Perfectly conserved sequences (PCS) between human and mouse are significantly enriched for small-protein coding sequence

Lucia Zifcakova[1], Md Abrar Jahin[1,2]

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

[2]Department of Industrial Engineering and Management, Khulna University of Engineering and Technology, Bangladesh

Introduction

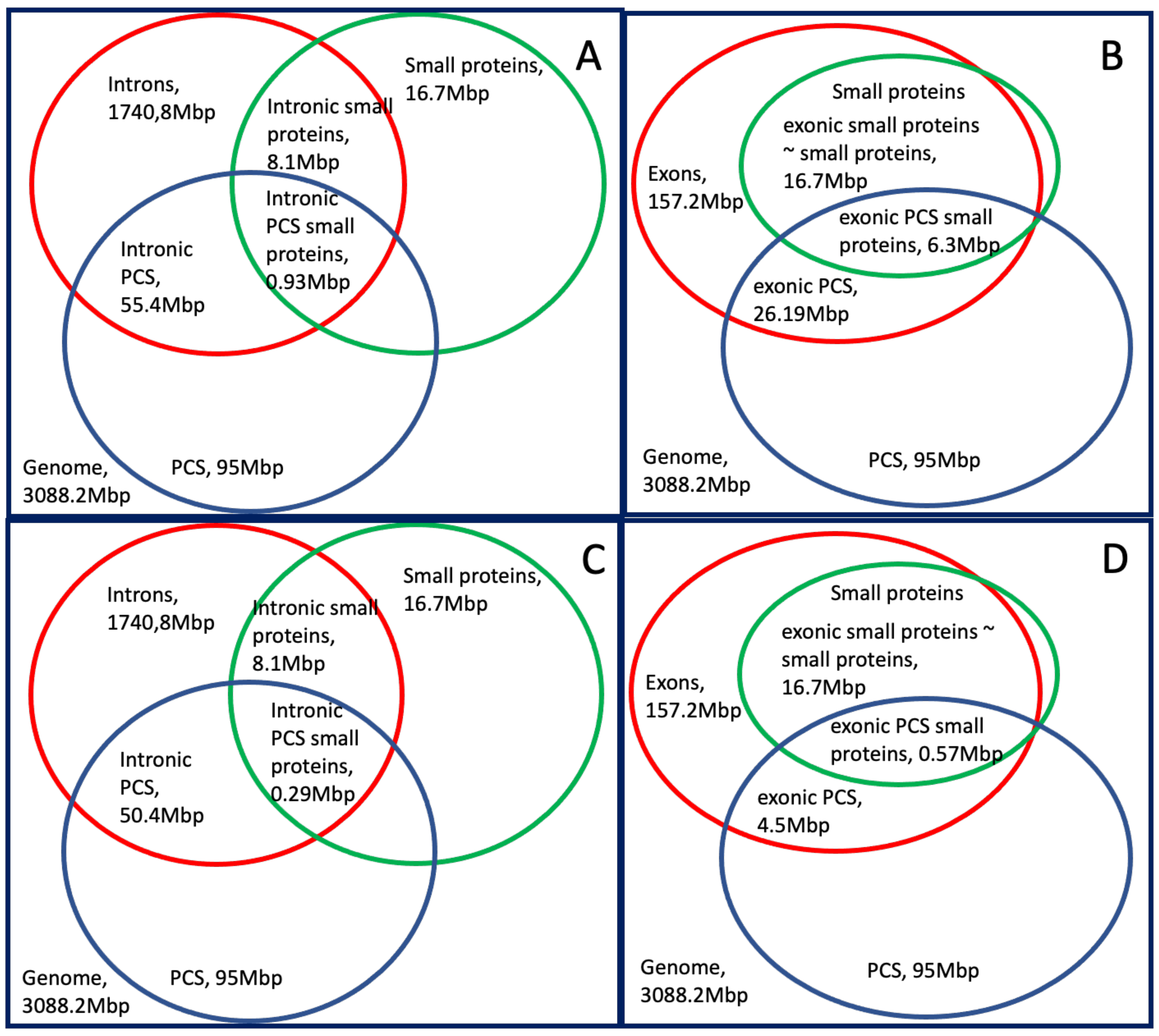
It has been observed that the lengths of perfectly conserved sequence (PCS) shared by a variety of genome pairs, such as human and mouse, follow a power-law distribution with exponent -4, down to lengths as short as 10 bases [1, 2]. In the absence of selection, an exponential length distribution would be expected, and it was conjectured that the power-law reflects selection on the sequences that compose it [1, 2]. We hypothesized that the observed power law distribution of PCS reflects strong evolutionary constraint on these sequences, which is reflected as enrichment of PCS in some specific functionalities.

Results

Exonic PCS were enriched in small proteins (Fisher's and hypergeometric test p value < 0.05) compared to non-PCS features. Intronic and random PCS were not enriched in small proteins

The natural exonic PCS of lengths of 20-300bp enriched in small proteins were significantly (p value < 0.05) associated with GO terms of human phenotype, such as e.g.: "Autistic behavior", "Delayed gross motor development", "Abnormal muscle tone", "Abnormal joint physiology", "Abnormal lip morphology". Both, random and natural exonic PCS annotated as small proteins were associated with few GO terms e.g.: "Strabismus", "Delayed speech and language development", "Neurodevelopmental delay". Almost 90% of natural intronic PCS annotated as small proteins were also annotated as known enhancers. Those PCS of length 20-30bp, even though not significantly enriched in small proteins, were associated with e.g.: "Abnormality of the hair" and of length 31-40bp with "Eczema" in GREAT enrichment analysis. Random intronic PCS annotated as small proteins were not associated with any human disease.

Figure 1. All quantities are in Mbp. A - natural PCS (blue circle), genomic introns (red circle), genomic small proteins (green circle), with corresponding overlaps quantified, human genome (blue square); B - same as A but for genomic exons and natural PCS; C - same as A but for random PCS; D - same as B for random PCS.



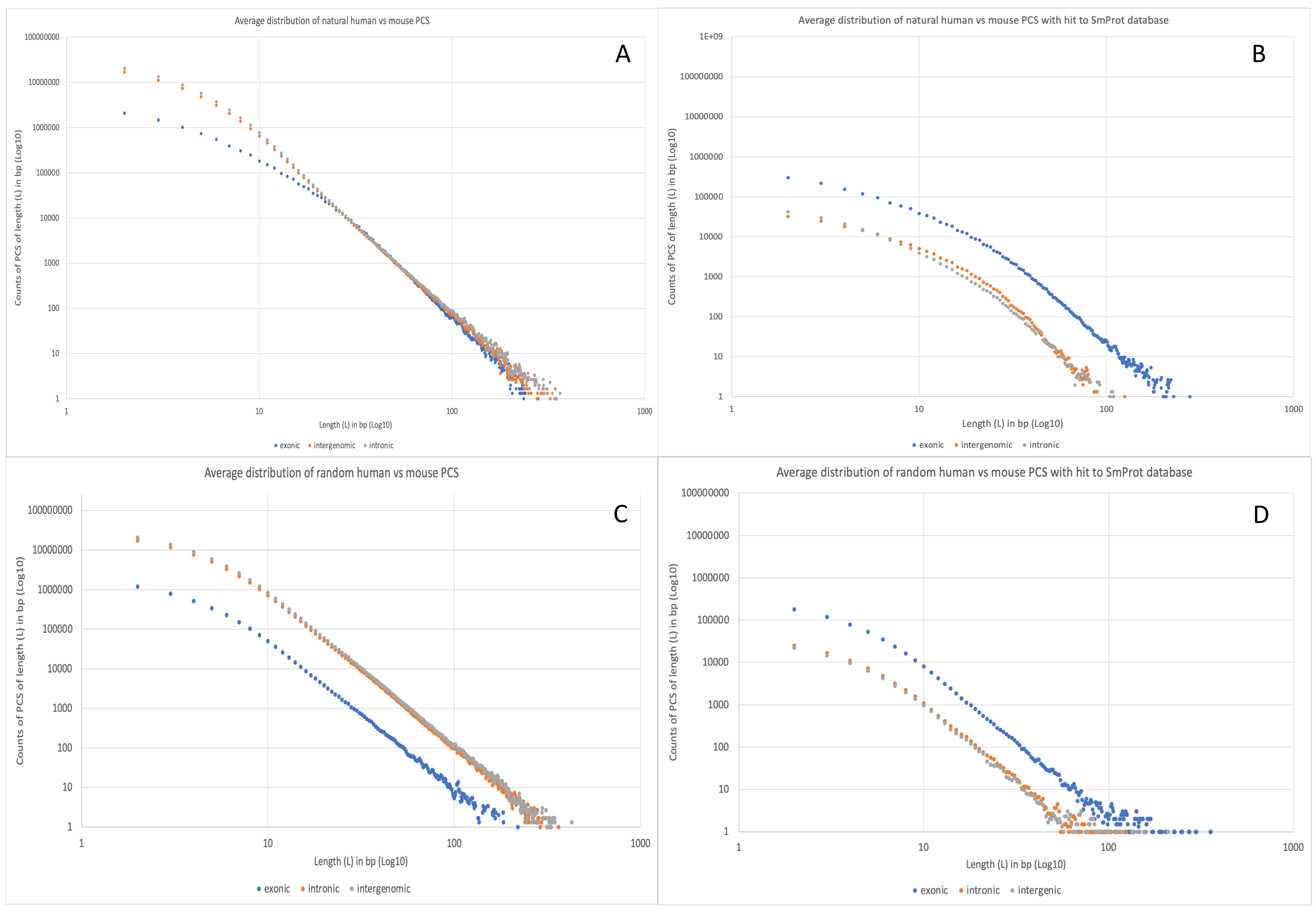
Conclusion

Natural PCS are enriched in small proteins compared to random PCS sets, suggesting selective evolutionary pressure for specific PCS function. In addition, natural exonic and intronic PCS annotated as small proteins are associated with human diseases in enrichment analysis.

Methods

We extracted PCS from the UCSC human and mouse genome alignment after removal of repetitive sequence. We leveraged RefSeq, SmProt and Enhanceratlas databases for PCS annotation by all known human genes, small proteins and enhancers, respectively. We have created 1000 sets of "random PCS" each with the same length distribution as natural PCS but randomly located in the non-repetitive part of the genome. To test for enrichment of small proteins in PCS we applied Fisher's exact test and the hypergeometric test, phyper, in R. gProfiler (for coding regions) and GREAT (for non-coding, cis-regulatory regions) was used to find enriched Gene Onthology (GO) terms in PCS annotated as small proteins. Both enrichment analyses use correction of p-value for multiple hypothesis testing.

Figure 2. There point averaged counts vs lengths distributions, blue dots are exonic, orange are intronic and grey are intergenic PCS. A - natural PCS; B - natural PCS annotated as small proteins; C - same as A for random PCS; D - same as B for random PCS.



Acknowledgments and References

This study was supported by the Physics and Biology Unit and OIST subsidy funding
We thank Prof. Miller for initial idea of the project and Dr. Biller for help with statistical methods.

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.