**Regulation of gene expression in an Emirati population with type 2 diabetes**

*Katie L Burnham, Katherine Elliott, Evelyn Lau, Lawrence Petherbridge, Julian C Knight*

**Abstract**

**Introduction**

Type 2 diabetes (T2D) is a major public health priority, notably in the Middle East where an elevated incidence is observed. Genome-wide association studies (GWAS) in predominantly northern European ancestry individuals have identified over 400 independent disease associations for type 2 diabetes, but it is unknown to what extent these risk variants are shared across or specific to particular populations. Additionally, understanding the functional basis of such associations is challenging, as regulatory interactions are frequently context specific. Mapping expression quantitative trait loci (eQTL) may lead to the identification of putative regulatory targets for risk variants, and transcriptomics can highlight biologically relevant genes that differ in expression between patients and controls. A number of quantitative traits have also been characterised in relation to T2D, notably fasting insulin and fasting glucose levels, combined glucose homeostasis indices and BMI, which permits the analysis of a larger number of individuals across a range of disease phenotypes.

The nature of genetic variation in the Emirati population and the Middle East more broadly is currently poorly characterised, although data is becoming available for some Levantine populations and the Qatar peninsular *(+ ref to Kate’s paper if poss)*. The present study includes individuals recruited through the Imperial College London Diabetes Centre (ICLDC) in Abu Dhabi, with the overall goal of characterising genomic variation in the Emirati population and investigating its impact on risk of developing type 2 diabetes (*+ ref GWAS if poss/published separately)*. Here, we specifically aimed to profile genome-wide gene expression and map eQTL in whole blood from carefully phenotyped patients with T2D and population controls to establish the extent and nature of transcriptomic variation in diabetes and how this may be modulated by genomic variation in a specific population.

--*How is the GWAS going to be published?*

**Methods**

*Study design and recruitment*

Individuals attending ICLDC were recruited to the study upon clinical assessment for measures of T2D related traits. Exclusion criteria included confirmed non-Emirati descent. Type 2 diabetes was diagnosed … at the time of sampling. Patients were classed as pre-diabetic if at the time of sampling … Patients with other forms of diabetes were categorised as “Other diabetes”.

*Sample collection and processing*

A venous blood sample was taken and 8.5ml placed in a PAXgene Blood DNA tube to stabilise whole blood RNA. Samples were extracted in batches of 24 using the QIAsymphony.

DNA …

*RNA sequencing and data processing*

Strand-specific ribo-depleted RNA libraries were generated for 552 samples using the Illumina TruSeq Stranded Total RNA Library Prep Kit with GlobinZeroHT. Libraries were sequenced to a median depth of 35 million 75-bp-paired-end reads. RNA-seq reads were aligned to the human genome (GRCh38) using STAR, with up to 6 mismatches permitted. Gene-level counts were generated using HTSeq-counts on properly paired reads based on GENCODE v21 annotations. Samples were excluded from the gene expression data on the basis of low mapping rate (2), acute infections (7), white cell abnormalities (3), pregnancy (12), and a lack of clinical information (1). Post-QC, data were available for 527 samples and 22,458 transcripts. Differential expression analysis was performed using the R package DESeq2.

*Genotyping data generation and processing*

Genotyping at >1.7 million sites was performed for 650 samples using Illumina Infinium Multi-Ethnic Global BeadChips. Data QC and filtering was carried out in PLINK v1.9 following Anderson *et al.* (2010). Specifically, individual QC was performed including sex check (3 excluded), excess missingness or heterozygosity (4 excluded), IBD (14 individuals excluded from 14 pairs with Pi\_hat >=0.5), and population outliers on MDS plots (43). SNPs were excluded for excess heterozygosity (HWE p<0.00001), missingness >5%, and MAF<1%. Post-QC, data were available for 586 individuals and 897,948 SNPs.

*eQTL mapping*

Gene expression and genotyping data were available for 471 individuals. One pair of samples was excluded due to a lack of concordance between genotyped variants and those called from the RNA-seq data using the GATK RNA-seq variant calling pipeline, suggesting a potential sample mismatch. The genotyping data were further filtered for MAF <1% and missingness >5% in the 470 individuals used for eQTL mapping, such that 899,740 variants were retained. The gene expression data were restricted to autosomal transcripts and normalised using the TMM method (edgeR).

Each transcript was tested for an association with SNPs within 250kb of the transcriptional start site (TSS) using a linear model including 25 gene expression principal components and 4 genotyping principal components. Significance of eQTL was determined following the approach described in (GTEX). Briefly, for each gene the minimal p-value (minP) was identified i.e. the most extreme p-value observed, and a null distribution of minP generated through 10,000 permutations of the gene expression data. From this, an empirical p-value was computed for the lead SNP per gene. Significant eGenes were then identified using multiple-testing corrected p-values (Benjamini-Hochberg) FDR < 0.05, and associated SNPs by nominal p-values less than the per-gene threshold corresponding to an empirical p-value at q<0.05.

*Data integration*

**Results**

We generated RNA-seq data for 110 T2D patients together with 273 controls, 121 pre-diabetics, and 23 individuals with other forms of diabetes.

*Cohort description, clinical features etc.*

Unsupervised clustering and PCA did not reveal any substructure within the data set associating with disease status or quantitative traits relating to diabetes.

*Differential gene expression*

897 transcripts were significantly differentially expressed between T2D and controls, adjusting for BMI and sex. The most significantly enriched pathway for these genes was TORC1 signallling

*Examples, pathways, quantitative trait analysis*

*eQTL*

Significant associations with local variants (<250kb from the transcriptional start site) were detected for 12,273 transcripts (56.9% of those tested). Of all protein-coding genes tested, 64.7% were eGenes, and 51.9% of lincRNAs.

*Interaction with disease status*

*Conditional analysis? Overlap with TSS, epigenetic data? Pathways and GWAS overlap*

*Trans effects – enriched in different regions?*

*ASE for specific examples?*

*Population-specific effects*

Comparison with European eQTL - Allele frequency/pop-specific alleles

*Data integration*

GWAS overlap

**Discussion**

Stranger *et al.* (2012) identified a clear impact of ancestry on global gene expression, although for the majority of eGenes a significant association was detected across populations.