Perfectly conserved sequences (PCS) between human and mouse are significantly enriched for exonic small proteins

Lucia Zifcakova[1], Md Abrar Jahin[1], Jonathan Miller[1]

[1]Physics and Biology Unit, Okinawa Institute of Science and Technology Graduate University, Japan

Introduction

The length distribution of perfectly conserved sequnces (PCS) follows -4 power law for sufficiently large L [1., 2.]. We hypothesized that the observed power law distribution of PCS reflects strong constraint on these sequences. The random PCS serve as a negative control.

Results

~95Mbp of both natural and random PCS, respectively:

7.7% (~7.3Mbp) of natural PCS were annotated as small proteins vs 0.53% (~16.5Mbp) small proteins in human genomic => natural PCS are 14x enriched in small proteins

0.7% (~ 0.6Mbp) of random PCS were annotated as small proteins => natural PCS vs random PCS are 11x enriched in small proteins

- 86% (~ 6.3Mb) of natural exonic PCS were small proteins
- 80% (~0.5Mbp) of random exonic PCS were small proteins
- 5.4% (~0.4Mbp) of natural intronic PCS were small proteins
- 10% (~0.07Mbp) of random intronic PCS were small proteins
- 7.9% (~0.5Mbp) of natural intergenomic PCS were small proteins
- 9% (~0.06Mbp) of random intergenomic PCS were small proteins

Conclusion

Natural PCS are 11x enriched in small proteins vs random PCS

Methods

We studied the whole-genome alignment of human and mouse (UCSC versions hg19 and mm10 2bit files from UCSC website, https://genome.ucsc.edu) for perfectly conserved nucleotide sequences (PCS). We classified PCS as "exonic" when their coordinates overlapped coordinates of an exon as defined in the RefSeq database gtf file (RefSeq (https://www.ncbi.nlm.nih.gov/refseq/); "intronic" when their coordinates overlapped a gene but not an exon, and "intergenomic" when their coordinates did not overlap any gene in the RefSeq database. For each of these three classes of natural PCS, we constructed a distinct cognate set of random PCS that preserved the histogram of that class of PCS as a function of length by randomly shuffling the coordinates of natural PCS within the corresponding parts of the genome for each class to create three sets of "random" PCS. We applied a lower PCS length (L) cutoff of 10bp for both natural and random PCS.

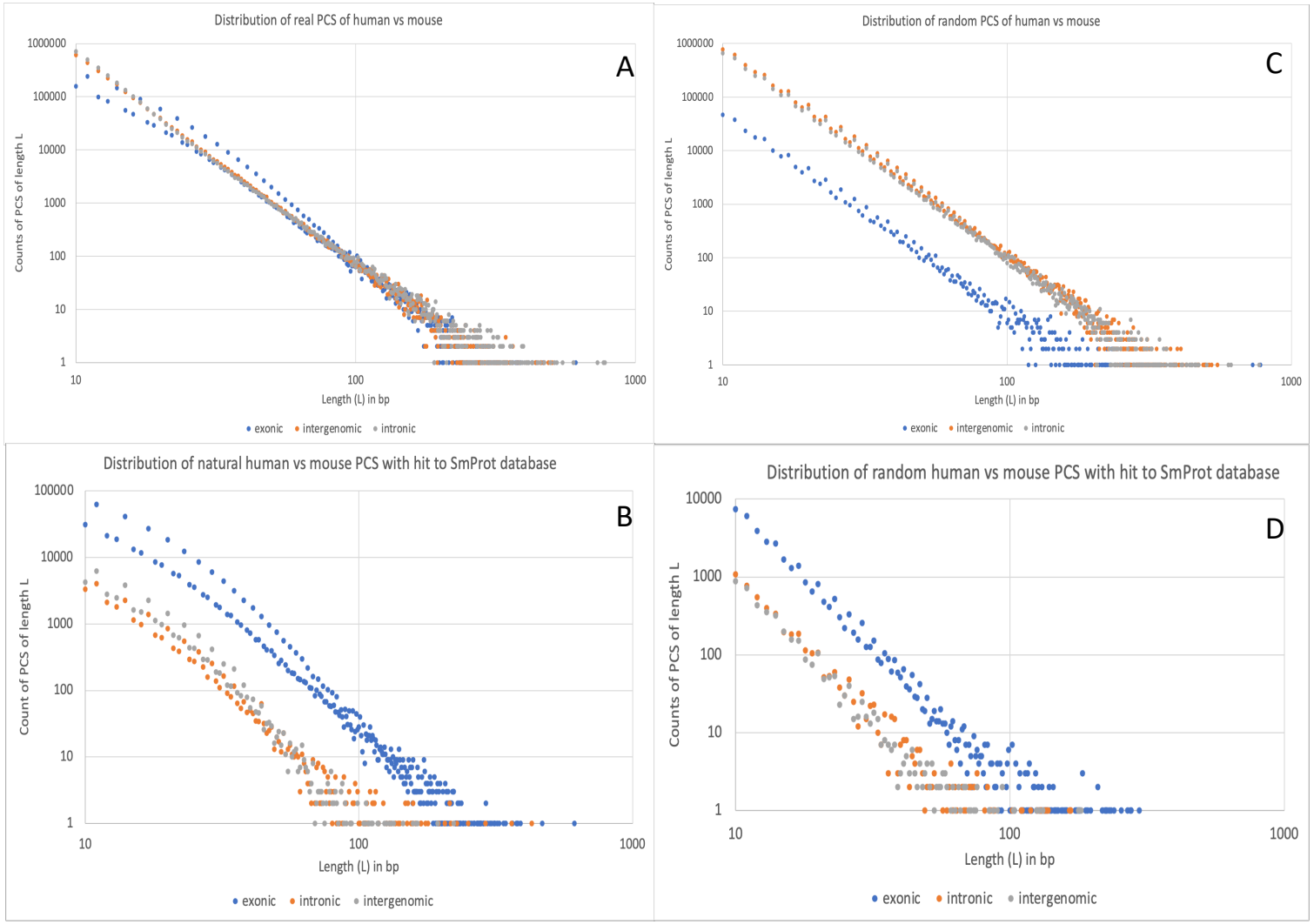

Figure 1. A – count vs length distribution of natural PCS annotated as exonic, intergenomic or intronic by RefSeq database; C - same as A for random PCS; B - count vs length distribution of natural exonic, intergenomic or intronic PCS with hit to SmProt database; D – same as B for random PCS.

1. Salerno, W., Havlak, P., Miller, J.: "Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments." *PNAS* 103, no. 35 (2006): 13121–25.
2. Massip, F., Sheinman, M., Schbath, S., Arndt, P.F.: How evolution of genomes Is reflected in exact DNA sequence match statistics, *MBE*, Vol. 32, Issue 2, Feb. 2015, 524–535,
3. Nash, A.J., Lenhard, B.: A novel measure of non-coding genome conservation identifies genomic regulatory blocks within primates, *Bioinformatics*, Vol. 35, Issue 14, Jul. 2019, 2354–2361