Characterizing transcriptome variation in human populations at single-cell resolution

Ruben Chazarra-Gil*, Marta Melé*
ruben.chazarra@bsc.es, marta.mele@bsc.es
*Barcelona Supercomputing Center, Pl. Eusebi Güell, 1-3, Barcelona, Spain

Keywords
Ancestry, single cell RNA-seq, eQTL

EXTENDED ABSTRACT

Phenotypic diversity in human populations is a direct consequence of genetic variation, which acts in conjunction with environmental and behavioral factors to produce phenotypic variation, from eye color and height to disease susceptibility and responses to drugs (1). High population-specific variability in disease's prevalence have been described, including multiple examples where a disease is strongly overrepresented in a single population, for instance sickle cell anemia in Africans, hemochromatosis in Northern-Europeans or familial Gaucher's disease in Ashkenazi Jews (2). In addition, population differences in response to drugs have been documented, for instance, 5-Fluorouracil (cancer chemotherapeutic), Warfarin (anticoagulant for preventing thrombosis and embolism) or nicotine (3).

In this context, there has been a growing interest in profiling the molecular causes underlying infectious disease-related phenotypic differences across individuals from populations of different genetic backgrounds. Studies using RNA-seq data from primary monocytes, as a model of an innate immunity, have shown that human populations differ in their transcriptional responses to immune challenges, which are largely controlled by genetics and have been shaped by natural selection (4). In addition, a similar work focusing on alternative splicing characterisation upon immune activation highlights the contribution of positive selection to diversify the splicing landscape of human populations (5).

Moreover, additional works using single-cell RNA-seq indicate that most of the ancestry effects on the immune response are cell type specific, exceptuating interferon (IFN) response which is strongly correlated with European ancestry after infection with influenza A virus (*Figure 1*) (6). Also in line with previous evidence, it has been seen that eQTLs explains > 50% of population differences in response to infection, stressing the key role played by genetics in shaping population differences in immune responses (6).

This evidence suggests that genetic ancestry is a main driver of inter-individual differences in response to infection. In turn, these findings highlight the importance of studying the effect of human population genetic variation over disease and disease response for developing effective treatments, and in order to lay the foundations for the establishment of personalized medicine. Moreover, the characterisation of the transcriptome differences derived from human genetic variation can provide further insights into the evolution of human populations.

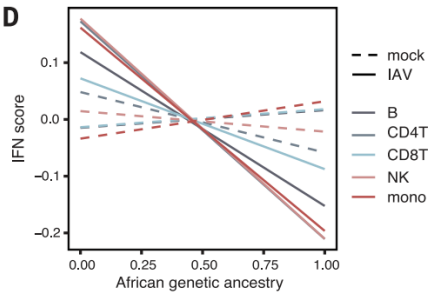

Figure 1*. Correlation between African genetic ancestry proportion and IFN score in mock (dotted lines) and IAV-infected conditions (solid lines) for different peripheral blood mononuclear cell types. A consistent reduction in the IFN response is observed as the estimated african ancestry of individuals increases. INF score represents the average expression of interferon genes, while the proportion of african genetic ancestry is estimated based on polymorphism presence. (Adapted from Randolph et al., (Science, 2021)).*

Data

This project is defined within the eQTLGen Consortium, a large-scale, international collaborative effort with the aim of finding disease-related genetic variants in individual immune cell types. This study will perform **single-cell RNA-seq** and **ATAC-seq** in a collection of peripheral blood mononuclear cells (**PBMCs**) from different African populations, including rural and urban populations, and first generation immigrants. In addition, *ex vivo* immune stimulation will be performed. **Genome variation data** is also expected to be present.

Objectives

The main aim of this project is performing a **detailed characterisation of the transcriptomes of different human populations at single cell resolution**. Some additional objectives are exposed below.

A key step in any scRNA-seq data analysis involves conferring identity to each cell. For this aim, we will use available machine learning methods which are able to project cell type labels from existing annotated references onto our query data. Fortunately, initiatives like the Human Cell Atlas are a source of curated and well annotated scRNA-seq reference datasets across many tissues. In addition, we will manually confirm the assigned cell type identities by checking the expression of established markers for each cell population present.

For the **estimation of genetic ancestry** of each individual we propose the usage of methods (like ADMIXTURE) based on polymorphism datasets.

In order to characterize genetic ancestry effects on gene expression across cell types, we will perform **differential expression analysis** (DEA) to determine population differentially expressed genes (popDEG). In this regard, it will be interesting to confirm the observation of ancestry effects being cell type specific in a non-viral stimulation context (6). Pseudo-bulking (aggregating all the cells from the same cell type and donor into a single profile) is often considered in single-cell DEA to increase statistical power and improve effect sizes. We

will consider both standard single-cell DEA and pseudo-bulking, in case the first strategy is unable to capture population differences. Next, to identify functionally enriched pathways with genetic ancestry, we will perform **gene set enrichment analysis** with the popDEGs. We will consider established sources of terms as GO, KEGG and additional gene signature databases as MSigDB.

In addition, we will assess the contribution of genetic variation to genetic ancestry-associated differences by performing **expression Quantitative Trait Loci** (eQTL mapping). We will search for cis-eQTLs (SNPs within 100kb distance from the gene) and trans-eQTLs (SNPs regulating gene networks over long genomic distances).
We will also identify genetic associations with epigenetic variation (**chromatin accessibility QTLs**) and **response to stimulation QTLs** (reQTLs). It has been reported that cis-eQTLs explain a large fraction of the variance in ancestry-associated expression differences (6). In order to validate this, we aim to quantify the fraction of population differences in gene expression that can be attributed to genetics as done in (6).

Moreover, in order to screen for **polygenic selection**, we will perform GSEA on the intersection of popDEGs and genes with an eQTL (eGenes). In this context, encountering consistent population differences in the expression of genes within the same pathway, can be due to genetic drift, or due to polygenic selection. We will employ different strategies to assess if natural selection, as opposed to genetic drift, has contributed to differences between populations.

Additional analysis we envision to perform include **gene regulatory networks** and **trajectory analysis** comparison across populations.

References

[1] Rahim, N.G., Harismendy, O., Topol, E.J. *et al.* Genetic determinants of phenotypic diversity in humans. *Genome Biol* 9, 215 (2008). https://doi.org/10.1186/gb-2008-9-4-215

[2] Xiao, Q., Lauschke, V.M. The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *npj Genom. Med.* **6,** 41 (2021). https://doi.org/10.1038/s41525-021-00203-x

[3] Bachtiar, M., Lee, C.G.L. Genetics of Population Differences in Drug Response. *Curr Genet Med Rep* **1,** 162–170 (2013). https://doi.org/10.1007/s40142-013-0017-3

[4] Quach, Hélène, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* 167.3 (2016): 643-656 https://doi.org/10.1016/j.cell.2016.09.024

[5] Rotival, M., Quach, H. & Quintana-Murci, L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat Commun* **10,** 1671 (2019). https://doi.org/10.1038/s41467-019-09689-7

[6] Randolph, Haley E., et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* 374.6571 (2021): 1127-1133.https://doi.org/10.1126/science.abg0928

Author biography

Ruben Chazarra Gil obtained his BSc in Biotechnology from the Polytechnic University of València in 2018. Motivated by his interest in bioinformatics, he performed a 1 year Erasmus internship at the Sanger Institute, (Cambridge, UK). Here he became interested in single-cell transcriptomics, and method development. His project consisted in building up a pipeline for benchmarking single-cell RNA-seq data integration methods which was published in NAR journal. Following his interest in gene expression and method development, he performed a second internship at the European Bioinformatics Institute focusing on cell-type annotation methods in the Array Express infrastructure. Next, he spent one year in the Core Bioinformatics Team at the Cambridge Stem Cell Institute, performing extensive single-cell expression data analysis on disease progression datasets in Liver and Lung. In addition, he developed the team's bulk RNA-seq pipelines. He has further interests in multi-omics data integration and application of machine learning in cell-type classification. In addition he is a fan of sports (cycling, running, swimming) and music, playing guitar often and singing in the University of Barcelona choir.