Interrogation of transcriptome data in Rheumatoid Arthritis for identifying disease susceptibility loci and predictors of treatment response

Rhodri Smith1,2, Nicola Goodson1,2, Robert Moots1,2, Helen L Wright1

(1) Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool UK (2) University Hospital Aintree, Liverpool, UK.

**Background:** Rheumatoid arthritis (RA) is a chronic autoimmune disease affecting approximately 1% of the Caucasian population worldwide causing significant morbidity. The genetics of disease pathogenesis remains poorly understood despite recent advances in high throughput genotyping and sequencing. Biological agents (e.g. TNF inhibitors, TNFi) have significantly impacted on disease management, however 30–40% of RA patients do not respond to this therapy. The aim of this study was to use transcriptome sequencing (RNA sequencing) data from human neutrophils to identify variants in RA that may underpin disease pathogenesis and predict response to biologic therapy.

**Methods**: RNA sequencing (RNA-seq) data from peripheral blood neutrophils isolated pre-TNFi treatment was analysed from 27 RA patients and 6 healthy controls. 21 RA patients subsequently responded to TNFi therapy (change in DAS28 >1.2). RNA-seq reads were mapped to the human genome (hg19) using TopHat2 and annotated using Cufflinks. Data was combined, calibrated and filtered using the Genome Analysis Tool Kit (GATK) to create a file of identified variants. These variants were subsequently interrogated using the VCFtools program package. Quality control parameters were applied in accordance with guidance and available literature, excluding variants that were: PHRED < 30, Minimum read depth < 4 and a loci sequencing success rate < 80%, with SNP clusters and indels also removed. Tajima D was used as a statistic for identifying regions of interest within the RNA-seq data. Identified variants were annotated and interrogated using the UCSC bioinformatics platform and pathway analysis of identified genes predicted through Ingenuity Pathway Analysis (IPA).

**Results**: GATK analysis identified 536,668 variants, which were refined to 5230 variants following application of QC parameters as specified with over 99% of variants excluded. RA patients had a mean Tajima-D score of 0.51 vs -0.19 in the controls (p<0.0001) and furthermore had significantly more regions of transcriptome with extreme positive Taj D values (p<0.0001). Bioinformatics analysis identified the variants with high Tajima D scores to be within a number of biologically relevant loci, including NCF1, which has been associated with autoimmune diseases including SLE and is predictor of RA severity in rat models. IPA revealed that a number of the highest scoring variants were within loci that were linked via a gene network regulated by activation of Fcgamma receptors (FCGR1A/B/C, FCGR2A/B, FCGR3B) and p38 MAPK.

**Discussion**: This study suggests that interrogation of transcriptome data has a role in elucidating the components underpinning RA pathogenesis, identifying a number of interesting loci that may contribute towards its missing heritability. However such preliminary data will require validation through direct sequencing of variants and investigation in independent data sets as well sub-group analysis of treatment response to biological therapy.