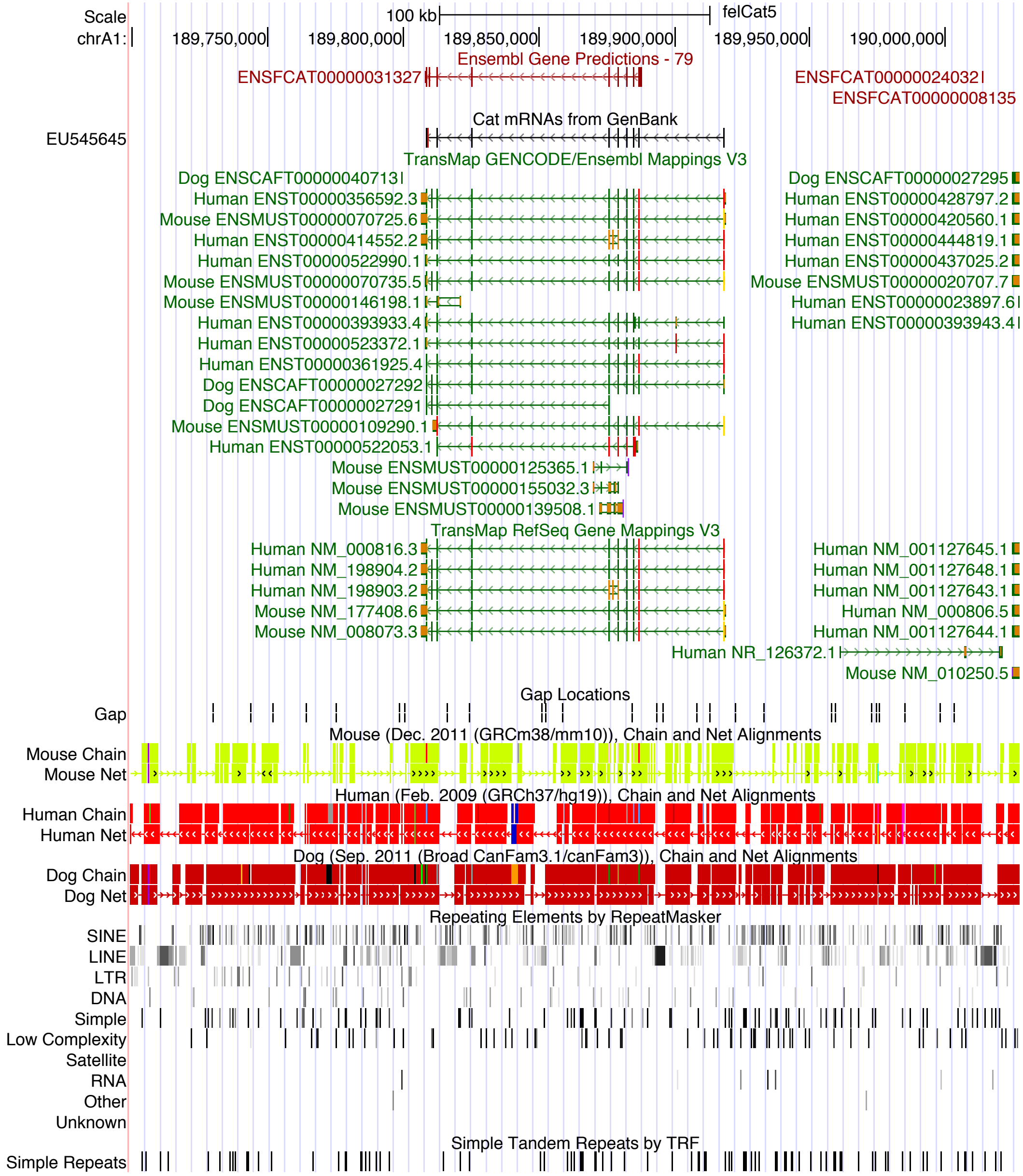

Figure 2: Another example of exon loss in gene ENSFCAT00000031327. Again, mRNA and TransMap are evidence of the coding region.

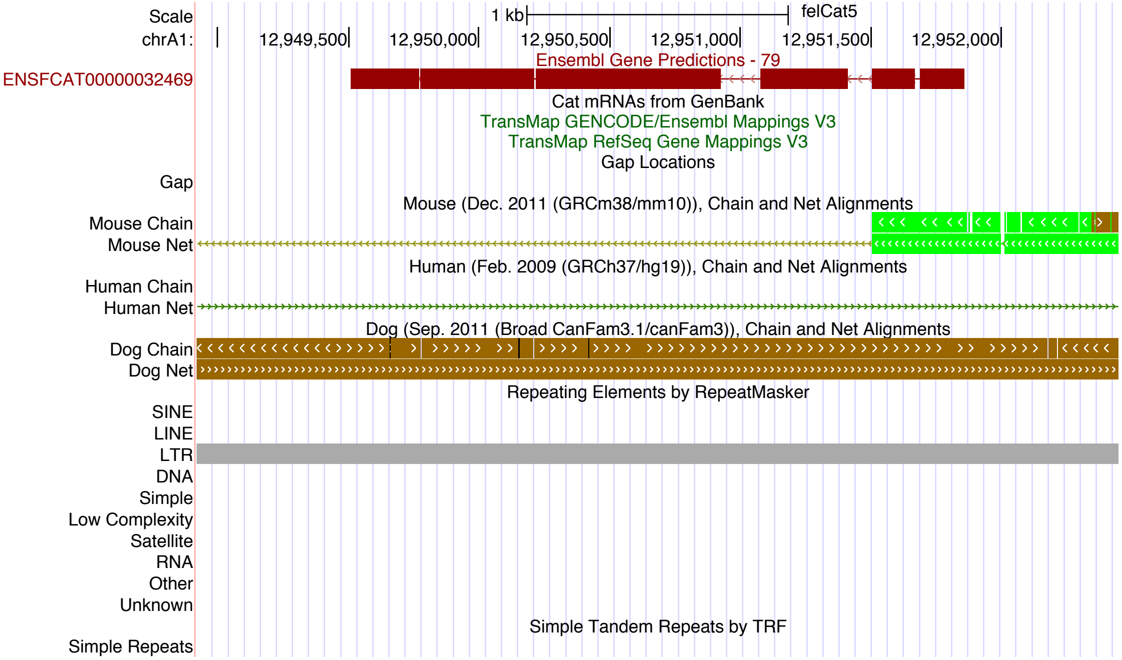
Scale
1 kb
felCat5
chrA1:
12,949,500
12,950,000
12,950,500
12,951,000
12,951,500
12,952,000
Ensembl Gene Predictions - 79
ENSFCAT00000032469
Cat mRNAs from GenBank
TransMap GENCODE/Ensembl Mappings V3
TransMap RefSeq Gene Mappings V3
Gap Locations
Gap
Mouse (Dec. 2011 (GRCm38/mm10)), Chain and Net Alignments
Mouse Chain
Mouse Net
Human (Feb. 2009 (GRCh37/hg19)), Chain and Net Alignments
Human Chain
Human Net
Dog (Sep. 2011 (Broad CanFam3.1/canFam3)), Chain and Net Alignments
Dog Chain
Dog Net
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Simple Tandem Repeats by TRF
Simple Repeats

Figure 3: Gene ENSFCAT00000032469 was predicted in the region annotated as LTR.