P1473 CREATING A RANDOM FOREST CLASSIFIER TO PREDICT HBF LEVELS IN SICKLE CELL DISEASE PATIENTS

Topic: 26. Sickle cell disease

Callum Rakhit1, Stephan Menzel2, Amina Nielsen Nardo-Marino³, John Strouboulis2, David Rees⁴, ⁵, Barnaby Clark1, ²

¹ Precision Medicine, Viapath LLP, London, United Kingdom; ² Red Cell Biology Group, King’s College London, London, United Kingdom; ³ Department of Haematology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark; ⁴ Red Cell Biology Group, Kings College London, London, United Kingdom; ⁵ Haematological Medicine, King’s College Hospital, London, United Kingdom

Background:

Sickle cell disease (SCD) is caused by a genetic variant in the HBB gene, forming HbS. Expression of fetal globin (HbF), reduces the severity of SCD. Three main quantitative trait loci (QTL) have been identified that influence HbF expression; chr 2 (BCL11A), Chr 6 and the globin loci. Drugs such as Hydroxyurea are given to SCD patients as they increase HbF levels and reduce clinical symptoms.

Being able to predict the HbF level of an individual is complex as the number of QTLs is large. Having an accurate HbF prediction may enable triage of affected children, starting effective treatments and management earlier.

We have built a Random Forest model to predict the HbF level of an individual using genome wide SNP array data. This creates decision trees to find the most informative markers that are associated with the trait (HbF). With successive decision trees the data set is split until as many individuals with the same HbF level are grouped. Moving down a decision tree, the entropy level is reduced and the information gain increased, this is quantifiable. Genetic markers that have high predictive value of HbF are likely to be in LD to genomic regions affecting HbF expression. Therefore the predictive model can be used to identify novel regions associated with HbF levels.

Aims:

(1) Create a Random Forest machine learning model that is able to predict the HbF level of an individual affected by sickle cell disease using genome wide SNP array data. (2) Assess the accuracy of the models prediction (3) Review the genetic loci that are contributing to the prediction of HbF

Methods:

Random Forest analyses were primarily performed in the R programming language using the h2o.randomForest algorithm. We created a prediction model using an unaffected twin cohort (1825 individuals) in which both HbF and Illumina MEGA array data was held.

A different model was created for a SCD cohort where HbF values and array data was available. A total of 752 individuals were used to create the sickle model. Both models were adapted to remove overfitting, reducing the number of decision trees to 8 and only using the top 1000 predictive markers.

Results:

In a non-sickle model, 653 cases had HbF prediction. The observed HbF level was plotted against the predicted and the model linearity assessed, R² = 0.45. The three main QTLs, involved in HbF regulation were identified in the top 10 ranked variants, plus other published variants, with a weak association. Plotting the relative importance of each SNP showed the top 10 variants contributed substantially to the model prediction.
A separate sickle model was created. The number of genetic variants that contributed to the HbF prediction in the sickle model was increased compared to the non-sickle cohort. Although the model was more complex the overall prediction accuracy was similar, R²=0.38. The genetic loci used in the prediction included BCL11A but also included NOS1 and ARG2. In addition, intragenic regions were identified at loci very distal to any known gene. Re-running the model, consistently saw the same top predicting SNPs, indicating they were significant findings. We have mapped and described the top performing variants used in the sickle model.

Summary/Conclusion:

We created a Random Forest classification model that can predict HbF levels using genome-wide SNP array data (R²=0.38). The model identified known and novel loci, potentially associated with HbF expression. The model may prove useful for predicting HbF% clinically. We are currently validating the model on a second sickle cell cohort.