Processing whole genome sequence data for a panel of world-wide human populations to infer the genetic bases of their biological adaptations

Supervisor: Prof. Marco Sazzini

Research Project

Background: It is well established in the scientific community that populations of *H. sapiens* evolved in Africa and then migrated in the rest of the continents according to the "Out of Africa" model (McEvoy et al., 2011). This resulted in our species having experienced extremely different climatic and ecological conditions (e.g., endemic pathogens and specific availability of food resources) that tightly depended on the geographical areas/environments occupied. Complex/polygenic adaptations are increasingly supposed to represent most of the adaptive responses evolved by modern human populations to cope with such a diversified range of selective pressures (Pritchard & Di Rienzo, 2010). In this context, the improvement of technologies aimed at generating Whole Genome Sequences (WGS) offer the proper type of data suitable to identify the genetic bases underlying these polygenic adaptations. However, studies regarding this topic still lack specifically defined methodologies able to formally test a realistic approximation of the model of polygenic adaptation. The majority of the approaches adopted so far are in fact based on Genome-Wide Association Studies (GWAS), which present several conceptual limitations such as the ascertainment bias affecting the genetic variants typed in these studies, the failure of the Common Disease/Common Variant hypothesis that they are suitable to test, and the necessity to perform association tests on very large cohorts of individuals representative of the population of interest.

Objectives: To overcome these limitations, the present project aims at investigating the genetic bases of polygenic adaptations evolved by a large panel of modern human populations in order to contribute to clarify the complex adaptive history experienced by our species in response to a multitude of selective pressures. To this end, we plan to exploit WGS data generated for several human populations across the globe to implement a combination of methodologies able to infer complex demographic and adaptive events thus contributing to shed light on the evolutionary dynamics that shaped *H. sapiens* genomic variation.

Activity Plan

The activities will be organized according to the following specific tasks:

Task 1 – Assembling whole genome sequence datasets

Publicly available whole genome sequences such as those collected in the context of the 1000 Genomes and HGDP projects (Auton et al., 2015; Bergström et al. 2020) will be assembled in order to obtain a unique starting dataset that would be the most informative as possible about present-day human genomic variably. Furthermore, WGS data for human populations still scarcely represented in such large-scale reference panels will be add to the dataset and processed with the pipeline of analyses described in Task 2 and Task 3. In

particular, these data will include WGS data for populations from Sicily, Sardinia and Corsica generated in the context of the PRIN2020 - Crossing the Sea project, from North and South Italy (Sazzini et al., 2020), as well as from non-admixed Native American populations made available tanks to the collaboration with the research group led by Dr. Andrés Moreno Estrada at the *Human Population Genomics Lab of the Centro de Investigación y de Estudios Avanzados*, LANGEBIO, Irapuato, Mexico.

Task 2 – Performing population structure analyses

After filtering genomic data according to stringent quality check procedures, multiple analyses aimed at delineating genetically homogeneous clusters of individuals/populations will be applied on the assembled datasets. In detail, fine-scale apportionment of genetic variation within and between populations will be firstly evaluated through Principal Component Analysis to check for data consistency after the merging procedure and then by applying more sophisticated clustering approaches, such as those based on ADMIXTURE and Chromopainter/fineSTRUCTURE algorithms.

Task 3 – Implementing selection scans on whole genome sequence data

Filtered WGS data will be then subjected to an integrated pipeline of analysis aimed at pinpointing complex biological functions/genes putatively involved in the modulation of complex adaptations, as previously described in Ferraretti et al., 2025. More in detail, such an approach will rely on the combination of the likelihood-based method developed by Harris & DeGiorgio (2020), which is able to distinguish chromosomal intervals characterized by variation patterns indicative of the action of natural selection from those conformed with a neutral evolutionary scenario, with the network-based algorithm presented by Gouy et al. (2017), which is capable to reconstruct network of genes contributing to those biological functions that appeared significantly impacted by the action of selection. The obtained results will be further validated by adopting more sophisticated approaches, such as that implemented in the study by Mughal & DeGiorgio (2019), which is able to test complex statistical models that consider the demographic history of single populations to infer adaptive events.

Task 4 – Investigating haplotype/genotype frequencies and signatures of archaic introgression at candidate adaptive loci

Genes/complex functions supported by all the selection scans performed will be further investigated in order to explore frequency patterns of related eQTLs (Expression Quantitative Trait Loci) variants, as well as the role of archaic introgression in shaping adaptive responses in our species. To achieve such a purpose, we plan to use specific functions/packages implemented in both the PLINK and R software with the aim of exploring haplotypes/genotypes frequencies across different populations. Moreover, methods such as Sprime (Browning et al., 2018) and VolcanoFinder (Setter et al., 2020), which are capable to infer chromosome intervals presenting variation patterns compatible with events of introgression and adaptive introgression respectively, will be adopted to further process the available WGS data. Finally, the haplotype structure of the candidate

adaptive introgressed genes will be reconstructed and evaluated in function of the similarities with the available archaic genome sequences (i.e. Neanderthal and Denisovan genomes) in order to further confirm the obtained results.