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** | 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** | 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

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** | 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** | 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** | 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** | 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

---

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** | Donate Weghorn |
| **Email** | dweghorn@crg.eu |
| **Institution** | 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** | 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** | Arnau Sebe-Pedros |

| | |
|---|---|
| **Email** | arnau.sebe@crg.es |
| **Institution** | 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** | Roderic |
| **Email** | roderic.guigo@crg.cat |
| **Institution** | CRG |
| **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