Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Michael Tress
Email	mtress@cnio.es
Institution	CNIO
Website	https://bioinformatics.cnio.es
Group	Biocomputing Group

Project

Computational genomics

Project Title:

Anotación de los genomas de humano y ratón

Keywords:

Alternative splicing, genome annotation, proteomics, genetic variation, protein function

Summary:

Nuestro grupo forma parte del consorcio GENCODE1. El objetivo de GENCODE es anotar todas las características funcionales de todos los genes (protein coding, non-coding y pseudogenes) de humano y de ratón, con el fin de comprender mejor los dos genomas. GENCODE trabaja conjuntamente con Ensembl2 (la mayoría de los genes y isoformas en Ensembl son anotados por GENCODE) y con UniProt3. Además, forma parte del proyecto internacional ENCODE4. Nuestro trabajo dentro de GENCODE consiste en la validación computacional de los modelos de los genes anotados en el consorcio, centrándonos en los genes que codifican para proteínas. El trabajo se lleva a cabo con cuatro herramientas básicas: las propias anotaciones de Ensembl y UniProt, las anotaciones de estructura y función disponibles en nuestra propia base de datos, APPRIS5, el análisis de experimentos de proteómica6 y la variación genética de proyectos como 1,000 genomas7. Integrando múltiples experimentos de proteómica hemos sido capaces de validar la expresión de péptidos para el 60% de los genes humanos. Con la información derivada de estos péptidos y a las anotaciones de APPRIS, Ensembl y UniProt para cada gen, hemos estimado que el genoma humano tiene solo 19,000 genes protein coding8-10. Gracias a nuestros trabajos GENCODE ha recalificado mas que mil genes que antes se consideraban protein-coding. También hemos investigado la expresión a nivel de proteína de las isoformas alternativas de splicing actualmente descritas para el genoma humano. Nuestros resultados indican que la mayoría de los genes tienen una isoforma dominante11-13, que la gran mayoría de las isoformas alternativas se expresan en cantidades no detectables12-14 y que el tipo de splicing conocido como “mutually exclusive splicing” es el más conservado en términos evolutivos y también el más expresado al nivel de proteínas6,14,16. Tenemos dos herramientas para la predicción de isoformas alternativas, APPRIS selecciona las isoformas dominantes para cada gen, y TRIFID predice la importancia funcional de cada isoforma. El trabajo que proponemos requerirá comparar las isoformas de los genes codificantes en humano y ratón. Nuestras herramientas APPRIS y TRIFID pueden predecir las isoformas principales y las isoformas alternativas funcionales en las dos especies, y queremos saber si las isoformas son equivalentes en humano y ratón. Vamos a intentar ver si las predicciones de APPRIS y TRIFID están validadas utilizando datos de proteomica y de variación genética. Queremos cuantificar el nivel de conservación de las isoformas en otras especies y, cuando posible, su importancia clínica.

References:

1. Frankish A, et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47:D766-D773. 2. Cunningham F, et al. (2019) Ensembl 2019. Nucleic Acids Res., 47:D745-D751. 3. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158-D159. 4. ENCODE Project Consortium. (2020) Perspectives on ENCODE. Nature. 583:693-698. 5. Rodriguez JM, et al. (2018) APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res. 46:D213-D217. 6. Ezkurdia I, et al. (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol. Biol. Evol. 29:2265-2283. 7. 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. Nature. 526:68-74. 8. Ezkurdia I, et al. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 23:5866-5878. 9. Abascal F, et al. (2018) Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. 46:7070-7084. 10.

Humans May Have Fewer Genes Than Worms, http://www.popsci.com/article/science/humans-may-have-fewer-genes-worms 11. Ezkurdia I, et al. (2015) Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 14:1880-1887. 12. Tress ML, et al. (2017) Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci. 42:98-110. 13. Tress ML, et al. (2018) Most Alternative Isoforms Are Not Functionally Important. Trends Biochem Sci. 42: 408-410. 14. Abascal F, et al. (2015) Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Comput Biol. 11:e1004325. 15. Zahn LM. (2012) Isoform Identification. Science 336:520-521. 16. Abascal F, et al. (2015) The evolutionary fate of alternatively spliced homologous exons after gene duplication. Genome Biol Evol. 7:1392-1403.

Expected skills::

Good computing skills, specifically management of large-scale data sets

Possibility of funding::

No

Possible continuity with PhD: :

Yes