Apr 17th, 11:00 AM - 12:00 PM

Identifying Ancient Conserved Non-Coding DNA Elements

Jeremiah Lee Gaiser
University of Montana, Missoula, jg117416@umconnect.umt.edu

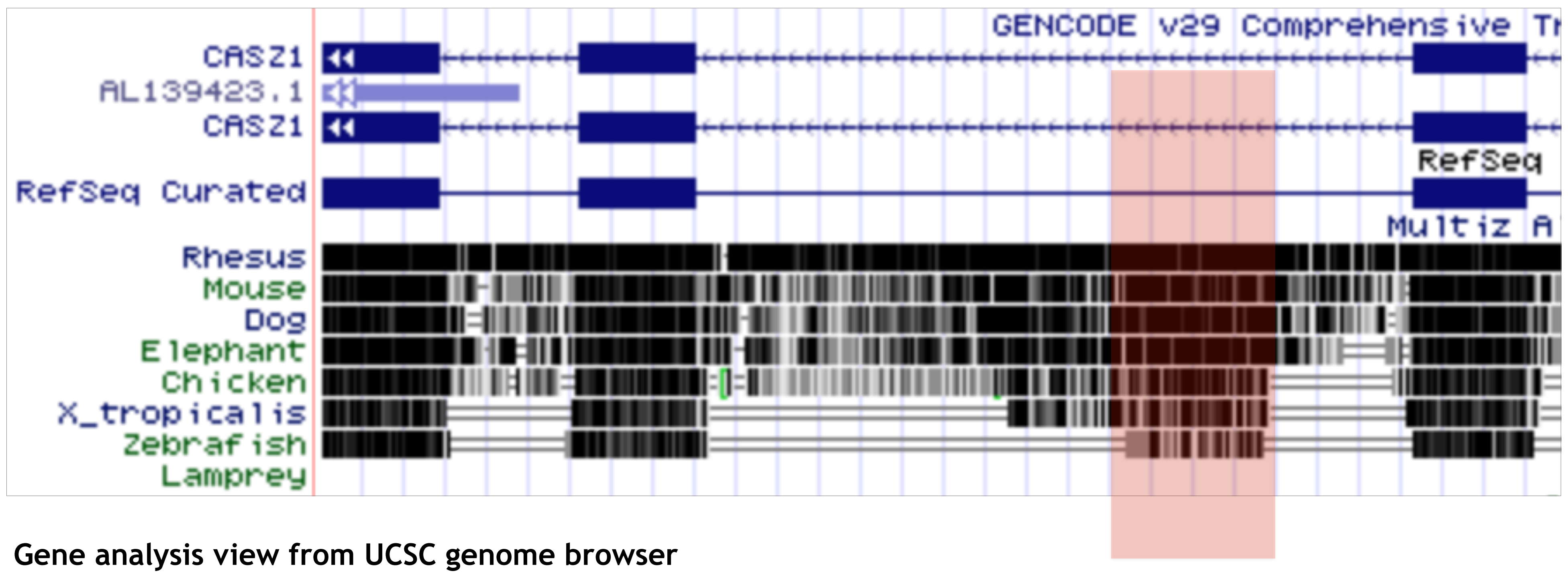
Identifying Ancient Conserved Non-Coding DNA Elements

Jeremiah Gaiser, Travis Wheeler
Dept. of Computer Science, University of Montana, Missoula, MT, USA

Conserved DNA Elements

DNA is the genetic material at the root of all life. It serves as the 'instructions' for the biomolecular machines that shape the bodies and synthesize the chemicals that comprise an organism. While DNA mutations and the forces of natural selection have resulted in the evolution of a tremendous diversity of species, there still exist many sequences of DNA that share remarkable similarity between organisms, even between species as different as humans and bacteria.

Here, we seek to understand DNA that remains highly conserved, perhaps over hundreds of millions of years, yet does not encode genes at all. The conservation of such DNA indicates some role that, while vital to species survival, remains to be understood. We employed open source computational tools and developed custom genomics analysis software to catalog these highly conserved non-coding sequences of DNA.



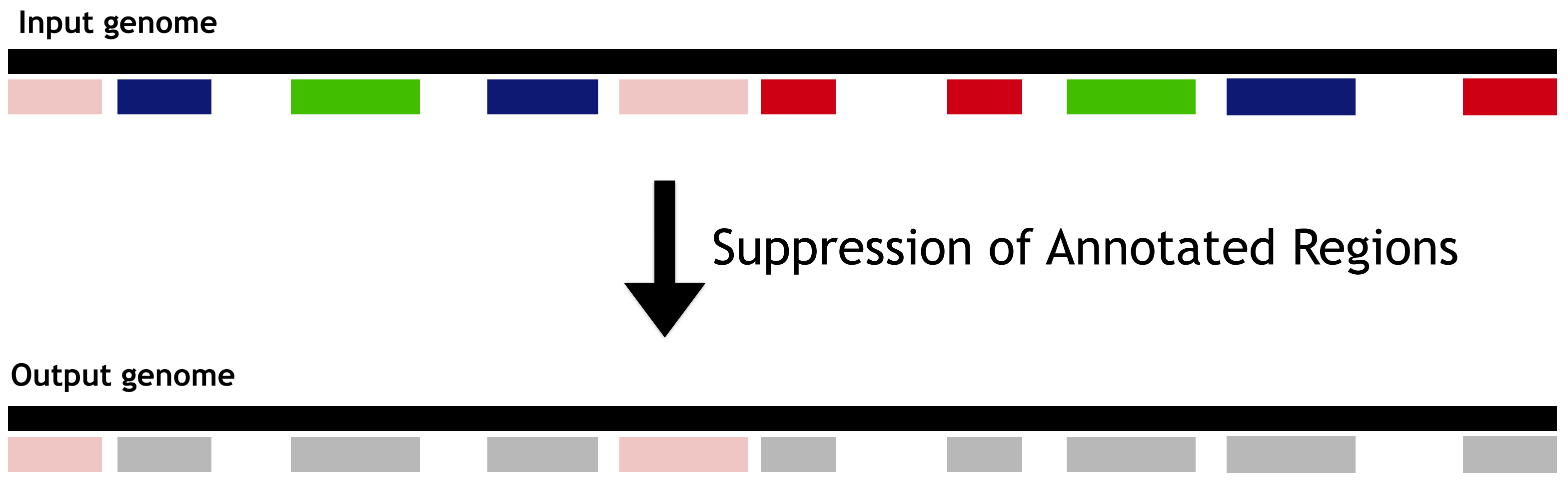
Gene analysis view from UCSC genome browser

■ Conserved Region 🟦 Exonic Region — Intronic Region ▨ Highly conserved non-coding region

Suppressing Protein and RNA Sequences

The first step in the software pipeline entailed finding and suppressing known protein and RNA expressing regions of DNA in the human genome (hg38). Using the alignment tool LAST[1], these coding regions were identified by finding significant local alignments between hg38 and gene/RNA sequences catalogued in databases such SwissProt and GtRNAdb. These annotated sequences were identified and subsequently masked. The result was the human genome absent of its known coding sequences. (This approach is in large part a reproduction of a software pipeline developed by collaborator Martin Frith.)

🟦 Protein-coding 🟥 RNA ▨ Masked
🟩 Pseudogene ▨ Conserved non-coding region



Here, a hypothetical region of a genome is depicted before and after being processed by the software pipeline. The colored blocks indicate regions of genomic elements. Masked regions are ignored by alignment software. Therefore, any alignments performed using the output genome will identify only conserved non-coding elements.

Aligning Human and Fugu fish Genomes

Using LAST alignment tool, the masked human and fugu fish genomes were sequenced to find significant alignments. The resulting hits represented the non-coding, anciently conserved sequences of interest.

```
  1 CTCTCCAGCGACAATAAAAAGAAACTTGAGTTTAAACAAAAAAAGTTACA    50
    |.||||.|.|||||||||.||||||||.||||||||| |||||||||.||
  1 CCCTCCTGTGACAATAAAGAGAAACTTCAGTTTAAAC-AAAAAAGTTCCA    49

 51 CCATATTTGCTCAGACTAACTATGATGAAAGGCAATGAAGACAAGGGCTC   100
    ||||||||||||||||||||||||||||||||||||||||||.|.|||.
 50 CCATATTTGCTCAGACTAACTATGATGAAAGGCAATGAAGACACGAGCTT    99

101 CTCATGTAGGTATCAATATAAATATTACTGGAAGGTCAATTAATATGTAA   150
    ||||||||||||||||||||||||||||||||||||||||||||||||||
100 CTCATGTAGGTATCAATATAAATATTACTGGAAGGTCAATTAATATGTAA   149
```

Alignment of a non-coding element shared between the human and fugu fish genomes

Post Processing

The hit locations were cross-referenced with the Genotype-Tissue Expression (GTEx) Database. This allowed us to retrospectively annotate any known coding sequences that the pipeline failed to mask, and identify proteins to which any highly conserved intronic sequences belong.

```
CHR     START   STOP    NAME            TYPE

chr1    135140  135895  RP11-34P13.15   processed_pseudogene
chr1    157783  157887  RNU6-1100P      snRNA
chr1    69090   70008   OR4F5           protein_coding
chr1    1173883 1197935 TTLL10          protein_coding
chr1    3658937 3668772 RP5-1092A11.5   antisense
chr1    9151667 9151777 MIR34A          miRNA
chr2    38813   46870   FAM110C         protein_coding
chr2    305110  314367  AC079779.5      lincRNA
chr2    692082  693235  AC092159.3      antisense
chr2    3017218 3017330 AC074264.1      miRNA
chr2    20606279        20606654        processed_pseudogene
```

A small example of the data provided by the GTEx Database.

Results

Our pipeline identified 13,338 unique conserved elements. Of these elements, 44% corresponded to intergenic regions without record in the GTEx database, while 44% were found to be within a known gene sequence. These are almost entirely non-expressing intronic regions of the gene. The remaining 12% were various flavors of pseudogenes and RNA, evidencing the need for improvement to our filtering pipeline.

Sequence Function	% of Hits
Intergenic	43.91
Protein Coding Gene	43.60
LincRNA	7.36
Antisense	3.18
Processed Transcript	0.67
Transcribed unprocessed pseudogene	0.25
Unitary pseudogene	0.21
Unprocessed pseudogene	0.16
Processed pseudogene	0.16
Transcribed processed pseudogene	0.15
Sense overlapping	0.15
rRNA	0.07
Misc RNA	0.04
snoRNA	0.03
miRNA	0.03
Sense intronic	0.02

Future Directions

Our work thus far serves as a jumping off point into a much larger effort to detect, characterize, and annotate anciently conserved non-coding DNA elements. Before doing so, however, the current software pipeline stands to be improved with the following steps:

- Identify and suppress uncaught RNA expressing sequences with sensitive RNA inference software
- Update post-processing to better characterize highly conserved intronic regions

Improving the pipeline will better serve future efforts to expand the project. From here, we will

- Develop annotation database of highly conserved noncoding regions
- Identify correlations between nonconserved elements and other genomic features in order to understand their function

1. Frith MC et al. Genome Research. January, 2011

Research funded by NIH, P20GM103546 (courtesy of a CBSD Junior Investigator grant)