Project: Annotate Your Chimp Sequences

In the last few labs, you've seen the basics of how to annotate genes and pseudogenes. You now have a chance to apply this knowledge to annotate some DNA sequences from chimpanzee.

What to Produce

The endpoint of the annotation process will be a map of all the features found within the sequences, along with a brief record of how the feature was annotated. You will be working in teams of three to annotate a given chunk of the chimp genome. The strategy will be to "divide and conquer;" after a group assessment, different members of the team should focus on different candidate genes. You will do a group PPT presentation and turn in individual papers. Here's what we would like to see:

1) At the beginning of the report show a figure with initial *Genscan* predictions; create a list of all gene-like features, including their endpoints in the sequence and a 1-2 line summary of what you think the feature is. The list should include genes and pseudogenes (though you may not be able to tell these apart with certainty). If your annotation refers to an existing GenBank or UniProtKB/Swiss-Prot record, be sure to include the accession number for this record. This list (or table) will present the combined efforts of the group.

Note that, even if you have multiple *BLAST* matches to a particular feature, you should give it just a single line. Your annotation should integrate the evidence from the various matches.

2) Follow the list with a concise but suitably detailed summary (paragraph-length) indicating what you know or believe about the feature. Give your conclusion as to whether it is a gene, a pseudogene, or something else, as well as showing what kind of evidence supports your hypothesis. If you can't make up your mind what the feature is, say so (but give evidence in favor of your favorite hypothesis). Combine results from *BLAST* with examination of the organization of the human and chimp genomes, plus any RefSeq mRNA alignments from *BLAT* on the Genome Browser. (Report your colleagues work in a brief summary, but do a detailed presentation of the evidence supporting your conclusions concerning the feature you annotate; figures showing key evidence are desirable.)

You should also annotate the longer transposable element relics, such as LINEs and retrovirus-like elements, found by *RepeatMasker*. These features provide good evidence that a particular part of the sequence has a known (if now inactivated) function. Don't bother drawing in all the Alus.

3) Produce a final map of the chunk, pooling results from team members. As discussed, work on these individually, but if you get stuck feel free to discuss with others regarding the proper interpretation of the data. Remember you will be graded mostly on the amount and type of evidence you find for any feature and how you interpret it.

Recommendations on How to Proceed

Start with the *Genscan* output as well as any significant *BLAST* hits to UniProtKB/Swiss-Prot and the human RefSeq RNA database. You could also perform *BLAST* searches against the non-redundant protein database (nr). As you find pieces worth investigating more closely, you can extract these pieces with the web-based *extractseq* tool if necessary and perform *BLAST* searches with the extracted sequence against the nucleotide databases. Follow up with an examination of the region in the human and/or chimp browsers.

For each feature, you should think about at least the following questions:

A) What family of genes or other sequences does the feature match?

B) Within a gene family, is there evidence as to which member gave rise to the feature? Remember to use evidence from mRNAs as well as protein to back up your answers.

C) Does the feature appear to match its human ortholog? Does it match at the right genomic location in human? Does it match human mRNAs as well as would be expected for an orthologous chimp sequence?

D) Is the match full-length at the protein level? At the DNA level? Are certain parts of the feature better conserved than others? Do you see evidence that is inconsistent with the feature producing a working version of its annotated protein (if any)?

E) What is your best guess as to the feature's boundaries? Can you identify the coding and untranslated regions of the feature using the mRNA and protein *BLAST* alignments?

F) When a single feature matches two or more discontiguous segments of the contig, are there repetitive elements in between them? (This can be good evidence of either an intron or a pseudogene, depending on other evidence, such as putative splice sites.)

G) How certain are you about your calls based on the available evidence? Remember that you have at your disposal *BLAST* itself, the HTML *BLAST* output, all its links to GenBank, *Genscan*, the genome browsers, and the UniProtKB/Swiss-Prot database and other databases. And, of course, there's always Google! If you're stumped as to what a particular feature might be, be creative in trying to gather evidence.

February 2014. Sarah C.R. Elgin