Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Sarah Djebali
Email	sarah.djebali@inserm.fr
Institution	IRSD, INSERM U1220
Website	www.en.irsd.fr
Group	Genetic and regulation of iron metabolism

Project

Computational genomics

Project Title:

Bioinformatics methods for the identification of enhancer/gene relationships in vertebrate genomes

Keywords:

enhancer/gene regulatory relationships; high-through functional sequencing data; chromatin structure; program evaluation; machine learning

Summary:

For many complex genetic diseases, the majority of the identified variants are located outside protein-coding genes [1], making it difficult to understand their function. And when variants are located far away from any gene, they are usually assumed to act on the nearest gene, which can often prove totally wrong [2]. The regulatory element that can explain this long distance action of the variant on the gene is the enhancer. Enhancers are genomic regions on which transcription factors bind, and which activate the expression of one or several genes by being brought close to (in 3D) the uptream regulatory elements (promoters) of those genes. Enhancers can therefore be far away from the genes they activate on the 1D genome, but being close to them in the 3D space of the nucleus. Today the best approaches to identify enhancer/gene relationships in the genomes are genetic screening [3] and targeted chromatin structure (3D), such as polymerase II ChIA-PET [4] or promoter capture HiC [5]. The problem is that the first one can only targets a handful of genes and the second one is very difficult and costly to generate. For this reason and because many international consortia such as ENCODE, FANTOM or Epigenome Roadmap have recently produced and made publicly available large quantities of functional 1D data (such as RNA-seq, ATAC-seq, histone marks or methylation data), the favoured approach is the integration of high-throuput functional 1D data. Although many programs exist to identify enhancer/gene relationships from functional 1D data [6,7,8], there is no consensus about what the best approach is. Here we would like to fill in this gap by assessing the different existing methods on reference sets and proposing a new method that uses a minimal amount of different data. The student will therefore have to: - Make a complete state-of-the-art of the existing 1D methods - Plan the evaluation * Define reference sets * Define criteria to include programs in the evaluation * Define the input data to use for each program to evaluate - Make the programs to evaluate work on small and real evaluation datasets - Determine the best approach and propose a new one that uses as few different input data as possible

References:

[1] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009

Jun 9;106(23):9362-7. [2] Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, Wei Y. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature genetics. 2017 Nov;49(11):1602. [3] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, Nguyen TH. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature Genetics. 2019 Dec;51(12):1664-9. [4] Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung WK, Ruan X, Ruan Y. ChIA-PET analysis of transcriptional chromatin interactions. Methods. 2012 Nov 1;58(3):289-99. [5] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics. 2015 Jun;47(6):598. [6] He B, Chen C, Teng L, Tan K. Global view of enhancer–promoter interactome in human cells. Proceedings of the National Academy of Sciences. 2014 May 27;111(21):E2191-9. [7] Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MT, Cheng C, Fan X, Gerstein M, Cheng AS. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature genetics. 2017 Oct;49(10):1428. [8] Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic acids research. 2019 Jun 4;47(10):e60-.

Expected skills::

Linux command line; Programming skills (bash, awk, python, ...); having already manipulated high-throughput (functional) sequencing data; know the basics of statistics and the R language; know how to run jobs on a cluster; understand written English well

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

PhD funding is not available yet but several options can be envisionned. This question also depends on the success of applications for funds that will be done during the fall.