Comparative Genomic Studies of ORFan genes in Mammalian Genomes

The main objective of this project is to investigate the function/s, if any, of the ORFan gene sequences identified in previous studies from the literature, particularly those found in the PNAS article by Clamp et al 2007(*Distinguishing protein-coding and noncoding genes in the human genome*) in a comparative genomic study across human, chimp, and dog/mouse.

The goals will be to investigate:

(a) Whether the above organisms actually carry the same sequence/s (i.e., ORFan genes),
(b) Whether these sequences (or genes) are protein or only RNA forming, and as such, investigate;
(c) What each of these genes are being utilize for, by these organisms.

Good protocol of scientific study usually begins with the literature based research initially and then duplicating the methodologies used by a landmark study chosen from the literature study. As such a review of articles by Clamp et. al 2007, Christen et al 2011, Change 2015, and Tautz 2015 should be initially conducted.

The purpose of the Clamp (2007) article was to test whether the non-conserved human ORFs represent bona fide human protein-coding genes or whether they are simply spurious occurrences in cDNAs. Although it has been broadly accepted that ORFs with strong cross-species conservation to mouse or dog are valid protein-coding genes, the authors state that no work has addressed the crucial issue of whether non-conserved human ORFs are invalid. The authors further state that one must reject the alternative hypothesis that the non-conserved ORFs represent (i) ancestral genes that are present in our common mammalian ancestor but were lost in mouse and dog or (ii) novel genes that arose in the human lineage after divergence from mouse and dog.
However, in our studies we will tests these hypothesis, particularly the latter wherein we will test to seek any specificity if any in novel genes formed in the human lineage.

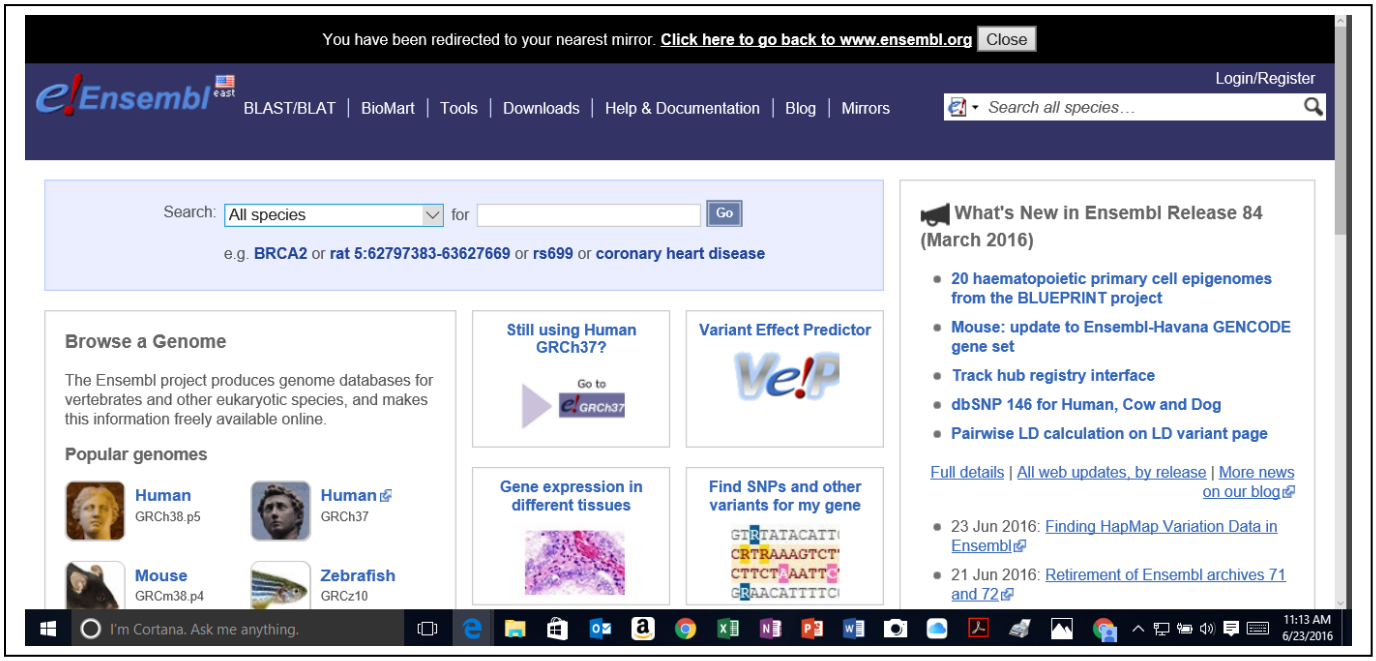
Analysis:
In the Clamp (2007) paper the analysis begins with a thorough reevaluation of available gene catalogs to identify conserved protein-coding genes and eliminate any putative genes resulting from those considered “clear artifacts”. They claim that they then studied the remaining set of non-conserved ORFs available at the time. They also claim that this analysis provides a scientifically valid methodology for evaluating future proposed additions to the human gene catalog.

Methods:
In this study the authors focused on the Ensembl catalog (version 35), which listed 22,218 protein-coding genes with a total of 239,250 exons. Their analysis had considered only the 21,895 genes on the human genome reference sequence of chromosomes 1–22 and X omitting the mitochondrial chromosome, chromosome Y, and ‘‘unplaced contigs,’’ which involve special considerations.

To do this the investigators had developed a computational pipeline protocol by which the putative genes are classified based on comparison with the human, mouse, and dog genomes using their specific "Pipeline" into which they submitted the sequences studied. Of the 21,895 genes they studied, using their algorithm (computational pipeline) they arrived at 1,285 ORFans.

Materials and Methods

All annotations for this study were based on the NCBI35 (hg17) assembly (2007) [current 2016 version depicted below with screen shot] and all genome alignments were taken from the pairwise BLASTZ alignment to mouse assembly NCBI36 (mm4) and dog Broad, Version 1.0 (canFam1; available from http://genome.ucsc.edu).



Objectives:

Although the authors assumptions and hypothesis may be different from ours, this is a landmark paper for our studies and investigations into the function of ORFan genes. The final goal will be to write a program (algorithm) such as the computational pipeline used in this study to be utilized for a genomic database systems study such as this. As such, a program could be developed with a novel algorithm (such as our "ORFanID") that can accomplish even more than what has been accomplished in this. In this project we will first "duplicate" this study in current databases in this Post ENCODE Project period which may easily provide more acceptance of alternative tested hypothesis and novel theories of significance.

References:

Clamp et al, Distinguishing protein-coding and noncoding genes in the human genome, *PNAS*, (2007), Vol 104:49, 19428 - 19433