Pseudogenes are defined as disabled counterparts of functional protein coding genes. Depending on their formation mechanism, pseudogenes can be classified into three groups: processed (created through a retrotransposition event), unprocessed (resulting from a gene duplication event) and unitary (originating from functional genes through a loss-of-function event). Processed and unprocessed pseudogenes have a functional protein-coding counterpart (parent gene) in the same organisms. By contrast, unitary pseudogenes have no parent gene in their host genome, and are characterized by the presence of functional homologs in a syntenic region in a different species or organism. The fourth type of pseudogenes is called polymorphic. These are defined by the presence of a disabling mutation that is not fixed in the population. Specifically, polymorphic pseudogenes are regarded as protein coding genes that possess a loss-of-function variant in the reference genome.

Our lab has experience in annotating pseudogenes. Over the years, using a combination of manual curation and computational analysis, we have identified and characterized pseudogenes in human [1-3], worm [1], fly [1], zebrafish [1], primates [1], and a wide range of mouse species [4]. By comparing human, worm, and fly pseudogene complements, we found that the numbers of pseudogenes are not proportional to the genome sizes or the numbers of coding genes in the genomes, suggesting that pseudogene evolution is following a species-specific pathway. This specificity is also reflected in pseudogene types, where processed pseudogenes dominate over duplicated ones in human more than in the other species. This result correlates with the known burst of retrotransposition events at the dawn of primate lineage [1].

Pseudogenes have long been considered non-functional elements. However, the last decade has seen a large number of studies reporting on pseudogene activity, highlighting them as key players in genome regulation. Specifically, pseudogenes have been shown to regulate the expression of their functional protein coding homologs by serving as a source of siRNAs, antisense transcripts, microRNA binding sites, or competing mRNAs [5-7]. Moreover, a close inspection of pseudogene disabling mutations reveals that the pseudogenization process is closely related to the loss-of-function events such as premature truncation of proteins, disruption of splicing, and loss of functional or structural domains [8-10].

We used annotation and functional genomics analysis to study the activity of pseudogenes in various organisms. We conducted systematic analyzes of human, mouse, worm, fly and zebrafish pseudogenes focusing on large pseudogene families [4, 11, 12]. We found that the majority of pseudogene complements are species specific. Moreover, despite the presence of numerous disabling mutations, transcriptome analyzes suggest that a large proportion of pseudogenes are transcribed, and some are even translated. Using RNAseq data from ENCODE, modENCODE, and mouseENCODE [13-15] projects we analyzed the expression patterns of pseudogenes. Overall, we found that on average 15% of pseudogenes are transcribed in each organism, and moreover, pseudogenes show higher tissue transcription specificity than their protein-coding counterparts.

Over the years we have created a variety of pipelines, databases and online resources for pseudogene analysis. Specifically, we have built an online repository (pseudogene.org) that provides information regarding the annotation and characterization of pseudogenes in human and other model organisms. The repository is host of a large number of pseudogene databases, datasets, and resources such as the human pseudogene database psiDR [2], model organisms’ pseudogene resource, psiCUBE [1], and Pseudofam, the online pseudogene family database. Both psiDR and psiCUBE also provide information regarding evolutionary and functional characterization of pseudogenes in the curated genomes. Pseudofam focuses on clustering pseudogenes into families based on their functional homolog protein family recorded in Pfam. Currently there are 10 eukaryotic genomes in Pseudofam.

References:

1. Sisu, C., et al., *Comparative analysis of pseudogenes across three phyla.* Proc Natl Acad Sci U S A, 2014. **111**(37): p. 13361-6.

2. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.

3. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.

4. Balasubramanian, S., et al., *Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes.* Genome Biol, 2009. **10**(1): p. R2.

5. Kalyana-Sundaram, S., et al., *Expressed pseudogenes in the transcriptional landscape of human cancers.* Cell, 2012. **149**(7): p. 1622-34.

6. Swami, M., *Epigenetics: Demethylation links cell fate and cancer.* Nat Rev Cancer, 2010. **10**(11): p. 740.

7. Poliseno, L., *Pseudogenes: newly discovered players in human cancer.* Sci Signal, 2012. **5**(242): p. re5.

8. Abyzov, A., et al., *Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division.* Genome Res, 2013. **23**(12): p. 2042-52.

9. Balasubramanian, S., et al., *Gene inactivation and its implications for annotation in the era of personal genomics.* Genes Dev, 2011. **25**(1): p. 1-10.

10. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science, 2012. **335**(6070): p. 823-8.

11. Zhang, Z.L., P.M. Harrison, and M. Gerstein, *Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes.* J Mol Biol, 2002. **323**(5): p. 811-22.

12. Liu, Y.J., et al., *Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity.* BMC Genomics, 2009. **10**: p. 480.

13. Mouse, E.C., et al., *An encyclopedia of mouse DNA elements (Mouse ENCODE).* Genome Biol, 2012. **13**(8): p. 418.

14. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

15. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-D801.