Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor Albert Jordan

Email ajvbmc@ibmb.csic.es

Institution IBMB-CSIC

Website https://www.ibmb.csic.es/en/department-of-molecular-genomics-dmg/chromatin-regulation-of-human-and-viral-gene-expression/#lab-presentation

Group Chromatin regulation of human and viral gene expression

Project

Computational genomics

Project Title:

Occupancy of histone H1 variants genome-wide and consequences of altering H1 levels on human chromatin organization and gene expression.

Keywords:

genomics, proteomics, epigenetics, histone H1, chromatin

Summary:

We focus our research on the control of gene expression in human cells by chromatin organization, components and modifications. The degree of compaction of chromatin affecting a gene promoter dictates accessibility to transcription factors and RNA polymerase, and many chromatin modifying enzyme families act to overcome difficulties imposed by chromatin. We investigate the role and specificity of histone H1 variants in chromatin organization and gene expression control. By RNA interference of the different human H1 variants we have found that they have different involvement in cellular processes such as cell cycle progression and gene expression. Knock-down of multiple H1 variants induce the interferon response due to derepression of heterochromatic repeats and endogenous retroviruses. We have also described a differential role of H1 variants in pluripotency and differentiation. Currently, we are investigating the occupancy of H1 variants genome-wide by ChIP-seq (NGS) and the consequences of altering H1 levels on chromatin organization (ATAC-seq, DNA methylation, hiC chromosome conformation-TADs, etc), with an extensive use of Genomics and Bioinformatics. Additionally, we are performing proteomics of H1 variants specific protein complexes in chromatin and nucleoplasm to identify its interacting partners.

References:

https://www.ibmb.csic.es/en/department-of-molecular-genomics-dmg/chromatin-regulation-of-human-and-viral-gene-expression/#selected-publications

Expected skills::

Previous experience in molecular biology, genomics or computational will be positively considered.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Alexandre Perera Lluna
Email	teresa.tarrago@exheus.com
Institution	Exheus SL
Website	https://exheus.com/
Group	Bioinformatics Department

Project

Computational genomics

Project Title:

Development and validation of predictive models for the analysis of transcriptomic data based on artificial intelligence.

Keywords:

transcriptomics, artificial intelligence, RNASeq, deep learning models, health ,

Summary:

The activity of Exheus S.L. focuses on analysing gene expression in blood samples by using Next Generation Sequencing (NGS) technologies together with artificial intelligence algorithms to produce reports on the health status of its users based on the expression levels of their genes. The information provided is actionable so that users can make changes in their eating habits, physical exercise, etc. and thus optimise their quality of life. In a first stage of the company's development, the methodology has been applied to the world of sport. At Exheus, we have a deep understanding of the effects of sport on gene expression, but now we intend to pursue new lines of research and identify more biomarker profiles with which to characterise other problems of the human organism such as the ageing process, cardiovascular diseases or metabolic diseases. The main objective of this project is to use information provided by the massive next-generation sequencing technology RNAseq to create a definition of normalised gene expression in a population, develop an artificial intelligence model with supervised learning and thus be able to characterise the differences that may pose a risk to health at an individual level, with the study of the deviation of that sample with respect to the rest of the population. To carry out this task, the starting point is to select the numerical method with which to quantify the differences at a statistical level, establish a normal population parameter, then build and train an artificial intelligence algorithm to differentiate between the different characteristic profiles of each case study and finally evaluate the effectiveness of the model to predict the particularities of a single sample. In other words, the big difference in the data processing proposed by Exheus, with respect to current solutions, is the generation of algorithms that group the data to analyse "what is normal" in an individual, and differentiate it from "what is not normal" to provide real-time information on the changes that occur in any organism.

References:

1- PLoS One. 2017 Oct 13;12(10):e0180322. doi: 10.1371/journal.pone.0180322. 2- Sergi Picart-Armada, Wesley K. Thompson, Alfonso Buil, Alexandre Perera-LLuna, The effect of statistical normalisation on network propagation scores (2020), Bioinformatics; btaa896,[Link] 3- Marín-Llaó J, Mubeen S, Perera-Lluna A, Hofmann-Apitius M, Picart-Armada S, Domingo-Fernández D. MultiPaths: a python framework for analyzing multi-layer biological networks using diffusion algorithms. Bioinformatics. 2020 Dec 26:btaa1069. doi: 10.1093/bioinformatics/btaa1069. Epub ahead of print. PMID: 33367476. 4- A Lopez-del Rio, M Martin, A Perera-Lluna, R Saidi, Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction (2020), Scientific Reports 10 (1), 1-14 5- Elizabeth Carolina Jiménez, Claudia Avella-Garcia, James Kustow, Sally Cubbin, Montse Corrales, Vanessa Richarte, Flavia Lorena Esposito, Imanol Morata, Alexandre Perera, Paloma Varela, Jose Cañete, Stephen V Faraone, Hans Supèr, Josep Antoni Ramos-Quiroga, Eye vergence responses during an attention task in adults with ADHD and clinical controls, Journal of attention disorders (2020), 1087054719897806 Sergio Picart-Armada, Steven J Barrett, David R Willé, Alexandre Perera-Lluna, Alex Gutteridge, Benoit H Dessailly, Benchmarking network propagation methods for disease gene identification (2019), PLoS computational biology 15 (9), e1007276 6- S Kanaan-Izquierdo, A Ziyatdinov, MA Burgueño, A Perera-Lluna, Multiview: a software package for multiview pattern recognition methods (2019), Bioinformatics 35 (16), 2877-2879 7- Josep Lupón, Giovana Gavidia-Bovadilla, Elena Ferrer, Marta de Antonio, Alexandre Perera-Lluna, Jorge López-Ayerbe, Mar Domingo, Julio Núñez, Elisabet Zamora, Pedro Moliner, Evelyn Santiago-Vacas, Javier Santesmases, Antoni Bayés-Genis (2019) Heart Failure With Preserved Ejection Fraction Infrequently Evolves Toward a Reduced Phenotype in Long-Term Survivors, Circulation: Heart Failure 12 (3), e005652

Expected skills::

Good organizational skills. • Team player. • Drive and determination. • Desire to learn. • Enjoy solving problems; engaged and motivated. • Ability to communicate effective both verbally and in writing.Strong skills in data visualization tools. • Technical proficiency, scientific creativity, collaboration with others and independent thought. Good level of statistical programming (R, Python, SQL and others). managing, processing, performing quality control and analysis of large amounts of Next Generation Sequencing data. Fluent in English.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

We are looking for a talented Bioinformatics student to join our growing team of Data Scientists and to grow with us.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Iñaki Martin-Subero
Email	imartins@clinic.cat
Institution	IDIBAPS
Website	https://www.clinicbarcelona.org/en/idibaps/research-areas/oncology-and-haematology/biomedical-epigenomics
Group	Biomedical Epigenomics

Project

Computational genomics

Project Title:

Dissecting the single cell chromatin architecture of normal and neoplastic B-cells.

Keywords:

Bioinformatics, Single-cell, CLL, B-cells, Epigenetics

Summary:

Over the last years, our group has been characterizing the molecular features of normal B-cell subpopulations and pools of leukemic cells from chronic lymphocytic leukemia (CLL), the most frequent leukemia in Western countries. We have explored the epigenetic, genetic and transcriptional relationships during normal B-cell differentiation and upon B-cell malignancies, and identified potential drivers that can be linked to the origin, progression and aggressiveness of the disease. Now, the recent development of single cell technologies has opened up the possibility of providing a detailed characterization of individual cells with unprecedented resolution. Therefore, we aim to use primary samples from healthy and oncologic patients to generate a comprehensive single cell epigenetic map to study the different maturation states of B cells, and the regulatory mechanisms that govern their tumorigenesis, to decipher the cellular diversity and clonal architecture of CLL. The successful candidate will be part of the BCLL@tlas project (2018 ERC Synergy Grant) to study how the chromatin architecture is modulated during normal B-cell differentiation and upon neoplastic transformation by applying an integrative computational analysis of single cell Multiome data (scRNA-seq and scATAC-seq). She/he will learn how to use the main bioinformatic tools to analyze and interpret single cell data such as Seurat/Signat, MOFA+ among others. We provide the opportunity to work in a highly dynamic environment with top-level collaboration (CNAG-CRG, Holger Heyn) to empower the candidate towards deeper learning of single cell data analysis with the possibility to transfer basic research into clinically relevant knowledge. Furthermore, we offer the possibility to extend the master's project to a future doctoral thesis.

References:

- Kulis, M., Merkel, A., Heath, S., Queirós, A. C., Schuyler, R. P., Castellano, G., ... & Martín-Subero, J. I. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nature genetics, 47(7), 746. - Queirós, A. C., Beekman, R., Vilarrasa-Blasi, R., Duran-Ferrer, M., Clot, G., Merkel, A., ... & Martín-Subero, J. I. (2016). Decoding the DNA methylome of mantle cell lymphoma in the light of the entire B cell lineage. Cancer cell, 30(5), 806-821. - Beekman, R., Chapaprieta, V., Russiñol, N., Vilarrasa-Blasi, R., Verdaguer-Dot, N., Martens, J. H., ... & Martin-Subero, J. I. (2018). The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nature medicine, 24(6), 868-880. - Massoni-Badosa, R., Iacono, G., Moutinho, C., Kulis, M., Palau, N., Marchese, D., ... & Heyn, H. (2020). Sampling time-dependent artifacts in single-cell genomics studies. Genome biology, 21, 1-16. - Duran-Ferrer, M., Clot, G., Nadeu, F., Beekman, R., Baumann, T., Nordlund, J., ... & Martin-Subero, J. I. (2020). The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nature Cancer, 1(11), 1066-1081. - Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., ... & Heyn, H. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature biotechnology, 38(6), 747-755. - Vilarrasa-Blasi, R., Soler-Vila, P., Verdaguer-Dot, N., Russiñol, N., Di Stefano, M., Chapaprieta, V., ... & Martin-Subero, J. I. (2021). Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. Nature communications, 12(1), 1-18.

Expected skills::

- Degree in relevant discipline (eg. Computational Biology, Genetics, Bioinformatics, etc) - Experience in computer programming (R, Python) - Knowledge on genome-wide data set analysis, including ATAC-seq, RNA-seq, ChIP-seq and single cell data will be a valuable asset. - Strong motivation for planning and executing research projects. - Initiative to acquire new bioinformatics, statistical or programming skills. - Communication skills to allow the efficient collaboration within the group and across multiple institutions. - Good English level

Possibility of funding::

No

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Antonio Julià
Email	toni.julia@vhir.org
Institution	Vall Hebron Institut de Recerca
Website	www.vhir.org
Group	Grup de Recerca en Reumatologia

Project

Computational genomics

Project Title:

Single-cell RNA-seq study of Juvenile Idiopathic Arthritis

Keywords:

single-cell; RNA-seq; autoimmune; TCR; juvenile arthritis

Summary:

Background: single-cell analysis technologies are revolutionizing our understanding of chronic autoimmune diseases. Juvenile Idiopathic Arthritis (JIA) is the most frequent chronic rheumatic disease during childhood, and is a leading cause of disability in the short and long term. JIA encompasses different clinical entities, where oligoarticular JIA (oJIA) is the most prevalent. However, despite involving more than 50% of the juvenile patients, the pathology of oJIA is much poorly understood. Objective: to identify the pathogenic cell subtypes associated with oJIA. Methods: samples of inflammatory synovial fluid and blood from patients with oJIA will be obtained, as well as from healthy control infants. Using single cell RNA-seq technology (10xGenomics), the different cell subpopulations will be identified in both tissues and associated to disease and clinical phenotypes. This project will using cutting-edge technology in a fast evolving and exciting research area.

References:

“Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry.”. Nature Immunology 2019. Massoni-Badosa, Ramon, et al. "Sampling time-dependent artifacts in single-cell genomics studies." Genome biology 21 (2020): 1-16. Prakken, Berent, Salvatore Albani, and Alberto Martini. "Juvenile idiopathic arthritis." The Lancet 377.9783 (2011): 2138-2149.

Expected skills::

Fluency in R and/or Python. Desire to learn

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jose A. Seoane
Email	joseaseoane@vhio.net
Institution	Vall d'Hebron Institute of Oncology (VHIO)
Website	https://scholar.google.es/citations?user=mWl7l_kAAAAJ
Group	Cancer Computational Biology

Project

Computational genomics

Project Title:

Epigenetic role in aromatase inhibitor resistant breast cancer

Keywords:

Breast cancer, epigenetics, ATAC-seq, machine learning, transcription signatures

Summary:

Hormone receptor positive breast cancer patients has improved largely their outcome over the last decades, mainly due to efficacy of hormone receptor treatments (HRTx). However, some of these patients relapse after several years. In the MSK-Impact metastatic breast cancer study, authors identify some mechanism of HRTx, however mechanism affecting around 60% of metastatic breast cancer are still unknown (Razavi et al. Cancer Cell 2018). We have identified some subgroups that are more likely to relapse after 10 years (Rueda et al. Nature 2019), and we think that some epigenetic related mechanism could be affecting at least one of these subtypes. In this project we will leverage the dataset from (Selli et al. Breast Cancer Research 2019) in order to find the patterns that distinguish dormant from resistant tumors. We will investigate the epigenetic enrichments on these patterns. Then we will develop a signature to identify these dormant tumors in other datasets, including TCGA. We will use the TCGA ATAC-seq data to identify peak and motifs that are associated with dormancy based on our signature. Finally, we will correlate the tumors identified by our signature with time to relapse and with current breast cancer subtypes.

References:

Razavi, P., Chang, M. T., Xu, G., Bandlamudi, C., Ross, D. S., Vasan, N., ... & Baselga, J. (2018). The genomic landscape of endocrine-resistant advanced breast

cancers. Cancer cell, 34(3), 427-438. Rueda, O. M., Sammut, S. J., Seoane, J. A., Chin, S. F., Caswell-Jin, J. L., Callari, M., ... & Curtis, C. (2019). Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature, 567(7748), 399-404. Selli, C., Turnbull, A. K., Pearce, D. A., Li, A., Fernando, A., Wills, J., ... & Sims, A. H. (2019). Molecular changes during extended neoadjuvant letrozole treatment of breast cancer: distinguishing acquired resistance from dormant tumours. Breast Cancer Research, 21(1), 1-15.

Expected skills::

Differential expression analysis, pathway enrichment, machine learning, ATAC-seq differential peak and motif enrichment.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

There will be funding for one PhD position associated with this project.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Marco Mariotti
Email	marco.mariotti.mm@gmail.com
Institution	Universitat de Barcelona
Website	https://www.mariottigenomicslab.com/
Group	Comparative Genomics and Recoding lab

Project

Computational genomics

Project Title:

Computational genomics of proteins containing the special amino acid selenocysteine

Keywords:

Comparative Genomics; Evolution; Gene analysis; Translation; Recoding

Summary:

Our lab employs comparative genomics to study the mechanisms of gene expression and protein synthesis. We focus in particular on “recoding” events, programmed exceptions to the genetic code (1). A remarkable example of recoding is selenocysteine: this special amino acid is present in human and many other species, but it is not among the canonical 20 residues of the genetic code. Instead, it is encoded by the UGA codon, which is normally a stop, but it is recoded to selenocysteine through a highly regulated “readthrough” mechanism occurring only in specific mRNAs (2). Selenocysteine is found in the catalytic site of specialized enzymes, where it provides enhanced biochemical properties, typically for improved redox catalysis. Due to recoding, the genes encoding for selenocysteine-containing proteins (“selenoproteins”) are often missed or wrongly annotated in genomes, since gene annotation programs only consider the canonical role of UGA as stop (3). Selenoprotein genes in human have various well-known essential functions (4, 5), but a large part of the tree of life remains unexplored in this sense. The student may participate in three projects, related to the interest of the lab: • Development of automated approaches to recognize and correctly annotate selenoprotein genes in nucleotide sequences. In practice, the student will combine and improve programs for gene prediction and RNA motif finding (e.g. see (6, 7)). This is particularly important in context of new species being sequenced at unprecedented speed. • Evolutionary analysis of selenoprotein evolution. The student will apply tools from phylogenetics and sequence analysis to selenoprotein genes from diverse organisms, tracing how the selenocysteine utilization pathways changed throughout lineages (e.g. see (8–10)). • Analysis of selenoprotein function and regulation in disease. In practice, the student will make use of large public datasets of human or mouse data to analyse patterns of selenoprotein expression across tissues and disease conditions (e.g. see (11)).

References:

1. Rodnina,M. V, Korniy,N., Klimova,M., Karki,P., Peng,B.-Z., Senyushkina,T., Belardinelli,R., Maracci,C., Wohlgemuth,I., Samatova,E., et al. (2019) Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Res., 10.1093/nar/gkz783. 2. Labunskyy,V.M., Hatfield,D.L. and Gladyshev,V.N. (2014) Selenoproteins: molecular pathways and physiological roles. Physiol. Rev., 94, 739–77. 3. Santesmasses,D., Mariotti,M. and Gladyshev,V.N. (2020) Bioinformatics of Selenoproteins. Antioxidants Redox Signal., 33, 525–536. 4. Kryukov,G.V., Castellano,S., Novoselov,S.V., Lobanov,A.V., Zehtab,O., Guigo,R. and Gladyshev,V.N. (2003) Characterization of mammalian selenoproteomes. Science (80-.)., 300, 1439. 5. Mariotti,M., Ridge,P.G., Zhang,Y., Lobanov,A. V, Pringle,T.H., Guigo,R., Hatfield,D.L. and Gladyshev,V.N. (2012) Composition and evolution of the vertebrate and mammalian selenoproteomes. PLoS One, 7, e33066. 6. Mariotti,M. and Guigó,R. (2010) Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes. Bioinformatics, 26. 7. Mariotti,M., Lobanov,A.V., Guigo,R. and Gladyshev,V.N. (2013) SECISearch3 and Seblastian: New tools for prediction of SECIS elements and selenoproteins. Nucleic Acids Res., 41. 8. Mariotti,M., Santesmasses,D., Capella-Gutierrez,S., Mateo,A., Arnan,C., Johnson,R., D’Aniello,S., Yim,S.H., Gladyshev,V.N., Serras,F., et al. (2015) Evolution of selenophosphate synthetases: Emergence and relocation of function through independent duplications and recurrent subfunctionalization. Genome Res., 25. 9. Mariotti,M., Lobanov,A. V, Manta,B., Santesmasses,D., Bofill,A., Guigó,R., Gabaldón,T. and Gladyshev,V.N. (2016) Lokiarchaeota Marks the Transition between the Archaeal and Eukaryotic Selenocysteine Encoding Systems. Mol. Biol. Evol., 33, 2441–53. 10. Mariotti,M., Salinas,G., Gabaldón,T. and Gladyshev,V.N. (2019) Utilization of selenocysteine in early-branching fungal phyla. Nat. Microbiol., 4. 11. Avery,J. and Hoffmann,P. (2018) Selenium, Selenoproteins, and Immunity. Nutrients, 10, 1203.

Expected skills::

Python and/or R; Basics of gene evolution and phylogenetics; Expression data analysis.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Bolognesi Benedetta
Email	bbolognesi@ibecbarcelona.eu
Institution	IBEC
Website	https://ibecbarcelona.eu/protein-phase-transitions-in-health-and-disease/
Group	Phase Transitions in Health and Disease

Project

Computational genomics

Project Title:

Mapping Amyloid Nucleation by Deep Mutagenesis

Keywords:

Amyloid, Neurodegeneration, Combinatorial Libraries, Deep Mutational Scanning

Summary:

One common principle among all forms of dementia is the aggregation of specific proteins into insoluble amyloid aggregates that deposit in the brains of patients. These same proteins are also mutated in inherited forms of dementia. Our lab has developed an assay that allows to measure the ability of thousands of protein sequences in parallel to form new aggregates. In this project you will design, build, test and analyse massive mutational libraries (thousands of mutations) that encompass all possible mutations in the main proteins that aggregate in a number of neurodegenerative conditions (Alzheimer's, Parkinson's, ALS). The result of this approach will be mutational map(s) through which we are able to predict the impact of mutations in these proteins, i.e. to predict if people carrying those mutations are likely to develop dementia later in life or not. These maps will also be used to gather mechanistic insights on the process of amyloid nucleation, a fundamental biophysical process whose determinants are still unknown. The project combines experiments and computational analysis but it can be tweaked depending on your interest and ambition. Get in touch to discuss this!

References:

1. M Seuma, A Faure, M Badia, B Lehner, B Bolognesi* The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. Elife 10, e63364 (2021) 2. Bolognesi B*; Faure A; Seuma M; Schmiedel J; Tartaglia GG; Lehner B. The mutational Landscape of a Prion-like Domain. Nature Communications (2019) 10(1),4162.

Expected skills::

R, Python, Basic Molecular Biology (PCRs, gels)

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

The project is flexible, meaning that the % of time dedicated to experimental or computational work can be tailored to the student's interest.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Anthony Mathelier
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway, University of Oslo
Website	https://mathelierlab.com
Group	Computational Biology & Gene Regulation / Mathelier Group

Project

Computational genomics

Project Title:

PREDICTING MASTER REGULATORS DISCRIMINATING BREAST CANCER SUBTYPES FROM RNA-SEQ DATA

Keywords:

breast cancer; gene regulation; transcription factors; multi-omics

Summary:

An internship for a research Master student in the field of Computational Biology / Bioinformatics is available at the Computational Biology & Gene Regulation group, Centre for Molecular Medicine Norway, University of Oslo, led by Anthony Mathelier. The group develops computational methods and tools to analyze the regulation of gene expression and the mechanisms by which it can be disrupted in human diseases such as cancers. See https://mathelierlab.com/ for further information. The project aims at identifying transcription factors (TFs) that are discriminating breast cancer subtypes. Molecular classifications of breast cancer tumours provide intrinsic subtyping classification in addition to predictive and prognosis value, which can be helpful in assessing treatment options. The identification and understanding of such molecular classifications is paramount as they highlight biological aspects of tumour behaviour, function, and identity. Breast cancers are classically distributed across five intrinsic subtypes (luminal A and B, HER2-enriched, basal-like, and normal-like) based on gene expression profiling of 550 genes, mainly reflecting the estrogen receptor (ER) and HER2 status of the tumours. A 50-gene classifier (PAM50), approved by the US FDA (Prosigna test), was subsequently designed to assign tumours to luminal A, luminal B, HER2-enriched, or basal-like subtypes. This stratification has been useful to assess the likelihood of efficacy from neoadjuvant chemotherapy. As TFs represent special class of proteins that plays a key role in the transcriptional regulation of gene expression machinery, we hypothesize that their expressions and activities are contributing to the critical differences between breast cancer subtypes. While ER+ breast cancers are known to be driven by the TFs ESR1, GATA3, and FOXA1, the main drivers of ER- breast cancers are still unknown. In this project, the selected student will analyze large-scale omics data sets (e.g. from RNA-seq, ATAC-seq, and ChIP-seq) to highlight the TFs specifically regulating the genes differentially expressed between breast cancer subtypes. During the course of the project, the student will use state of the art computational approaches to predict where TFs bind to the DNA and methods to infer master TFs regulating the differentially expressed genes. Moreover, the work will expose the trainee to computational approaches for the management, analyses, and interpretation of large-scale, next generation sequencing data. We seek a highly motivated individual with programming skills, knowledge of computational tools development dedicated to the analysis of high throughput sequencing data. Knowledge in statistical methods and/or a biological background is a plus. We are looking for applicants excited about combining life sciences and computation to analyze gene expression regulation. The candidate will be collaborating with researchers at the Oslo University Hospital, where the group is also affiliated. The successful candidate will be collaborative, independent, with strong enthusiasm for research, and should be fluent in at least one of the following programming languages: Python, R, or bash. Being familiar with gene expression regulation in general and transcription factor binding is an advantage.

Expected skills::

Python, R, or bash

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Josep F. Abril
Email	jabril@ub.edu
Institution	Universitat de Barcelona
Website	https://compgen.bio.ub.edu/
Group	Computational Genomics Lab

Project

Computational genomics

Project Title:

Improving planarian interaction network with transcriptomes meta-analysis

Keywords:

planarian, transcriptomics, gene and protein networks, k-mer analysis, genome re-annotation

Summary:

In collaboration with the Barcelona Planarian Lab (UB), we are integrating transcriptomic data from different sequencing technologies into PlanNet/PlanEXP interaction network browser, as well as performing differential gene expression (DGE) analyses for planarian species to understand regeneration and developmental biological processes. With the advent of Single-Cell sequencing protocols, we have to improve the means by which molecular biologists can manage, analyze and interpret new gene expression data, as it has been done in PlanEXP. Up to the date, many RNA-seq experiments in planarian have been described and published, in which different transcripts may have been only expressed under specific conditions. We want to gather as much sequence information as possible from those data sets in order to distill an improved reference transcriptome sequence set, together with the latest genome assembly, which will have an impact on future DGE based on traditional RNA-seq or single-cell approaches, by improving gene structure resolution and better modeling genic features.

References:

PlanNet/PlanEXP (https://compgen.bio.ub.edu/PlanNET/)

Expected skills::

Candidate should have computational analysis skills using bash, unix tools, and scripting languages (perl, python, R). Further skills in database management or web development will be taken into consideration.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Depending on candidate curriculum and interests, there will be possibilities to apply for pre-doctoral fellowship to follow up with the project.

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

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Eulàlia de Nadal
Email	eulalia.nadal@irbbarcelona.org
Institution	IRB Barcelona
Website	https://www.irbbarcelona.org/en/research/cell-signaling
Group	Cell Signaling

Project

Computational genomics

Project Title:

Unraveling the intrinsic determinants of dynamic gene expression variability

Keywords:

Gene expression variability, single-cell RNA-seq, cellular stress

Summary:

Virtually all organisms interact with the environment. Sudden changes in the external environment, such as variations in nutrients, oxygen concentration or osmolarity, compromise cell fitness. To survive, cells have developed cellular responses ranging from metabolism tuning to gene expression reprogramming. Upon stress, Saccharomyces cerevisiae (yeast) undertakes a massive gene expression reprogramming involving at least 10 % of the genome. Despite being tightly regulated, gene expression changes differ between cells in a genetically homogeneous population, a phenomenon known as gene expression variability. This variability underlies relevant processes such as stress responsiveness, antibiotic and chemotherapy resistance or aging. The advent of single-cell RNA-seq (scRNA-seq) technologies allowed the quantification of gene expression variability at genome-wide level. Recently, we and others have developed a range of scRNA-seq methods (1-7) for interrogating single-cell yeast transcriptomes under several different conditions (e.g., oxidative, osmotic, and heat stress, galactose media, nitrogen limiting media or aging). For this project, we propose creating a yeast single-cell atlas (yscAtlas), a centralized resource comprising newly generated and published single-cell yeast datasets to profile the dynamics of gene expression variability. Leveraging on the yscAtlas, we will stratify all the yeast genes by their gene expression variability profile under changing environmental conditions. Based on this stratification, we will determine the biophysical and biochemical features associated with specific gene expression variability dynamics. To link specific gene features with gene expression variability dynamics, we will use state-of-the-art scRNA-seq data integration methods (e.g., MNNs, CCA), visualization methods (e.g., tSNE, UMAP), a set of complementary gene expression variability metrics and data modelling approaches.

References:

1. Gasch, A. P. et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLoS Biol. 15, e2004050 (2017). 2. Nadal-Ribelles, M. et al. Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations. Nat. Microbiol. 4, 683–692 (2019). 3. Saint, M. et al. Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation. Nat. Microbiol. 4, 480–491 (2019). 4. Zhang, Y. et al. Single-cell RNA-seq reveals early heterogeneity during ageing in yeast. BioRxiv (2020) doi:10.1101/2020.09.04.282525. 5. Jariani, A. et al. A new protocol for single-cell RNA-seq reveals stochastic gene expression during lag phase in budding yeast. elife 9, (2020). 6. Tsuyuzaki, H. et al. Time-lapse single-cell transcriptomics reveals modulation of histone H3 for dormancy breaking in fission yeast. Nat. Commun. 11, 1265 (2020). 7. Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R. & Gresham, D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. elife 9, (2020).

Expected skills::

We are looking for an enthusiastic master student to work in our group with basic knowledge of any programming language (preferably R or Python) and interest in functional genomics.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution

Project

Computational genomics

Project Title:

The genomic mechanisms of ichthyocarbonates precipitaion

Keywords:

fish, climate change, carbon cycle, genome, microbiome,

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The project Calcium carbonate released by teleost fish in marine environments (AKA ichthyocarbonates) represents one of the main carbon sinks in the open ocean, so a mitigator of climate change. The ichthyocarbonate pellets released by fish have an impact on the global carbon cycles and based on the most recent predictions of temperature increase and acidification as a result of climate change its importance will increase in the future. The formation of ichthyocarbonates in the gut of the teleost fish is also a key mechanism for the fish survival because allows them to maintain their osmotic balance. However, despite its physiological importance and its role as an alternative carbon sequestration method, little is known about the genomic mechanisms involved in the precipitation of ichthyocarbonates. Using genomics and transcriptomics data from the Gulf Toadfish (Opsanus beta), a model organism for the study of osmoregulation, such genes involved in the calcium carbonate precipitation have not been found neither in any other fish genome as far as we know. Classically it has been thought that the responsible for the precipitation of calcium carbonate was the fish, but recently microorganisms have been reported on the surface of the ichthyocarbonates opening the door to the possibility that the fish microbiota might be playing a role in this process. So, it is possible that the genes directly involved in the precipitation of ichthyocarbonates are present in the microbiome. The aim of this project is to characterize the complete mechanism of ichthyocarbonates precipitation targeting at the same time the piscine host and its microbiome. We will compile a set of reference genomes of teleost fish and re-analyze them using alternatives approaches that would allow us to obtain a better assembly to minimize the loss of information and to assemble the genomes of the most abundant associated microbes using binning strategies on the “contaminant” fraction of the raw genomic data. We hope that using this strategy will allow us to reconstruct the complete carbonate precipitation pathway.

References:

Wilson, R. W. et al. 2009. Contribution of Fish to the Marine Inorganic Carbon Cycle Science 323, 359–362.

Expected skills::

R, Python, Genome Assembly and Annotation, Phylogeny, Binnig Strategies, Database Management

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)

Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution

Project

Computational genomics

Project Title:

A Ribosomal Operon Reference Database

Keywords:

rrn, metabarcoding, eDNA, microbiome, biomonitoring

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The proposed project Metabarcoding has been for many years a useful approach to study the diversity and distribution of micro and macroorganisms across environments. Furthermore, metabarcoding is currently being implemented successfully as a biomonitoring tool. This methodology is used for diagnosis of microbial pathogens, to study the health of lakes, rivers and beaches, to track the presence of invasive or endangered species, etc. However, the current metabarcoding methodologies present certain limitations, being the most significant the lack of phylogenetic resolution. The most popular metabarcoding approach is the use of short read barcodes generated using Illumina. These fragments, that are commonly not longer than 400 bp, despite providing very useful information cannot reach the level of detail that would allow us to infer from them species or strain identities (the latest in the case of microbes). We propose the use of the whole rRNA operon (rrn) as a barcode for life. Many fragments of the rrn such as the 16S and 23S in bacteria, the 18S, ITS1, ITS2 and 28S in eukaryotes, or fragments of them, are commonly used as barcodes. By using the rrn we are using a barcode that is many times longer than the current barcodes and that also includes many of them. So, it does not have only the advantage of providing more phylogenetic resolution but also allows to bring previous information generated using other rrn derived barcodes under the same phylogenetic and taxonomic framework. In order to stablish the rrn as a barcode the first thing we need to generate is a reference database. As we are just starting to generate now the first rrn amplicons using third generation sequencings (Nanopore, PacBio) we still do not have access to this type of data to generate such a reference database. However, genomes and metagenomes can be sources of rrn that can be used as references after placing them in a phylogenetic tree in order to assign them an identity. We propose to use extracted rrn from publicly available genomes and metagenomes and build a phylogenetically aware reference database using R and MySQL.

References:

Guillou, L. et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597-604 del Campo, J. et al. (2018) EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLOS Biol. 16, e2005849 Jamy, M. et al. (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity Molecular Ecology Resources 20, 429–443

Expected skills::

R, HMMER, Python, MySQL, phylogeny

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution.

Project

Computational genomics

Project Title:

The Microeukaryotic Virome

Keywords:

Virome, Giant Viruses, Microeukaryotes, Viral Endogenization

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The proposed project Virus have been reported as a significant component of the nuclear genomes of different microeukaryotes from algae to heterotrophic protists. These viruses have been proved to be relevant for different aspects of microeukaryotic biology, shaping the genome of their algal host or protecting the host from other viral infections. The number of viruses described from microeukaryotes is relatively low compared to those infecting bacteria or macroorganisms. The microeukaryotic virome is a source of novel viral diversity, particularly of giant viruses. The aim of this project in collaboration with Professor Richard A. White from the University of North Caroline Charlotte is to characterize the viral landscape of the unicellular eukaryotes. Initially we will build a comprehensive database of microeukaryotic genomes and transcriptomes. Using this dataset, we will proceed to extract the viral signal from the different organisms' genome and proceed to their characterization using phylogenetic trees. We expect in this project to unveil a significant amount of viral diversity. As a byproduct of it we will also generate a comprehensive microeukaryotic genomic database.

References:

Fischer, M. G. et al. (2016) Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540, 288–291 Moniruzzaman, M. et al. (2020) Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588, 141-145

Expected skills::

R, Phyton, Genome Analysis, Phylogenies, Database Management

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	marta melé
Email	marta.mele.messeguer@gmail.com
Institution	Barcelona Supercomputing Center
Website	https://www.bsc.es/discover-bsc/organisation/scientific-structure/transcriptomics-and-functional-genomics-lab-tfgl
Group	transcriptomics and functional genomics lab

Project

Computational genomics

Project Title:

Understanding human individual variation in splicing

Keywords:

Transcriptomics, differential gene expression, human populations, splicing, ribosome profiling, posttranscriptional processing, RNA binding proteins.

Summary:

Summary: The candidate will join Marta Melé's Transcriptomics and Functional Genomics lab in the Life Sciences Department at the Barcelona Supercomputing Center. The lab is interested in understanding how individual variation in gene expression and splicing profiles can explain phenotypic differences between individuals both in the context of health and disease. To address this question, we use large-scale transcriptomic analysis and the latest single-cell sequencing technologies combined with the development of novel methods to study gene expression, splicing and cell type composition variation across human tissues and phenotypes. In this project, we will perform a large-scale analysis of splicing variation between individuals with different phenotypes and from different ethnic groups. In previous studies, we observed that in certain contexts splicing varies more between individuals than between tissues. Also, we have found that ancestry contributes more to explain splicing differences between individuals than other traits such as age. Remarkably, we observe that ribosomal proteins have strikingly large splicing variation between individuals of different ancestries. This pattern could have functional consequences for the translation machinery that we will explore further. Ultimately, the question that we want to tackle in this project is what is the role of splicing in determining why human individuals are different from one another. What you will learn: Development of computational pipelines to analyse and interpret large omics datasets such as RNA-Seq, single-cell RNA-seq, ribosome profiling, and CLiP-seq. Working in a High Performance Computing environment. Scientific collaboration in the context of international consortia, effective communication of research findings in internal and external meetings, scientific writing, and critical thinking. Also the master student will join the Melé lab journal clubs, lab meetings and lab lunches to talk about science but also have fun and discuss non-science related topics with the group.

References:

Melé, M. et al. The human transcriptome across tissues and individuals. Science (80-.). 348, 660–665 (2015).

Expected skills::

Strong programming skills in bash, python, R, perl, or similar. Excellent communication skills in spoken and written English. Capacity to contribute to research projects with novel research ideas and analysis. Capacity to work as a team in a highly collaborative and diverse environment. Experience working in HPC clusters will be appreciated. Experience with Next Generation Sequencing data analysis will be appreciated. Availability to start in July 2020 is preferred.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Camille Lassale
Email	classale@imim.es
Institution	IMIM
Website	
Group	Cardiovascular Epidemiology and Genetics (PI: Roberto Elosua)

Computational genomics

Project Title:

Diet quality, DNA methylation and transcriptomics: an integrated translational approach

Keywords:

Nutrition; Epigenetics; DNA methylation; EWAS; Mendelian Randomization

Summary:

The aim of this project is to discover novel epigenetic biomarkers modified by lifestyle, that can identify patients at elevated cardiovascular disease (CVD) risk and act as new therapeutic targets. CVD is a major cause of death worldwide and cost the European Union economy over €210 billion per year. Prevention strategies are targeted at patients at increased risk of future CVD, imperfectly identified by CVD risk prediction algorithms that need improvement. Furthermore, the biological mechanisms through which diet and physical activity act on cardiometabolic health (obesity, diabetes, dyslipidemia) are not fully understood. The epigenome, which controls the differential expression of genes, is both heritable and modifiable by the environment, but little is known on how diet influences epigenetic mechanisms. In this project, we will use independent discovery and validation population cohorts from different countries (Spain, USA) with dietary, lifestyle and clinical data, as well as cutting-edge multi-omics data: GWAS, epigenome-wide DNA methylation, and transcriptomic (gene expression). Using an epigenome-wide association study (EWAS) design, we will assess the effects of food, nutrients and overall dietary patterns on DNA methylation, discovering novel diet-related methylated loci (CpG sites). We will then evaluate the impact on gene expression, and if diet- and physical activity-related CpGs causally relate with CVD and risk factors, in particular obesity, dyslipidemia and diabetes. Finally, the added predictive value of these biomarkers over established CVD risk prediction scores will be evaluated. Complex novel analytical approaches will be used, including integration of multi-omics data through meta-dimensional and multi-staged analysis, and Mendelian Randomization to assess causality. This multidisciplinary project, combining epidemiology, nutrition, genetics and omics, will provide new insight in the aetiology of CVD and identify novel predictors to improve CVD precision medicine.

Expected skills::

Advanced R programming; basic knowlegde in epidemiology; interest in nutrition

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

Co-supervision with Dr Elosua

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Silvia G. Acinas & Pablo Sanchez
Email	sacinas@icm.csic.es
Institution	Institute of Marine Science (ICM) CSIC
Website	https://www.icm.csic.es/en
Group	Ecology and Genomics of Marine Microbes

Project

Computational genomics

Project Title:

Revealing active prokaryotic genomes from the global Deep Ocean microbiome.

Keywords:

Malaspina Expedition, deep ocean, metagenomes, metatranscriptomes, metagenomic assembled genomes, web development, functional analyses, comparative genomics.

Summary:

The Malaspina expedition has represented a milestone in the Spanish science: a unique expedition that has collected biological samples at a global scale. We have sequenced more than 200 metagenomes and 50 metatranscriptomes from all the temperate oceans across the globe, with emphasis in the bathypelagic (between 1000 and 4000 m deep), comprising 7 Terabases and 30.000 million read pairs. This unprecedented dataset has allowed us to generate a marine microbial gene catalog (M-geneDB) of more than 50 milion of distinct genes, many of them with still unknown function. We have also generated a first collection of 317 high-quality metagenome-assembled genomes (MAGs) from deep-ocean bacteria and archaea that has allowed us to get a glimpse on the metabolic potential of the keystone prokaryotes in the bathypelagic (Acinas et al., 2019). The MAG collection include potentially chemolithoautotrophic microorganisms - capable to incorporate inorganic carbon in the absence of light, as well as other microorganisms capable of nitrogen fixation - an essential element for life scarce in the ocean. We still don't know how active these microorganisms are in the deep ocean, and therefore their relevance in the biogeochemical cycles as active keyplayers remains unkown. The analysis of the metatranscriptomes of the Malaspina expedition will be crucial to better understand microbial diversity and functioning in this unexplored ecosystem on Earth. The goals of this TFM project would be: 1) to analyze the Malaspina bathypelagic metatranscriptomes, mapping them to the MAG collection and to explore their metabolic activity, 2) to develop a web front-end to query and visualize our MAGs collection from the Malaspina global expedition, 3) to assist in increasing the MAG collection by assemblying, binning and annotating metagenomes.

References:

1.Acinas, S. et al. Metabolic Architecture of the Deep Ocean Microbiome. bioRxiv 635680 (2019). doi:10.1101/635680

Expected skills::

The applicant should be proficient in the unix command line and R. Knowledge of shiny and/or web development tools and MySQL will be a plus. She or he should be pro-active and able to acquire new skills autonomously.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

The candidate would be integrated within the team of Dr. Acinas under the close supervision of the bioinformatician Dr. Pablo Sánchez and Dr. Acinas. The candidate would be also interacting with other members of the Acinas' lab and other PIs of the department. The candidate would have an active scientific environment attending to lab meetings and seminars at the ICM.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Luis A Pérez Jurado & Carlos Ruiz Arenas
Email	luis.perez@upf.edu / carlos.ruiza@upf.edu
Institution	UPF
Website	
Group	Genetics Unit

Project

Computational genomics

Project Title:

Combination of common and rare genetic variants to improve the diagnosis of complex disease

Keywords:

Human genetics Polygenic risk score Pathogenic variants Complex diseases

Summary:

Complex diseases, such as obesity or autism, are caused in most cases by a combination of environmental and genetic factors. Genetic factors can be classified as pathogenic or disease susceptibility variants, depending on the strength of their association with the phenotype. On one hand, pathogenic variants are genetic variants enough to cause the disease with high penetrance. These variants are normally ultra rare (<0.1%), preventing the application of association studies and usually requiring family studies and/or other functional studies for their validation. On the other hand, disease susceptibility variants increase the risk to suffer the disease but they are not enough by themselves to cause it, requiring additive effects of other concurrent genetic or environmental factors. These susceptibility variants tend to be relatively common in the population (from 0.1% to 50%) and association studies, comparing their frequency in cases and controls, can be used to define and quantify their relation to disease. The results of these association studies can be collapsed in Polygenic Risk Scores (PRS). PRS are a measure of the risk alleles for a disease carried by an individual. Thanks to the availability of big public datasets, PRS have improved their performance and can identify individuals with high susceptibility to disease (Khera et al 2018). In recent years, both approaches have been independently applied to study complex diseases. For instance, the genetic heritability of autism has been estimated to be 3-10% due to de novo rare variants, 3-10% to inherited rare variants and around 50% due to common variants (Alonso-Gonzalez et al, 2018). Despite this success, a significant proportion of heritability is still missing. We hypothesize that missing heritability is mainly due to rare variants with low penetrance, i.e. variants that are only pathogenic in a specific genetic background (Fahed et al, 2020). In this project, we propose to combine the analysis of common and ultra rare variants to improve our understanding of some common diseases. We propose three main tasks: - Compare PRS between controls and cases with and without high penetrant variants - Prioritize variants considering PRS - Propose new candidate genes We will use autism as an example of a complex disease and we will apply our

methods to data from public repositories (dbGAP, SFARI, UK Biobank) and internal data. The applicant who will work in this project will learn to perform variant calling, prioritize genetic variants, define and compute polygenic risk scores (PRS), work with public data, develop analysis pipelines and work with software containers.

References:

Alonso-Gonzalez A, Rodriguez-Fontenla C, Carracedo A. De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis. Front Genet. 2018 Sep 21;9:406. doi: 10.3389/fgene.2018.00406. PMID: 30298087; PMCID: PMC6160549. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, Lai C, Brockman D, Philippakis A, Ellinor PT, Cassa CA, Lebo M, Ng K, Lander ES, Zhou AY, Kathiresan S, Khera AV. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020 Aug 20;11(1):3635. doi: 10.1038/s41467-020-17374-3. PMID: 32820175; PMCID: PMC7441381. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018 Sep;50(9):1219-1224. doi: 10.1038/s41588-018-0183-z. Epub 2018 Aug 13. PMID: 30104762; PMCID: PMC6128408. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, Samocha KE, Goldstein JI, Okbay A, Bybjerg-Grauholm J, Werge T, Hougaard DM, Taylor J; iPSYCH-Broad Autism Group; Psychiatric Genomics Consortium Autism Group, Skuse D, Devlin B, Anney R, Sanders SJ, Bishop S, Mortensen PB, Børglum AD, Smith GD, Daly MJ, Robinson EB. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017 Jul;49(7):978-985. doi: 10.1038/ng.3863. Epub 2017 May 15. PMID: 28504703; PMCID: PMC5552240.

Expected skills::

Good level of bash and R scripting and a good background in human genetics.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

Lab experiments to confirm the project results might be considered.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	María
Email	maria.soler@vhir.org
Institution	Vall d'Hebron Research Institute (VHIR)
Website	http://www.vhir.org/portal1/grup-equip.asp?t=psiquiatria-salut-mental-i-addiccions&s=recerca&contentid=186872
Group	Psychiatry, Mental Health and Addiction

Computational genomics

Project Title:

Polygenic and trasncriptomic risk score analyses in a sample of school children: assessing the effect of the genetic background of neurodevelopmental disorders on school performance

Keywords:

polygenic risk score, GWAS, TWAS, ADHD

Summary:

Access to education is a predictor for a wide range of later life outcomes such as employment, income, and health. While it is well established that school performance is an indicator of future social success, there is a need for understanding why some children have difficulty learning. About 30-50% of children have poor school performance, which impacts on individual, family, education and social spheres, and predicts worse life outcomes, including health compromising behaviors and physical, mental, and emotional problems. In this project we will aim to explore to which extent the genetic background associated with neurodevelopmental disorders, such as attention deficit/ hyperactivity disorder (ADHD), impacts on school performance outcomes in 3500 children from primary and secondary schools with genotype data available. We will use publicly available summary statistics from genome-wide and transcriptome-wide association studies of neurodevelopmental disorders to generate risk scores in our sample of children and assess their effect on school performance (1,2). We will build these risk scores in two different ways: using genotype data and gene expression data. First of all, we will impute genotypes in our sample against a reference panel to increase the number of genetic variants available, and we will also impute gene expression data, using reference panels linking genotypes with gene expression in relevant tissues. Then, we will build the risk scores and test their effect on school performance in our sample. This strategy may lead us to identify vulnerable groups of individuals with higher risk for poor school performance, and could inform preventive strategies for school failure and to promote population-wide positive development.

References:

1. Demontis, D., Walters, R.K., Martin, J. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet 51, 63–75 (2019). https://doi.org/10.1038/s41588-018-0269-7 2. Pujol-Gualdo N, Sánchez-Mora C, Ramos-Quiroga JA, Ribasés M, Soler Artigas M. Integrating genomics and transcriptomics: Towards deciphering ADHD. Eur Neuropsychopharmacol. 2021 Mar;44:1-13. doi: 10.1016/j.euroneuro.2021.01.002. Epub 2021 Jan 23. PMID: 33495110.

Expected skills::

R programming, flexibility to work with different softwares

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

Any queries please contact maria.soler@vhir.org

Master project 2021-2022

Personal Information

Supervisor	Silvia G. Acinas & Felipe Hernandes Coutinho
Email	sacinas@icm.csic.es
Institution	Institute of Marine Sciences (ICM), CSIC
Website	https://www.icm.csic.es/en
Group	Ecology and Genomics of Marine Microbes

Project

Computational genomics

Project Title:

Revealing novel viruses among sentinel polar prokaryotes by single cell genome analyses

Keywords:

Viruses, prokaryotes, polar regions, single cell genomics, single amplified genomes, microbial metagenomes, auxiliary metabolic genes, Arctic ocean, Antarctic ocean.

Summary:

The polar oceans are under severe threats brought by climate change. Preserving these habitats and the ecosystems therein requires an understanding of the many biological entities that reside there. Microorganisms are the main drivers of the biogeochemical cycles that sustain life in the polar oceans. In turn, viruses of microorganisms play a significant role in this process by selective killing of their hosts and by altering their metabolisms during infection. Therefore, a comprehensive understanding of the biodiversity of Arctic viruses is fundamental for the preservation of these ecosystems. Recent findings from our group revealed that members of the genus Polaribacter are sentinel species, meaning that they quickly respond to the environmental changes that threaten the Arctic (Royo et al. 2020). Yet very little is known about viruses infecting these organisms and their environmental roles. The aim of this project is to reveal novel viruses of Polaribacter through the analysis of Single-cell Amplified Genomes (SAGs). We have previously obtained a set of 91 Arctic and Antarctic Polaribacter SAGs. The project will involve using state-of-the-art bioinformatics tools to identify viruses associated with these genomes, either as prophages or free viruses. In addition, the project will focus on the genetic diversity of these viruses, specifically their repertoire of auxiliary metabolic genes that can be used to alter the host metabolism during infection. The discovered viral genomes and that of their hosts will be analysed in metagenomes from the polar oceans to determine their abundance patterns and associations with physical, chemical and biological parameters, which will shed light on how these factors, together with viruses, control the abundance of this keystone prokaryotic taxa in polar oceans. Finally, the discovery of such viruses has direct implications for the health sciences, as they are likely to encode lysins that could be engineered to be active against Flavobacterium (a close relative of Polaribacter) and other pathogens. Thus, these viruses and their lysins might represent alternatives to antibiotic therapy and are of great biotechnological potential.

References:

Marta Royo-Llonch, M, Pablo Sánchez, Clara Ruiz-González, Guillem Salazar, Carlos Pedrós-Alió, Karine Labadie, Lucas Paoli, Tara Oceans Coordinators, Samuel Chaffron, Damien Eveillard, Eric Karsenti, Shinichi Sunagawa, Patrick Wincker, Lee Karp-Boss, Chris Bowler, Silvia G Acinas. Ecogenomics of key prokaryotes in the arctic ocean. bioRxiv 2020.06.19.156794; doi: https://doi.org/10.1101/2020.06.19.156794

Expected skills::

Basic understanding of a scripting language (e.g. Python or Perl). Basic understanding of statistical analysis with R or MATLAB. Basic understanding of bacterial and viral genomics.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

The candidate would be integrated within the team of Dr. Acinas under the close supervision of the bioinformatician Dr. Hernandes Coutinho and the microbial ecologist Dr. Acinas. The candidate would be also interacting with other members of the Acinas' lab and other PIs of the department. The candidate would have an active scientific environment attending to lab meetings and seminars at the ICM.

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Donate Weghorn
Email	dweghorn@crg.eu
Institution	Centre for Genomic Regulation (CRG)
Website	http://weghornlab.net/
Group	Evolutionary Processes Modeling

Project

Computational genomics

Project Title:

Selection on cancer genomes exerted by the immune system

Keywords:

cancer genomics, immune evasion

Summary:

Cancer is a genetic disease, caused by DNA mutations that accumulate in cells of the human body over the course of time. One of the most important lines of defense against cancer is the immune system. Consequently, detectable cancer tumors must have been able to evade the body's immune surveillance. We expect this feature of successful tumors to leave a footprint of selection in the cancer genome. The aim of this project is to investigate differences in selection between cancer tumors that evolved under different strengths of the immune response. To this end, we will use somatic mutations detected in over 10,000 tumors as well as the tumors' gene expression data. The project has a bioinformatics, a statistical data analysis, and a population genetics component. The student will learn all the corresponding techniques and tools regarding data analysis, partly in collaboration with other lab members.

References:

https://www.nature.com/articles/ng.3987 https://www.nature.com/articles/s41588-020-0687-1

Expected skills::

Programming, logical-analytical thinking

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Chaysavanh Manichanh
Email	cmanicha@gmail.com
Institution	FUNDACIO HOSPITAL UNIVERSITARI VALL D'HEBRON - INSTITUT DERECERCA
Website	https://sites.google.com/site/manichanhlab/
Group	Microbiome Lab

Project

Computational genomics

Project Title:

Development of bioinformatics and statistical tools to integrate meta-omics data to decipher the human microbiome

Keywords:

Human Microbiome; Metagenomics; Metatranscriptomics; Metabolomics; Composition and functions

Summary:

Meta-omics approaches have been intensively used over the last 20 years to study the composition and functions of the human microbiome (the other Human Genome) in health and disease conditions. The aim of the present work is to develop and/or implement bioinformatics tools to analyze and integrate metagenomics, metatranscritptomics and metabolomics data. • You will work in the dry-lab conducting bioinformatics and biostatistical research. You will be integrated in a young and collaborative environment: medical doctors, nutritionist, molecular biologists, bioinformaticians, statistician. • You will learn from your colleagues, and take responsibility, in writing your conclusions into academic papers, which eventually will be published in High Impact Journals. We want to help you build solid foundations on the research method, so you will be assisted by more experienced colleagues.

References:

https://sites.google.com/site/manichanhlab/our-publications

Expected skills::

Fluent in English (most of our team are foreigners, thus English is our language); Theoretical and practical knowledge of classical statistical inference and Machine Learning; Strong coding experience

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

We are looking for a motivated student who is seeking to pursue his/her career in research. The candidate will be remunerated 900 euros/month (gross salary) during his master internship and will be offered the possibility to apply for a PhD fellowship (INPhINIT “la Caixa”, FPU, AGAUR, VHIR...).

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Robert Castelo & Irene Madrigal
Email	robert.castelo@upf.edu
Institution	Universitat Pompeu Fabra / Hospital Clínic
Website	https://functionalgenomics.upf.edu
Group	Functional Genomics / Molecular Genetics

Project

Computational genomics

Project Title:

Deciphering genetics of hereditary hemorrhagic telangiectasia

Keywords:

WES, variant calling, variant filtering and interpretation

Summary:

Hereditary hemorrhagic telangiectasia (HTT) is an autosomal dominant vascular dysplasia leading to epistaxis, telangiectasia and visceral arteriovenous malformations. Pathogenic variants in ENG and ACVRL1 are the main genetic cause responsible of the disease. Historically, genetic testing for HTT consisted of the analysis of ENG and ACVRL1. Nowadays whole exome sequencing (WES) has been introduced as diagnostic tool in patients with this disease. WES allowed the identification of several pathogenic genetic variants; nevertheless the proportion of unresolved exomes is much higher than expected. Particularly in HTT, in which the clinical phenotype is very specific, WES did not reveal, in the studied genes (ACVRL1, ENG, EPHB4, GDF2, RASA1 and SMAD4) , any pathogenic variant in 75% of patients. We assume the existence of other responsible genes o genetic mechanisms in HTT. The main objective of the project is to develop new algorithms for WES analysis in order to detect new candidate genes for HTT. This project will be jointly supervised with Dr. Irene Madrigal from the Molecular Genetics department at the Hospital Clínic de Barcelona, who is in charge for finding a genetic diagnosis for these patients.

Expected skills::

basic knowledge of human genetics, programming and analysis of next generation sequencing data

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Lorenzo Pasquali

Email	lorenzo.pasquali@upf.edu
Institution	UPF
Website	https://www.endoregulatorygenomics.org/
Group	Endocrine Regulatory Genomics

Project

Computational genomics

Project Title:

Genetics and regulatory genomics of glucose metabolism diseases

Keywords:

Regulatory genomics, pancreatic islets, diabetes, chromatin, regulatory functions

Summary:

In the present project we will characterize the dynamics of tissue-specific cis-regulatory networks in tissues central to the glucose metabolism. The project will include the analysis and integration of chromatin data such as open chromatin profiles (ATAC-seq), histone modifications (ChIP-seq), 3D chromatin structure (4C-seq/Hiseq) and transcriptomic maps (RNA-seq), with the aim of identifying unexplored paths in the context of the molecular mechanisms that maintain tissue-specific functions and cell fate.

References:

Ramos-Rodríguez et al. DOI: 10.1038/s41588-019-0524-6 Eizirik el al. doi: 10.1038/s41574-020-0355-7

Expected skills::

High motivation, team work, knowledge of R, experience with Unix operating systems, basic knowledge of regulatory genomics, expertise in statistical analysis.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Master project 2021-2022

Personal Information

Supervisor	Ramiro Logares
Email	ramiro.logares@icm.csic.es
Institution	Institute of Marine Sciences (CSIC)
Website	https://www.log-lab.barcelona
Group	log-lab and Ecology of Marine Microbes

Project

Computational genomics

Project Title:

The ocean microbiome: population dynamics and evolution in a changing planet

Keywords:

metaomics, microbiome, ocean, populations, evolution

Summary:

Oceans are vital for the functioning of the Earth and the regulation of climate. The tiny organisms it contains are crucial for global ecosystem function. For example, microbial phytoplankton in the ocean fix as much carbon from the atmosphere as land plants, and other heterotrophic microbes guarantee that a part of the fixed carbon is circulated through food webs sustaining animal life. In a context of global change, it is fundamental to increase our understanding of marine microbes and how they may be affected by the warming ocean. The genomic machinery of the ocean microbiome has been investigated intensively during the last 15 years thanks to the advent of High Throughput Sequencing. Millions of new genes have been retrieved from the ocean and connected to metabolisms and genomes. Yet, we still have a poor understanding of the fine-grained genomic variation that is normally associated to populations. These populations may show specific adaptations to different oceanic conditions allowing us to comprehend the action of natural selection and to discover gene variants that may code for proteins with functional variation. In addition, we still have a rudimentary comprehension of how that fine variation behaves over time or space, due to selection or drift, leading to evolutionary change. All in all, population genomics and fine evolutionary change are possibly two of the greatest challenges in ocean microbiome research for this decade. The proposed master project aims to 1) determine the population variation of selected ocean microbes (using Single Nucleotide Polymorphisms [SNPs]) and 2) find out whether some of the previous variation is due to evolutionary processes that occurred relatively recently in geological time. This project will occur within the recently funded project MINIME (Microbial Evolution and population genomics in a changing ocean. PI: Ramiro Logares, financed by the Spanish Research Agency). The project will use short (Illumina) and long read (PacBio) metagenome datasets from two global ocean expeditions (Tara Oceans and Malaspina) as well as monthly metagenomes from two coastal microbial observatories in the Mediterranean Sea over 7 years. In combination, these datasets include Terabytes of genomic data being possibly the best representation we have so far of the diversity and function of marine microbes. We will build metagenome-assembled genomes (MAGs) and then map short metagenomic reads from the global ocean or the time-series to a number of selected MAGs of ecological importance. Afterwards, we will perform a SNP calling analysis, aiming to dive into populations genomics. Populations will be determined and we will aim to link them with environmental or geographic features. SNPs analyses will indicate whether part of the detected variation has emerged through adaptive evolution. Most of the work will involve bioinformatics, statistics and machine learning. Analyses will be performed at our marine bioinformatics platform Marbits https://marbits.icm.csic.es as well as at the Finisterrae II supercomputer at CESGA in Galicia via CSIC agreements. This project is well-suited for a motivated student that is up for the challenge to work at the interface between microbiology, metaomics, bioinformatics & oceanography. Work in this project can open future opportunities in other more applied projects within the EU prioritized research area of blue biotech via bioprospecting gene variants of the ocean microbiome that could be used in industry (https://ec.europa.eu/maritimeaffairs/policy/biotechnology_en). There are possibilities of economical support via the CSIC's JAE intro programme (next deadline 12th April 2021, more calls coming in the future (https://sede.csic.gob.es/intro2021) or via other projects (to be discussed). The project is designed so it can occur even in a scenario of Covid19 restrictions. The log-lab has hosted 5 master students from the master in bioinformatics for health sciences in the past and most of them continued with PhD studies at the Institute of Marine Sciences (ICM-CSIC) or abroad or work as bioinformaticians at the ICM-CSIC. The ICM-CSIC is the largest center of marine research in Spain and a leading in its field that has recently received the Severo Ochoa excellence distinction. The ICM has a dynamic, motivating and multidisciplinary research environment that aims at promoting the career development of young researchers (https://www.icm.csic.es/en).

References:

Falkowski, P. The power of plankton. Nature, 2012. 483(7387): p. S17-20. Logares R, et al. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. Microbiome 8:55 Santos-Júnior CD, etal. (2020) Uncovering the genomic potential of the Amazon River microbiome to degrade rainforest organic matter. Microbiome 8:151 Sunagawa, S., et al., Structure and function of the global ocean microbiome. Science, 2015. 348(6237): p. 1261359. Carradec, Q., et al., A global ocean atlas of eukaryotic genes. Nat Commun, 2018. 9(1): p. 373. de Vargas, C., et al., Eukaryotic plankton diversity in the sunlit ocean. Science, 2015. 348(6237): p. 1261605.

Expected skills::

To be familiar with Bash and R and good communication in English

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Feel free to contact me for more details or if you have any questions

---

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	David Comas
Email	david.comas@upf.edu
Institution	UPF
Website	https://www.biologiaevolutiva.org/dcomas/
Group	Human Genome Diversity

Project

Computational genomics

Project Title:

Analysis of the human genome diversity: unravelling demographic and genomic processes

Keywords:

Genome diversity, human populations, demography, adaptation

Summary:

The interests of our research are focused on the human genome diversity analysis in order to infer the (genomic and population) processes responsible for this diversity and try to establish the (population and epidemiological) consequences of the human genetic variability. Thus, our main research lines are focused on aspects of human genome diversity, population genetics, genome variation and disease susceptibility, and genome evolution and disease. 1. Population processes: Concerning population processes that have modeled the human genetic diversity, we have focused our research on the use of molecular tools to reconstruct the human population history through the phylogeny of genetic markers. Our interest has been focused on the genetic consequences at population level of human migrations and admixtures. The use of well-established phylogenies in the mitochondrial and Y-chromosome human genomes allowed us to unravel the population history of several populations. Nonetheless, we have recently used whole genome variation in the autosomes in order to establish the structure of human populations. 2. Genomic processes: Concerning genomic processes that have modeled the human genetic diversity, our research has have been focused on the relationship between human diversity and complex traits, including complex diseases. The genetic analysis in human populations of genes of biomedical interest might shed light on the evolution of these genes. In this context, we have focused our research in the analysis of genes that have been previously associated to complex diseases, such as psychiatric and immunological diseases. The analysis of these genes has allowed us to conclude that some of the failures in replicating genetic associations are due to extreme genetic differences between populations. In addition, we are also interested in other complex traits, such as height, not directly related to disease.

References:

1. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, Fadhlaoui-Zid K, Pimenoff VN, Soodyall H, Zalloua P, Marques-Bonet T, Comas D (2019) Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biology 20:77. 2. Font-Porterias, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, Calafell F, Comas D (2019) European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. PLoS Genetics 15(9): e1008417. 3. Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D (2019) Heterogeneity in Palaeolithic population continuity and Neolithic expansion in North Africa. Current Biology 29:3953-3959. 4. Castro e Silva MA, Nunes K, Lemes RB, Mas-Sandoval A, Amorim CEG, Krieger JE, Mill JG, Salzano MS, Bortolini MC, da Costa Pereira A, Comas D, Hünemeier T (2020) Genomic insight into the origins and dispersal of the Brazilian coastal natives. Proceedings of the National Academy of Sciences USA 117 (5) 2372-2377. 5. Bianco E, Laval G, Font-Porterias N, García-Fernández C, Dobon B, Sabido-Vera R, Sukarova Stefanovska E, Kučinskas V, Makukh H, Pamjav H, Quintana-Murci L, Netea MG, Bertranpetit J, Calafell F, Comas D (2020) Recent Common Origin, Reduced Population Size, and Marked Admixture Have Shaped European Roma Genomes. Molecular Biology and Evolution 37(11):3175-3187.

Expected skills::

Computational skills to manage and analyze genotype and DNA sequence data from whole genomes

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor Rory Johnson

Email	rory.johnson@ucd.ie
Institution	University College Dublin
Website	gold-lab.org
Group	Laboratory for Genomics of Long Noncoding RNAs in Disease

Project

Computational genomics

Project Title:

Computational genomics of CRISPR-Cas, noncoding RNAs and cancer

Keywords:

CRISPR; cancer; lncRNA; drug discovery; genomics

Summary:

One of the biggest biological surprises of the last decade has been the discovery of a completely new class of genes in the human genome – long non-coding RNAs (lncRNAs). These RNA transcripts are not translated into protein, but instead seem to function as regulatory molecules that control the expression of other genes. As part of the international ENCODE consortium, our group has helped catalogue >10.000 of these genes, 99% of which remain completely uncharacterised. LncRNAs represent an extremely promising source of new drug targets. The objective of our lab is to develop a new generation of anti-cancer therapies based on designed lncRNA inhibitors. We identify lncRNA targets via in-house developed, interdisciplinary strategies combining bioinformatics with CRISPR-Cas9 genome-engineering tools. We can offer a variety of tailor-made projects to interested students. These may be, for example, integrative data analysis to make testable predictions about lncRNA functionality, or creation of pipelines for identification of cancer-causing lncRNAs from our CRISPR screens. Previous MSc students have gone on to publish first-author papers based on their MSc thesis, and have successful scientific careers with us and other groups (eg from UPF: Carlevaro-Fita, Pulido-Quetglas, Mas-Ponte, Lanzos – see Pubmed). Students in our lab get exposed to latest bioinformatic and experimental practices on a daily basis. They get closely mentored and have numerous opportunities to present their work internally. Our lab is involved in several international collaborations and consortia, including the International Cancer Genome Consortium (https://www.nature.com/collections/afdejfafdb) and we were recently awarded a prestigious Future Research Leaders grant from the President of Ireland (https://www.sfi.ie/research-news/news/president-higgins-honours/). If you are motivated to work at the forefront in computational cancer genomics and genome-engineering in a fun, supportive and motivated team, then contact us for more information! See also: gold-lab.org https://twitter.com/GOLDLab_Bern

References:

Selected recent papers: Rheinbay E... PCAWG Consortium (including Johnson R, Carlevaro-Fita J, Lanzos A) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020 Feb;578(7793):102-111. Bergadà-Pijuan J, Pulido-Quetglas C, Vancura A, Johnson R#. CASPR, an analysis pipeline for single and paired guide RNA CRISPR screens, reveals optimal target selection for long noncoding RNAs. Bioinformatics. 2019 (In Press) Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R#. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Research 2019 Feb;29(2):208-222 Joana Carlevaro-Fita, Rory Johnson#. Global Positioning System: Understanding long noncoding RNAs through subcellular localisation. Molecular Cell 2019 Mar 7;73(5):869-883 Roberta Esposito, Núria Bosch, Andrés Lanzós, Taisia Polidori, Carlos Pulido-Quetglas, Rory Johnson#. Hacking the cancer genome: Profiling therapeutically-actionable long noncoding RNAs using CRISPR-Cas9 screening. Cancer Cell 2019 Apr 15;35(4):545-557. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R#, Johnson R#. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nature Genetics 2017 Dec;49(12):1731-1740 Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R#. Towards a complete map of the human long non-coding RNA transcriptome. Nature Reviews Genetics 2018 Sep;19(9):535-548.

Expected skills::

Any of Unix / python / R

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Josep F. Abril
Email	jabril@ub.edu
Institution	Universitat de Barcelona
Website	https://compgen.bio.ub.edu/
Group	Computational Genomics Lab

Project

Computational genomics

Project Title:

SARS-CoV-2 genomic analyses from metagenomic samples

Keywords:

next generation sequencing, genome characterization, taxonomy imputation, variant analysis, phylogenetic reconstructions

Summary:

The candidate will carry on already defined computational protocols to analyze SarsCov2 genomic novel data from clinical and environmental metagenomic samples, in order to characterize variation and to determine taxonomical classification of the viral strains found on those sequenced samples. Environmental samples will be processed by the Viruses Contaminants of Water and Food Lab (VirCont) at Universitat de Barcelona, where clinical samples will be processed by the Respiratory Viruses Unit at Vall d'Hebrón Research Institute. Highthroughput sequenced reads obtained from those samples will be mapped and assembled to reconstruct viral genomes and to extract the specific nucleotide and amino acid variants. In function of the sequencing coverage and the type of samples, intrahost variability of those will be also taken into consideration, as well as different epidemiological factors, to deepen into the mutations driving into different pandemic parameters like patogeneity or infectivity.

References:

Work in progress.

Expected skills::

Candidate should have computational analysis skills using bash, unix tools, and scripting languages (perl, python, R). Further skills in database management or web development will be taken into consideration.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

We have been awarded with a "Pandèmies 2020" AGAUR grant, so we expect to be able to hire some part time technician or pre-doctoral fellow, starting probably next semester. Depending on candidate curriculum and interests, there will be possibilities to apply for pre-doctoral fellowship to follow up with the project.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Arnau Sebe-Pedros
Email	arnau.sebe@crg.es
Institution	Centre for Genomic Regulation (CRG)
Website	https://www.sebepedroslab.org/
Group	Single-cell genomics and evolution

Project

Computational genomics

Project Title:

Investigating animal cell type diversity, evolution and regulation using single cell genomics and epigenomics approaches

Keywords:

Evolutionary biology; Single-cell genomics; Genome regulation; Animal phylogenetics; Comparative genomics

Summary:

Projects and specific tasks We are looking for students to join our team to work on a computational project involving integrative analysis of high-throughput single-cell genomics and chromatin data in different animals and unicellular relatives of animals. You will analyse single-cell datasets from different species and perform comparative genomics analyses. The goal is to reconstruct the evolutionary origin and diversification of animal cell types. We also have a second position to work on the development of a phylogenetics pipeline to infer genome-wide gene orthologies. You will learn about phylogenetics methods, protein alignment tools, and gene family evolution. The goal is to set-up a robust orthology framework to integrate single-cell atlases from diverse organisms; as well as to focus on the evolution of particular multi-gene families that are important for animal multicellularity and cell type differentiation (e.g. transcription factors). About the group Our group studies genome regulation from an evolutionary systems perspective. In particular, we are interested in deciphering the evolutionary dynamics of animal cell type programs and in reconstructing the emergence of genome regulatory mechanisms linked to cell type differentiation (from transcription factor binding through chromatin states to the physical architecture of the genome). To this end, we apply advanced single-cell genomics and chromatin experimental methods to molecularly dissect cell types and epigenomic landscapes in phylogenetically diverse organisms. We also develop computational tools to integrate these diverse data sources into models of cell type gene regulatory networks and we use phylogenetic methods to comparatively analyze these models. Our recent work has provided the first whole-organism cell type atlases in different species and mapped key regulatory genome features underlying these cellular programs (see Sebé-Pedrós 2018, Cell, Sebé-Pedrós 2018 NEE, Sebé-Pedrós 2016 Cell). By sampling additional species and chromatin features at single-cell resolution, we now aim at dissecting the evolution of cell types and their underlying gene regulatory networks.

References:

Check our website: https://www.sebepedroslab.org/

Expected skills::

We are seeking for creative and highly motivated students with an interest in evolutionary biology, genome regulation and/or comparative genomics. We are preferentially looking for dry/computational candidates, but there is also a possibility to work on dry+wetlab projects. Basic bioinformatics skills (command-line terminal, R/python scripting) are highly desirable, while ability to work in collaborative projects is a must. Possibility to continue with PhD after the master.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Geòrgia Escaramís
Email	gescaramis@ub.edu
Institution	Dpt. Ciències Biomèdiques, Universitat de Barcelona
Website	https://www.ub.edu/portal/web/dp-biomedicina/genomica-funcional-de-malalties-neurodegeneratives.-ip-eulalia-marti-puig
Group	Genòmica Funcional de Malalties Neurodegeneratives

Project

Computational genomics

Project Title:

Small-RNAs in neurodegenerative diseases: identification of deregulated species as potential biomarkers and therapeutic targets

Keywords:

small RNA, extracellular vesicles, neurodegenerative disaese, deregulation patterns, biomarkers

Summary:

Background: Neurodegenerative diseases (ND) are debilitating and largely untreatable conditions, whose prevalence dramatically increases in late life. Thus, early diagnosis, patient's stratifications and identification of presymptomatic individuals at higher risk of developing dementia represent important unmet needs. Transcriptional deregulation in ND occurs long before clinical symptoms and pathological hallmarks become evident, and circulating, extracellular RNA (exRNA) has been revealed as a new source of little invasive biomarkers. Our preliminary data show that plasma sub-fractionation provides specific patterns of distribution of protein brain markers and exRNAs in extracellular vesicles (EVs) and non-EV compartments. These results highlight the possibility that both plasma EVs and/or non-EVs compartments are informative layers in biomarker discovery. Moreover, diverse data including our most recent findings suggest that exRNA influence the homeostasis and signaling pathways in brain cells. Objective: The objective of this project is to analyze the exRNA transcriptome and identify new deregulated species in neurodegenerative diseases using small-RNA-seq data. We will evaluate the performance of specific pipelines generated in the lab in comparison with another commonly used analytical platform. Methods: We will use data generated in the lab and/or publicly available data. Transcripts mapping will be performed using STAR toolkit. Quantification and annotation will be carried out using in-house bioinformatic tools: Seqbuster (Pantano et al, 2010) and seqcluster pipelines (Pantano et al, 2011) as well as the exceRpt processing toolkit of the NIH Extracellular RNA Communication Consortium (ERCC) (Rozowsky et al., 2019; doi:10.1016/j.cels.2019.03.004). The integration of the different pipelines will allow us to better characterize the profiles of different small RNA (sRNA) types, such as microRNA, tRNA, snoRNA, piRNA or circularRNA. Differential expression of each sRNA between cases and controls will be carried out using negative binomial generalized linear models through DESeq2 bioconductor R package. Machine learning approaches such as Random Forest or Suport Vector Machines will be performed to evaluate weather global patterns of sRNA types can discriminate disease versus control condition. Functional Genomics of Neurodegenerative Diseases group: Applicants will be integrated into the research group "Functional Genomics of Neurodegenerative diseases" (P.I. Eulàlia Martí), that belongs to the Institute of Neurosciences. We are based at the Department of Biomedical Sciences (Universitat de Barcelona, Campus Clinic), with multidisciplinary research teams tackling different aspects of Neurosciences. This very friendly environment provides a number of different theoretical skills and experimental expertise as well as it gathers many young and successful research leaders. Non-coding RNAs (ncRNAs) generally act as gene expression regulators; and are particularly abundant and diverse in the brain, showing highly dynamic and specific expression patterns. The accurate expression pattern of ncRNAs is fundamental for the correct function of the nervous system and deregulation of ncRNA pathways underlies human disease. Our hypothesis is that the ncRNA profiles reflect in a very precise manner fine-tuned changes in neuronal sates. Our lab has been working for more than 10 years in the understanding of disease-driven deregulation of ncRNAs and their role in neuronal dysfunction. Major challenges in translational biomedicine that are the core research of the lab are (i) to evaluate the potential of ncRNAs as disease-specific, peripheral non-invasive biomarkers and (ii) to understand the functional and pathogenic relevance of these species, which may help to unravel disease mechanisms and identify therapeutic targets. We perform these activities using state-of-the art functional genomics approaches, in mouse models and cell cultures. We have also strong expertise in RNA-seq data mining algorithms to detect candidate ncRNA species. We develop these activities in a network of national and international collaborations to address multidisciplinary aspects of RNA biology in Neurosciences.

Expected skills::

basic bash programming and R programming language knowledge

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Amelie Baud
Email	abaud@ebi.ac.uk
Institution	CRG
Website	https://www.crg.eu/en/programmes-groups/baud-lab
Group	Amelie Baud

Project

Computational genomics

Project Title:

Dissecting the genetic basis of handling-induced micturition in BXD recombinant inbred mice

Keywords:

Genotype to phenotype path; Complex traits genetics; Systems genetics; Animal models

Summary:

We have observed significant and strong differences in handling-induced micturition/urination between two inbred strains of mice, C57BL/6J and DBA2/J. We have collected phenotype data (micturition) on a large number of recombinant inbred strains derived from C57BL/6J and DBA2/J (BXD recombinant inbred mice), and a wealth of additional phenotypes as well as sequence data are available for these mice (http://www.genenetwork.org/). The project aims at dissecting the genetic basis of this phenotype, namely quantifying the proportion of phenotypic variation explained by genetics (heritability), mapping the underlying genomic loci (quantitative trait loci), and identifying phenotypes that are genetically correlated with micturition, in order to better understand what this phenotype represents (e.g. Is it a response to stress? Does it instead reflect morphological differences in the urinary system of the mice?).

References:

http://www.genenetwork.org/ (database and analysis toolkit to study BXD recombinant inbred mice); https://doi.org/10.1016/j.cels.2020.12.002; DOI 10.1007/978-1-4939-6427-7_4

Expected skills::

Experience programming in R would be a plus

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Anthony Mathelier
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway, University of Oslo
Website	https://mathelierlab.com
Group	Computational Biology & Gene Regulation / Mathelier Group

Project

Computational genomics

Project Title:

PREDICTING EVOLUTIONARY CONSERVED PRIMARY MICRORNA TRANSCRIPTION START SITES ACROSS SPECIES

Keywords:

microRNA; evolution; gene regulation; transcription factors

Summary:

An internship for a research Master student in the field of Computational Biology / Bioinformatics is available at the Computational Biology & Gene Regulation group, Centre for Molecular Medicine Norway, University of Oslo, led by Anthony Mathelier. The group develops computational methods and tools to analyze the regulation of gene expression and the mechanisms by which it can be disrupted in human diseases such as cancers. See https://mathelierlab.com/ for further information. MicroRNAs (miRNAs) represent a class of small (~22nt) RNAs that post-transcriptionally regulate gene expression through mRNA degradation and/or translational repression. They are derived from primary miRNAs (pri-miRNAs), which are long RNAs (up to hundreds of kilobases) transcribed by RNA polymerase II. Expression of miRNAs must be accurately controlled as they are involved in key cellular processes and deregulation is associated with diseases such as cancer. Despite decades of active research on the identification of miRNAs, the understanding of their transcriptional regulation is very limited due to a lack of precise pri-miRNA annotations and the location of corresponding promoters driving their transcription. As part of the FANTOM consortium, we and others recently predicted transcription start sites (TSSs) of several pri-miRNAs, dedicated to controlling miRNA expression across hundreds of human and mouse samples. This work highlighted that pri-miRNA promoters seem to be more evolutionary conserved than protein-coding gene promoters. Moreover, distances between pri-miRNA TSSs and miRNAs seems overall conserved between the two species. Taking advantage of these findings, and the recently available high-quality miRNA complements of 45 species, stored in the state-of-the-art MirGeneDB (http://mirgenedb.org) database, the selected student will investigate how to use evolutionary conservation to predict pri-miRNA TSSs across species. The analyses will be complemented with the analysis of cis-regulatory conservation across species by looking at transcription factors binding at the determined

promoter regions using the computational tools and resources developed in the Mathelier group. This work is part of a collaborative effort between the Mathelier (NCMM, UiO, Oslo, Norway) and the Fromm (UiT, Tromsø, Norway) groups for the annotation of primary miRNAs across major metazoan groups. The work will expose the trainee to the analysis of genomics data and the development of computational tools.

Expected skills::

Python, R, or bash

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Roderic
Email	roderic.guigo@crg.cat
Institution	Center for Genomic Regulation
Website	https://genome.crg.cat/
Group	Bioinformatics and Genomics

Project

Computational genomics

Project Title:

: Efficient gene annotation across the entire phylogenetic spectrum

Keywords:

Bioinformatics, gene finding, transcriptomics,

Summary:

Understanding Earth's biodiversity and responsibly administrating its resources is among the top scientific and social challenges of this century. The Earth BioGenome Project (EBP) aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of 10 years (http://www.pnas.org/content/115/17/4325). The outcomes of the EBP will inform a broad range of major issues facing humankind, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services. It will contribute to our understanding of biology, ecology and evolution, and will facilitate advances in agriculture, medicine and in the industries based on life: it will, among others, help to discover new medicinal resources for human health, enhance control of pandemics, to identify new genetic variants for improving agriculture, and to discover novel biomaterials and new energy sources, among others. The value of the genome sequence depends largely on the precised identification genes. The aim of the research project is to develop a gene annotation pipeline that produces high quality gene annotations that can be efficiently scaled to more than one million species. Our group has a long-standing interest in gene annotation. Roderic Guigo developed one of the first computational methods to predict genes in genomic sequences (geneid, Guigó et al, 1992), which has been widely used to annotate genomes during the past years. On the other hand, we are part of GENCODE, which aims to produce the reference annotation of the human genome. Within GENCODE we have developed experimental protocols to efficiently produced full-lengh RNA sequences. Our pipeline will be based on identifying the genes that can be precisely predicted computationally in a given species, subtract them from RNA samples, and produced high quality RNA sequences for the genes that are more difficult to annotate. The master student will work specifically on the identification of selenoprotein genes

Expected skills::

Good programming skills python, C, or similar. Good unerstandgin of molecular biology concets

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Mario Cáceres
Email	mcaceres@icrea.cat
Institution	Institut de Biotecnologia i de Biomedicina (IBB), UAB
Website	https://invfest.uab.cat
Group	Comparative and Functional Genomics Group

Computational genomics

Project Title:

Functional and evolutionary impact of polymorphic inversions in the human genome

Keywords:

Structural variants, Human Genetics, Functional effects, Phenotypic traits, Evolution

Summary:

The master student will integrate in a young, interdisciplinary and highly-dynamic group and the project could vary according to the interest and background of the candidate. In particular, the proposed tasks span a diverse range of themes focused in the functional and evolutionary impact of inversions, which are a little studied class of genomic variants and their effects have been missed in most current genomic studies. 1. Identification and genotyping of inversions from bioinformatic analysis of long-read sequences to study their association with phenotypic traits and disease susceptibility (in collaboration with the CNAG) 2. Comparative study of known human inversion regions in other mammal species genomes to determine if there are inversion recurrence hotspots conserved over long evolutionary distances that might indicate a potential functional role. 3. Development of new functionalities and visualization tools for our human polymorphic inversion data base InvFEST (http://invfestdb.uab.cat/), the world reference of human inversions.

References:

F. A. M. Maggiolini et al. Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. Genome Research 30: 1680-1693 (2020). M. Puig et al. Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. Genome Research 30: 724-735 (2020). C. Giner-Delgado et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nature Communications 10: 4222 (2019). D. Vicente-Salvador et al. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. Human Molecular Genetics 26:567-581 (2017). A. Martínez-Fundichely et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Research 42 (D1): D1027-D1032 (2014).

Expected skills::

Expected skills depend on the actual line of research chosen, but should include scripting/programming skills (python, bash, R and/or perl) and experience in genomic variants and functional analysis. Knowledge of MySQL and PHP would also be helpful for working with the InvFEST database.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Comments:

Depending on the degree of experience of the candidate and the task performed it is possible to obtain financial support for the master practice, Also, at the end of the master there is the possibility to apply for a PhD fellowship.

Master project 2021-2022

Personal Information

Supervisor	Mar Albà
Email	mar.alba@upf.edu
Institution	Hospital del Mar Medical Research Institute (IMIM)
Website	evolutionarygenomics.imim.es
Group	Evolutionary Genomics

Project

Computational genomics

Project Title:

Using Nanopore RNA sequencing to explore the transcriptome

Keywords:

Transcriptomics; transcript discovery; long reads; Nanopore; gene expression.

Summary:

The field of transcriptomics is rapidly advancing thanks to the development of high throughput RNA sequencing techniques. However, most studies are based on Illumina short reads, which limits our understanding of the transcriptome. Short reads can often be mapped to different gene isoforms, which generates ambiguous results, and the use of a fixe set of gene annotations prevents detecting transcripts that are specific of a given individual or disease. Long read sequencing tecnologies, such as Nanopore, can generate reads that correspond to the whole transcript, and thus can solve many of the problems currently associated with the use of short reads. For example, Nanopore reads have recently been employ to unravel the complexity of the transcriptome of SARS-CoV-2 (Kim et al., 2020) or to identify alternative splicing events in cancer (Tang et al., 2020). Whereas several programs are available for transcriptome reconstruction using long reads, we have recently developed the first method that does not need a reference genome and which can identify transcripts in genomes that have undergone structural rearrangements (de la Rubia et al., 2020). The aim of the project will be to investigate the complexity of the transcriptome using Nanopore RNA sequencing reads and compare the results to those obtained Illumina reads. We have already generated the sequencing data for 4 different yeast species, including S. pombe, which contains about 1,000 genes with multiple introns. The project provides an unique opportunity to explore the power of long read tecnologies for novel transcript discovery and quantification.

References:

de la Rubia, I., Indi, J.A., Carbonell, S., Lagarde, J., Albà, M.M., Eyras, E. (2020). Reference-free reconstruction and quantification of transcriptomes from long-read sequencing. bioRxiv, https://doi.org/10.1101/2020.02.08.939942 Kim, D., Lee, J-Y., Yang, J-S., Kim, J.W., Kim, V.N., Chang H. (2020). The architecture of SARS-CoV-2 transcriptome. Cell 181, 914–921. Tang , A.D., Soulette, C.M., van Baren, MJ., Hart, K.,.. , Brooks, A.M. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nature Communications 11(1):1438.

Expected skills::

Interest in transcriptomics and long read technologies; basic knowledge of a programming language; basic knowledge of R; good level of English.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Josefa Gonzalez
Email	josefa.gonzalez@ibe.upf-csic.es
Institution	IBE (CSIC-UPF)
Website	gonzalezlab.eu
Group	Evolutionary and Functional Genomics

Project

Computational genomics

Project Title:

Discovering new targets for malaria vector control strategies in urban settings

Keywords:

malaria, Structural variants, adaptation, urbanization, transposable elements

Summary:

Malaria is a deadly disease that kills ~400.000 people per year mostly in Africa, but also in other worldwide regions. Urban environments were until recently considered to be unfit for Anopheles larvae development. However, during the last decades the two major African malaria vectors, Anopheles gambiae and An. coluzzii, have rapidly adapted to polluted habitats threatening current vector-control strategies. While genomic approaches have already been applied to develop vector-control strategies, so far they have focused on single nucleotide changes in coding regions applied to traits previously known to be relevant for the mosquito vector capacity, such as insecticide resistance. This project puts forward a new strategy based on the emergent field of urban adaptation to identify new genetic and epigenetic targets for malaria vector control. The project aims are (i) identifying all the genetic variants present in Anopheles genomes including SNPs, transposable elements, and copy number variants; (ii) identifying signatures of selection at the DNA level to pinpoint the most relevant genes for urban adaptation; and (iii) identifying the environmental factors more relevant for adaptation to urban environments. This project goes beyond the state-of-the-art by combining two emerging fields of research, urban adaptation and the functional role of structural variants, to tackle a relevant societal challenge that not only affects African countries, as re-emergence of malaria associated with climate change and increased human mobility is already being recorded in non-African countries.

Expected skills::

NGS data processing

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

An interview to further discuss the project is required before acceptance to the lab

---

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Biola M. Javierre
Email	bmjavierre@carrerasresearch.org
Institution	Josep Carreras Leukaemia Research Institute
Website	https://www.javierrelab.com https://www.carrerasresearch.org/es
Group	3D Chromatin Organization

Project

Computational genomics

Project Title:

Deciphering the role and regulation of spatial-temporal genome architecture in B cells and diffuse large B cell lymphomas

Keywords:

B lymphocytes, lymphoma, DLBCL, spatio-temporal genome architecture, transcriptional regulation

Summary:

During differentiation, B cells diversify - through mutation and translocations - their immunoglobulins loci to efficiently recognize a specific pathogenic insult, facilitating its neutralization and destruction. Additionally, alterations of B cell differentiation and immunoglobulins diversification cause the development of cancer, immunodeficiency, allergy and autoimmunity. For instance, off-target mutations and translocations during immunoglobulins diversification is a major cause of diffuse large B cell lymphoma. Despite of the clinical relevance of all these processes, we still do not have a complete understanding of the underlying molecular mechanisms that regulate them. To fill this gap of knowledge, I propose a multidisciplinary approach combining state-of-the-art omics strategies, computational biology, genome engineering and mouse experimentation to provide fundamental insights about B cells and their malignant transformation from the spatio-temporal genome architecture perspective. First, I will shed light on the regulatory factors and their underlying molecular mechanisms that spatially organize the B cell chromatin. Second, I will evaluate whether the spatio-temporal genome organization transcriptionally controls B cell differentiation and function, and whether it can protect the genome from collateral oncogenic damage during immunoglobulins diversification. This will require the development and implementation of a novel, low-input, genome-wide method for studying the gene promoter-centered genome architecture. Finally, I will clinically translate the mechanistic and functional insights to improve our understanding about diffuse large B cell lymphoma and its clinical manage. Collectively, we will provide unprecedented mechanistic insights into our understanding of how B cells function and protect their genome from intrinsic oncogenic damage, with clinical impact on regenerative medicine, immunotherapy, autoimmunity, allergy, immunodeficiencies and cancer.

References:

Javierre B.M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384.e19 (2016).

Expected skills::

• A high level of motivation and interest. • Proficiency in at least one scripting or programming language. • Proficiency in scripting environments for statistics and data analysis. • Competitive CV. • High level of collaborative and communicative skills. • Good level of English speaking and writing skills. • International mobility will be considered a major plus.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Ramon Massana
Email	ramonm@icm.csic.es
Institution	Institut de Ciències del Mar, CSIC
Website	https://emm.icm.csic.es
Group	Ecology and Genomics of Marine Microbes

Project

Computational genomics

Project Title:

Unraveling the diversity and genomic data of marine microbial predators

Keywords:

Single Cell Genomics, Illumina assembling, Gene prediction and function, Microbial eukaryotes, Phagotrophy

Summary:

Molecular surveys of microbial eukaryotic diversity in the past two decades have unveiled many novel and uncharacterized species that are major components of marine ecosystems. The extent of this novelty is particularly dramatic among the heterotrophic and mostly bacterivorous species. These eukaryotic predators play key trophic roles in marine ecosystems but there is little knowledge regarding their ecophysiology and the genomic basis of its bacterivorous activity. In our lab we address this question by using metabarcoding, metagenomics, metatranscriptomics and single-cell genomics on natural microbial communities. In this master project, the student will work in the analysis of a large collection of single-cell amplified genomes (SAGs) collected at the Blanes Bay Microbial Observatory, assisting in the assembly, gene prediction and functional annotation of the partial genomes and identifying the taxonomical affiliation of the SAGs retrieved. The student will also be involved in culturing assays with a set of eukaryotic predators where hypothesis built upon the genomic data can be tested experimentally, such as the role of rhodopsin in food vacuole acidification or the presence of chitin as a resting mechanism in some species. This training period will provide an ample overview of microbial ecology, genomics and bioinformatics to the selected student, which will surely benefit further stages in his/her career.

References:

Labarre, A., D. López-Escardó, F. Latorre, G. Leonard, F. Bucchini, et al. 2021. Comparative genomics reveals new functional insights of uncultured MAST species. ISME J. doi.org/10.1038/s41396-020-00885-8 Massana, R., A. Labarre, D. López-Escardó, A. Obiol, F. Bucchini, et al. 2021. Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. ISME J. 15:154–167. Labarre, A., A. Obiol, S. Wilken, I. Forn and R. Massana. 2020. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. Limnol. Oceanogr. 65:S149–S160.

Expected skills::

UNIX, bash, python, R programming. Notions of protein databases and functional assignations

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Funding can be provided through a JAE Intro fellowship (https://sede.csic.gob.es/intro2021), under a call with the same title. Deadline on 12 April 2021

Master project 2021-2022

Personal Information

Supervisor	Mar Albà
Email	mar.alba@upf.edu
Institution	Hospital del Mar Medical Research Institute (IMIM)
Website	evolutionarygenomics.imim.es
Group	Genòmica Evolutiva

Project

Computational genomics

Project Title:

Identification of novel classes of neoantigens in cancer

Keywords:

Transcriptomics; cancer; small ORFs; neoantigens; immunotherapy.

Summary:

Mutations affecting proteins expressed in cancer cells can generate neoantigens, which are small peptides presented by MHC receptors. These peptides can be recognized by T cells and lead to the destruction of cancer cells. The identification of neoantigens expressed by cancer cells is thus crucial to develop anti-cancer vaccines and predict the response to immunotherapy. Neoantigens derived from single mutations in coding sequences can be predicted using whole exome sequencing (WES), which has become a widely employed technique for the molecular characterization of cancer samples. There is one important class of neoantigens, however, that cannot be detected by WES. This class is made of peptides that are cancer-specific – absent from normal tissues - and that are usually not annotated in the databases. These neoantigens are predicted to be highly immunogenic and thus could improve the development of new treatments for cancer. They include peptides originated from the translation of ORFs in cancer-specific transcripts or peptides derived from frameshift mutations that escape nonsense mediated decay. This type of neoantigens can only be identified using RNA sequencing data (RNA-Seq) of cancer samples. The aim of the project is to identify different classes of neoantigens using available WES and RNA-Seq data from diverse cohorts of cancer patients that have been treated with immunotherapy (Litchfield et al., 2021). We will measure the prevalence of neoantigens carrying single mutations and of neoantigens that are cancer-specific, and investigate how their abundance correlates with the response to immunotherapy. In the group we have previously employed massive transcriptomics data to identify recently originated transcripts in human and mouse and predict novel peptides (Ruiz-Orera et al., 2015; Ruiz-Orera et al., 2018). Here we will use similar techniques to identify cancer-specific transcripts and neoantigens.

References:

Litchfield, K., Reading, J.L., Puttick, C., .., Quezada, S.A., McGranahan N., Swanton, C. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell, 184: 1–19. Ruiz-Orera, J., Grau-Verdaguer, P., Villanueva-Cañas, J-L., Messeguer, X., Albà, M.M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nature Ecology and Evolution, 2:890–896. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., Albà, M.M (2015) Origins of de novo genes in human and chimpanzee. Plos Genetics, 11 (12), pp. e1005721.

Expected skills::

Interest in computational genomics and transcriptomics; basic knowledge of a programming language; basic knowledge of R; good command of English.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Master project 2021-2022

Personal Information

Supervisor	Maria José Aranzana
Email	mariajose.aranzana@irta.cat
Institution	IRTA-CRAG
Website	https://www.cragenomica.es/research-groups/rosaceae-genetics-and-genomics
Group	Plant Genetics and Genomics

Project

Computational genomics

Project Title:

Machine learning methods in plant genomics

Keywords:

Machine learning, Deep Neural Networks, RNA-seq

Summary:

The team is looking for a high motivated student to confront questions in different areas of machine learning for biology under the supervision of two bionformaticians, members of the research team. The student will be able to join in one or more of the following projects: - Image Segmentation and analysis through deep neural networks - Pedigree reconstruction through graph flow and relationship building - RNA-seq analysis and differential expression in peach

Expected skills::

Phyton and R programing, interest in machine learning

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Núria López-Bigas
Email	nuria.lopez@irbbarcelona.org
Institution	IRB Barcelona
Website	bbglab.irbbarcelona.org
Group	Biomedical Genomics Group

Project

Computational genomics

Project Title:

Understanding cancer biology

Keywords:

Cancer drivers, selective advantage, mutational processes, tumorigenesis

Summary:

A tumor has between hundreds and thousands of mutations and only a few are directly involved in tumorigenesis, frequently called driver mutations. These mutations affect genes which when mutated confer the cell with a growth advantage with respect to its neighbors. Our lab has developed methods to identify these driver genes, and has analyzed tens of thousands of tumors, producing a catalog of the genes underlying tumorigenesis in the most frequent cancer types. Currently, we are interested in cataloguing the downstream effect that mutations affecting these driver genes have in different tumor types. While many mutations in driver genes are capable of driving tumorigenesis, some are not, and the range of driver mutations of a cancer gene varies between tumor types. Understanding the functional effect of driver mutations thus constitutes a key goal of cancer genomics research.

References:

Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H,Gonzalez-Perez A, Lopez-Bigas N 2020, 'A compendium of mutational cancer driver genes', Nature Reviews Cancer, 20, pages555–572. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A & Lopez-Bigas N 2019, 'The mutational footprints of cancer therapies', Nature Genetics, 51, 12, 1732-1740 Pich O, Muiños F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A & Lopez-Bigas N 2018, 'Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes', Cell, Vol. 175, no4, 1, pp 1074-1087.e18 Tamborero et al, 2018. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Medicine. 10:25 Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A & López-Bigas N 2017, 'Reduced mutation rate in exons due to differential mismatch repair', Nature Genetics, 49, 1684–1692 Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A & López-Bigas N 2016, 'Nucleotide excision repair is impaired by binding of transcription factors to DNA', Nature, 532, 7598, 264. Rubio-Perez C, Tamborero D, Schroeder MP, Antolin AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A & Lopez-Bigas N 2015, 'In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals unexploited targeting opportunities', Cancer Cell, 27(3):382-396. Gonzalez-Perez, Abel; Perez-Llamas, Christian; Deu-Pons, Jordi; Tamborero, David; Schroeder, Michael P.; Jene-Sanz, Alba; Santos, Alberto; Lopez-Bigas, Nuria 2013, 'IntOGen-mutations identifies cancer drivers across tumor types', NATURE METHODS, 10, 11, -.

Expected skills::

Basic programming, data analysis and statistics skills. Willing to learn

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Anthony Mathelier and Vipin Kumar
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway, University of Oslo
Website	https://mathelierlab.com
Group	Computational Biology & Gene Regulation / Mathelier group

Project

Computational genomics

Project Title:

Identification of positively selected breast cancer somatic variants in 3D

Keywords:

Cancer mutation; positive selection; non-coding mutation

Summary:

Context Most cancer alterations occur in the noncoding portion of the human genome, which contains important regulatory DNA segments acting as genetic switches to ensure gene expression occurs at correct times and intensities in correct tissues. However, the identification of critical noncoding cancer drivers has been hampered by the lack of accurate mapping and functional characterization of these DNA segments. Available methods that aim to detect noncoding cis-regulatory cancer driver variants in clinical data rely on a narrow representation of genomic disruption exclusively considering hotspots of recurring mutations in an abstract 1D description of the genome. These hotspots are likely cancer-drivers as they represent signals of positive selection in tumor mutations. Unfortunately, such 1D-centric approach omits mutations affecting DNA regions whose combined enrichment occurs in 3D notably through multiple enhancer-promoter coupling. Strategy More specifically, this project looks for enrichment across patients of noncoding variants enriched within 3D chromatin aggregates. To locate these enrichments, we will take advantage of the genome-wide 3D description produced by HiC to capture mutations whose aggregate effect in 3D gets scattered along the conventional 1D representation of the genome. We will first chart the tridimensional organisation of chromosomes for “normal” breast tissue cells using Hi-C data from the HMEC cells. This conformation will be used to derive a reference architecture that will contextualise the identification of driver mutations in breast cancer patients. The accurate reconstitution of chromatin 3D aggregates will be pivotal to capture these mutation patterns. To achieve this we will be using a state-of-the-art clustering method developed in-house that dynamically adjusts the resolution of HiC data to faithfully reflect chromosome architecture at all its scales. We will then use statistical models to evaluate how the spatial proximity between mutation events within the 3D chromatin aggregates coincides with coordinated mutation patterns among these mutations. The innovative part of this project is to join structural and functional characterisation of chromatin aggregates to pinpoint driver mutations. Finally, we will assess how the 3D-clustered mutations coincide with the dysregulation of their potential target genes in patients. Expected Outcome This project will contribute to the characterisation of noncoding alterations by examining the interplay between chromosome architecture and cancer associated variants to produce a more realistic contextualisation for the interpretation of noncoding alterations than current approaches. This multi-omics strategy will culminate with the development of computational tools identifying cancer-driving DNA alterations enriched in specific 3D chromatin aggregates.

Expected skills::

Python, R, or bash

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Julio Rozas

Email	jrozas@ub.edu
Institution	Universitat de Barcelona
Website	http://www.ub.edu/molevol/julio/
Group	Evolutionary Genomics & Bioinformatics

Project

Computational genomics

Project Title:

Comparative and evolutionary analysis of multi-gene families in spider genomes

Keywords:

Comparative genomics; Gene Families; Transposable elements; Repetitive elements; phylogenomics; Adaptive genomics; genome annotation

Summary:

Understanding the origin, amplification and functional role of repetitive sequences in eucaryotic genomes is a central question in Evolutionary Biology. Despite that modern high-throughput sequencing (HTS) technologies are currently accessible for many labs, the accurate identification and annotation of gene family is one of the major challenges in the field. This scenario will change in the near future thanks to the irruption of the so called third-generation sequencing technologies (i.e., long-read sequencing). In this sense, our research group is generating new high quality genomic data from a group of Canary Island endemic spiders (Chelicerata) using long-read sequencing technologies but also chromosome-scale assembly techniques, such as Hi-C and Chicago libraries. The objective of this TFM is to perform a comparative genomic study of the molecular evolution of 1) the major gene families involved in the chemosensory system (olfactory and gustatory), or 2) those encoding venoms and toxins or 3) the repetitive elements (transposable or other types of repetitive sequences) in chelicerates and, by extension, in arthropods. This research has very relevant biological characteristics with many applications, beyond evolutionary biology. For the analysis, we are using comparative genomics and transcriptomics approaches, under the theoretical framework of molecular evolutionary genetics to identify the genomic regions and gene functions driving diversification. We applied powerful bioinformatics tools to detect changes in coding and non-coding regions, in gene copy number, and in gene expression levels associated with speciation processes. The student will participate in the assembly, annotation and analysis in several spiders (and chelicerates) species. For that, he/she will use high quality genome sequences (data generated by our group based on third generation sequencing technologies, and sequences already available in databases), bioinformatics tools (software and scripts to manipulate and visualizte sequences and genomic annotations, to identify repetitive elements, to conduct evolutionary genetics analyses). The basic work-flow will consist in the identification and primary annotation of repeats, the determination of families, types and classes, the estimation of gene turnover rates, or the characterization of the distribution of these repetitive sequences across chromosomes or with respect to other genomic elements, such as protein-coding genes. Many of these analyses will be carried out in our high performance computer cluster.

References:

• Frías-López, C., Sánchez-Herrero, J. F., Guirao-Rico, S., Mora, E., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2016.DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. Bioinformatics 32: 3753-3759. doi:10.1093/bioinformatics/btw534. • Rispe, C. et al. 2020. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest. BMC Biol. 18: 90. doi: 10.1186/s12915-020-00820-5. • Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A., Rozas, J. 2019.The draft genome sequence of the spider Dysdera silvatica (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. GigaScience 8: 1-9. doi: 10.1093/gigascience/giz099. • Vizueta, J., Escuer, P., Frías-López, C., Guirao-Rico, S., Hering, L., Mayer, G., Rozas, J., Sánchez-Gracia, A. 2020. Evolutionary history of major chemosensory gene families across Panarthropoda. Mol. Biol. Evol. 37: 3601-3615. doi: 10.1093/molbev/msaa197. • Vizueta, J., Sánchez-Gracia, A., Rozas, J. 2020. BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. Mol. Ecol. Res. 20:1445-1452. doi: 10.1111/1755-0998.13202. • Vizueta, J., Rozas, J., Sánchez-Gracia, A. 2018.Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates Genome Biol. Evol. 10: 1221-1236. doi:10.1093/gbe/evy081. Research Group References: (http://www.ub.edu/molevol/julio/SelPublications.html)

Expected skills::

Basic knowledge on NGS data handling and analysis, especially in genome assembly and annotation, notions of comparative genomics and transcriptomics approaches and phylogenetic methods, and experience with Linux operating systems and some of the high level programing languages commonly used in bioinformatics (Perl, Python, R).

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Olivia Belbin
Email	obelbin@santpau.cat
Institution	IIB-SantPau
Website	www.santpaumemoryunit.com
Group	Neurobiology of Dementias

Project

Computational genomics

Project Title:

A multimodal study to determine the polygenic risk for cognitive, structural and functional brain changes in Alzheimer's disease patients.

Keywords:

Alzheimer's disease, synapse, polygenic risk, neuroimaging, proteomics

Summary:

Despite significant efforts over the last decade, there is still substantial 'missing heritability' for late-onset Alzheimer's disease (LOAD). As synapse loss is an early event in LOAD, we hypothesized that synapse-encoding genes will be enriched for LOAD risk-modifying loci that could account for this missing heritability. To test this hypothesis, we first characterized the synaptic proteome by combining data from proteomic studies of synaptic fractions isolated from mouse, rat and human brain tissue with gene ontologies from public databases. The resulting synaptic proteome comprised 537 proteins with a known synaptic function and that are expressed at the synapse (Lleó et al 2019). To construct the "synaptic PRS" model, we extracted summary statistics from the International Genetics of Alzheimer's Project genome-wide association meta-analysis of 74,046 patients for the 2,993 single nucleotide polymorphisms (SNPs) that reside within the 537 synapse-encoding genes. Synaptic PRS using these SNPs were calculated using PRSice-2 software (Choi and O'Reilly, 2019, Euesden, et al., 2015) as previously described (Chaudhury, et al., 2019, Lawingco et al., 2020). An unbiased threshold for p-values in the genome-wide meta-analysis was used to prioritize SNPs that gave the best fitting PRS model (highest Nagelkerke r2 value) and the β-statistic was used to generate weighting estimates for each SNP. The optimal model ("synaptic PRS") was tested in 2 independent data sets of controls and pathologically confirmed LOAD. The mean Synaptic PRS was 2.3-fold higher in LOAD compared to controls ($p<0.0001$) with a predictive accuracy of 72% in the target dataset (n=439) and 73% in the validation dataset (n=136), a 5-6% improvement compared to the current best known LOAD risk factor, APOE ($p<0.00001$) and a 3% improvement compared to an unrestricted model. The synaptic model comprises 8 variants from 4 previously identified (BIN1, PTK2B, PICALM, APOE) and 2 novel (DLG2, MINK1) LOAD loci involved in glutamate signaling ($p = 0.01$) or APP catabolism or tau binding ($p = 0.005$). As the simplest

PRS model with good predictive accuracy to predict LOAD, the synaptic PRS could be used to identify individuals at risk of LOAD before symptom onset. These data are published in Lawingco et al 2020. In the proposed project, the candidate will generate the synaptic PRS for 360 clinically diagnosed AD patients from the SPIN cohort (Alcolea et al 2019 Alzheimers Dement (N Y)) using PRSice-2 software, implemented in R, and will perform a multimodal study to determine the association of the synaptic PRS with baseline and longitudinal change in cognitive performance, brain atrophy (structural magnetic resonance imaging), brain glucose metabolism (Positron Emission Tomography) as well as cerebrospinal fluid markers of LOAD pathology (Aβ1-42, AB42:40, t-tau, p-tau) and axonal (NF-L) and synapse (VAMP-2) degeneration in the same individuals. Using R, the candidate will assess whether the predictive capacity of the synaptic PRS can be improved by inclusion of SNP*SNP statistical interaction terms into the model. The candidate will use the same methodology to generate and compare synaptic PRS models for other neurodegenerative diseases such as frontotemporal lobar degeneration related syndromes and Lewy body dementias. Finally, there is scope for the candidate to integrate the PRS with proteomic data from cerebrospinal fluid and synaptic fractions to further elucidate the functional basis of the PRS variants.

References:

Lleó, A., Núñez-Llaves, R., Alcolea, D., Chiva, C., Balateu-Paños, D., Colom-Cadena, M., Gomez-Giro, G., Muñoz, L., Querol-Vilaseca, M., Pegueroles, J., Rami, L., Lladó, A., Molinuevo, J., Tainta, M., Clarimón, J., Spires-Jones, T., Blesa, R., Fortea, J., Martínez-Lage, P., Sánchez-Valle, R., Sabidó, E., Bayés, À., Belbin, O. 2019. Changes in synaptic proteins precede neurodegeneration markers in preclinical Alzheimer's disease cerebrospinal fluid. Mol Cell Proteomics. 2019 Mar;18(3):546-560. Choi, S.W., O'Reilly, P.F. 2019. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8(7). Euesden, J., Lewis, C.M., O'Reilly, P.F. 2015. PRSice: Polygenic Risk Score software. Bioinformatics 31(9), 1466-8. Chaudhury, S., Brookes, K.J., Patel, T., Fallows, A., Guetta-Baranes, T., Turton, J.C., Guerreiro, R., Bras, J., Hardy, J., Francis, P.T., Croucher, R., Holmes, C., Morgan, K., Thomas, A.J. 2019. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. Transl Psychiatry 9(1), 154. T Lawingco, S Chaudhury, KJ Brookes, T Guetta-Baranes, R Guerreiro, J Bras, J Hardy, P Francis, A Thomas, O Belbin, K Morgan. Genetic variants in glutamate-, Aβ-, and tau-related pathways determine polygenic risk for Alzheimer's disease. Neurobiol Aging. 2020 Nov 12:S0197-4580(20)30391-2. D Alcolea, J Clarimón, M Carmona-Iragui, I Illán-Gala, E Morenas-Rodríguez, I Barroeta, R ribosa-Nogué, I Sala, M.B Sánchez-Saudinós, L Videla, A Subirana, B Benejam, S Valldeneu, S Fernández, T Estellés, M Altuna, M Santos-Santos L García-Losada, A Bejanin, J Pegueroles, V Montal, E Vilaplana, O Belbin, O Dols-Icardo, S Sirisi, M Querol-Vilaseca, L Cervera-Carles, L Muñoz, R Núñez, S Torres, M. Valle Camacho, I Carrió, S Giménez, C Delaby, R Rojas-Garcia, J Turon-Sans, J Pagonabarraga, A Jiménez, R Blesa, J Fortea, A Lleó. The Sant Pau Initiative on Neurodegeneration (SPIN) cohort: A data set for biomarker discovery and validation in neurodegenerative disorders. Alzheimer Dement (NY) 5 (2019) 597-609.

Expected skills::

Basic R programming skills, conceptual understanding of genome-wide association studies and polygenic risk scores, regression modelling (glm, nlme, rlr), correlative analyses and basic statistics. Prior knowledge of Alzheimer's disease genetics/biology would be highly valued.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor Geòrgia Escaramís

Email gescaramis@ub.edu

Institution Dpt. Ciències Biomèdiques, Universitat de Barcelona

Website https://www.ub.edu/portal/web/dp-biomedicina/genomica-funcional-de-malalties-neurodegeneratives.-ip-eulalia-marti-puig

Group Genòmica Funcional de Malalties Neurodegeneratives

Project

Computational genomics

Project Title:

Exploring the small transcriptome (sRNAs) in neurodegenerative diseases

Keywords:

small RNA, extracellular vesicles, neurodegenerative disaese, deregulation patterns, biomarkers

Summary:

Circulating microRNAs have proven to be reliable biomarkers of disease, due to their high stability, in vivo in the circulation. Most studies are focused in miRNA profiling in whole blood, plasma serum and more recently in extracellular vesicles (exosomes). The composition of the small RNA transcriptome is more complex than anticipated and other species strongly perturbed in disease status, including snoRNAs and tRNA fragments, may provide a novel source of bioactive compounds with potential as diagnostic and prognostic biomarkers. The objective of this project is to analyze the sRNA transcriptome in different blood compartments (plasma and/or EVs) and identify new deregulated species in neurodegenerative diseases. The role of the student will focus on the application of the tools for the quantification and detection of these species, and the subsequent downstream analysis through statistical learning methods to identify signature patterns that discriminate disease conditions versus controls.

References:

Sørensen SS1, Nygaard AB2, Christensen T. miRNA expression profiles in cerebrospinal fluid and blood of patients with Alzheimer's disease and other types of dementia - an exploratory study. Transl Neurodegener. (2016); 5:6 Max KEA, Bertram K, Akat KM, Bogardus KA, Li J, Morozov P, Ben-Dov IZ, Li X, Weiss ZR, Azizian A, Sopeyin A, Diacovo TG, Adamidi C, Williams Z, Tuschl T. Human plasma and serum extracellular small RNA reference profiles and their clinical utility. Proc Natl Acad Sci U S A. (2018); 115(23):E5334-E5343. Gámez-Valero A, Campdelacreu J, Vilas D, Ispierto L, Reñé R, Álvarez R, Armengol MP, Borràs FE, Beyer K. Exploratory study on microRNA profiles from plasma-derived extracellular vesicles in Alzheimer's disease and dementia with Lewy bodies. Transl Neurodegener. 2019 Oct 3;8:31. doi: 10.1186/s40035-019-0169-5. eCollection 2019. Pantano L, Estivill X, Martí E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Res. (2010); 38(5):e34 Pantano L, Estivill X, Martí E. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. Bioinformatics (2011); 27(22):3202-3 Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, Thistlethwaite W, Subramanian SL, Milosavljevic A, Gerstein M. exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst. 2019 Apr 24;8(4):352-357.e3.

Expected skills::

basic bash programming and R programming language knowledge

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed