Integrative analysis of gene expression and chromatin accessibility in African populations

Supervisors
EBI: Paul Flicek
WTSI: Manjinder Sandhu

Background
Gene expression studies within and between well-studied populations have been transformative in cataloging gene expression differences, expression quantitative trait loci with different types of regulatory variants, as well as allele-specific expression that underlie many disease associations.[1] Technological advances in RNA sequencing and transcript assembly have also enabled analysis of variation in transcript structure and regulation of alternative splicing. For example, splicing ratios can differ between distant populations even in the absence of expression differences. Such population-specific splicing differences have been shown to be involved in known disease-susceptibility genes that correspond with differences in disease prevalence. Additionally, thousands of unannotated transcripts have been identified within populations,[2,3] highlighting the difficulty in distinguishing population-specific transcripts that are functionally relevant versus those that simply arise from noisy splicing. Elucidating how gene expression regulation and splicing are impacted by historical human migrations, and adaptive selection will aid functional interpretation of the genome and improve our understanding of the transferability and evolution of genetic regulation across populations.

Despite large scale sequencing efforts such as the 1000 Genomes Project, DNA and RNA sequence data collected from the African continent and readily available to researchers is still limited. Capturing representative genetic diversity is important to provide a framework for medical genetics in Africa, as well as a resource for imputation and fine-mapping. In addition to facilitating GWAS, resources are needed to interpret associations in African populations, given that a substantial number of genetic associations may arise from population specific variation. Existing resources examining the transcriptomic landscape in African populations, specifically focusing on genetic variants associated with gene expression are inadequate. We are in the process of creating the African Transcriptomic Resource (ATR) using 1000 genomes cell lines to generate high coverage RNAseq data and accompanying ATAC-seq data from 6 different populations across Africa including the Luhya (LWK), Mende in Sierra Leone (MSL), Gambian in Western Divisions in the Gambia (GWD), Esan from Nigeria (ESN), Yoruba from Ibadan, Nigeria (YRI), and Maasai in Kinyawa, Kenya (MKK). Such a resource will help understand the transcriptional landscape in African populations, population differences in the diversity of splicing isoforms, as well as identify novel transcripts and exons across the genome. This will be also the first large-scale resource to examine eQTLs within Africa. Such a resource will be invaluable for the interpretation of findings from large-scale GWAS being conducted in Africa, supporting initiatives such as the H3Africa consortium.[4]

Summary
This ESPOD fellowship will take a leading role in the analysis of the ATR data. Specifically we are interested in the identification of population specific transcriptional isoforms including novel transcripts and exons used in the African samples as compared to existing European data sets. We are also interested in the identification and comparative population analysis of eQTL data as well as the integration of this regulatory information with ATAC-seq generated for a subset of 100 samples.

Research Plan

Data generation is on going and anticipate completion of all RNA-seq and ATAC-seq by the end of Summer 2017. These data will be mapped to the current human genome assembly in the context of the current reference GENCODE gene annotation and eQTLs calculated.

We will create a catalog of novel transcriptome structure to increase the number of known human transcripts and map transcriptome diversity in Africa and within the specific assayed African populations. We will test the hypothesis that the increased sequence diversity in Africa is connected to both increased transcriptome diversity and increased transcriptional regulatory diversity by joint analysis of the RNA-seq and ATAC-seq data. We also plan to identify and analyse allele specific expression to identify genetic variants associated with differences in gene expression. Each of these analyses will be carried out separately per population, and across all populations, controlling for confounding by population structure.[1]

Finally, we will assess the ATR as an imputation and discovery resource, for imputation of gene expression into GWAS data, to identify loci where imputed gene expression is associated with traits. This can be a powerful approach to identify genetic loci associated with disease, in the context of large transcriptome wide association studies (TWAS).[5,6]

Collaborative Partners

The analysis will be collaborative in nature and draw the expertise of groups participating in the ATR from WTSI, EMBL-EBI, University of Lausanne and Stanford University. We also anticipate distribution of data and some primary analysis via the International Genome Sample Resource (IGSR) at EMBL-EBI and feeding back results into the GENCODE project for refinement of the human genome annotation.

References

1. Lappalainen T, Sammeth M, Friedlander MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013; **501**(7468): 506-11.
2. Peters BA, St Croix B, Sjoblom T, et al. Large-scale identification of novel transcripts in the human genome. *Genome Res* 2007; **17**(3): 287-92.
3. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010; **464**(7289): 768-72.
4. The H3Africa Consortium. Research capacity. Enabling the genomic revolution in Africa. *Science* 2014; **344**(6190): 1346-8.
5. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016; **48**(3): 245-52.
6. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015; **47**(9): 1091-8.

Profile

An ideal candidate would have an analytical background (computer science, mathematics, physics, statistics, etc.) with an understanding of population genomics and appropriate programming experience. A key element of ATR is the assessment of the transcriptional landscape in African populations, comparisons in a global context and the impact of transcriptional variation on quantitative traits and diseases.