Progress report PSC Partners Seeking a Cure

The exome in PSC

Rinse K. Weersma MD, PhD, Professor, Department of Gastroenterology and Hepatology, University of Groningen and University Medical Centre Groningen

Prof. Dr. rer. nat. Andre Franke, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel

Marijn Visschedijk, MD, PhD student Department of Gastroenterology and Hepatology, University of Groningen and University Medical Centre Groningen

Rudi Alberts, PhD, postdoc bioinformatics, Department of Gastroenterology and Hepatology, University of Groningen and University Medical Centre Groningen

---

Both environmental and genetic factors are involved in the development and disease course of PSC, which makes it a complex disease. After a recent study of the International PSC study group, the total number of known PSC genes is 16. Part of the hidden heritability is thought to reside in rare coding variants residing in the exonic regions of the genome. The aim of this study is to identify and confirm exomic variants involved in PSC pathogenesis.

In this project, PSC patients and healthy controls have been genotyped using the Illumina Exome BeadChip. The Illumina Exome BeadChip contains 243,094 exomic variants. DNA of in total 1247 PSC patients has been used from Dutch, German and Norwegian patients. We have used exome chip data of 10038 healthy controls from the same three countries.

We have completed the Quality Control of the exome chip data set. First we have performed quality control on a per sample basis. We have removed the samples that have a bad call rate. Next we have checked whether reported gender of each sample corresponds to the genetic data, and have updated this information if necessary. Then, we have calculated the relatedness between individuals and removed highly related individuals (based on genotype). Finally, we have removed a few ethnic outliers from the data by projecting our data on the first few principal components in the 1000 Genomes genotype data. After the per sample quality control, we have performed a per SNP quality control. SNPs were removed from the dataset based on call rate, Hardy-Weinberg Equilibrium and minor allele frequency.

The exome chip data appeared more difficult to analyze than a standard GWAS chip. Because mainly rare genetic variants are interrogated, standard statistics can not be applied. A simple logistic regression with eventual inclusion of a few principal components to correct for population stratification could therefore not be done. Also, the fisher exact test could not be used because it does not correct for population stratification. As solution we have used the MLMA LOCO method to do single-marker genome-wide association analysis (Yang *et al*.). This method can properly be applied to rare variants and also corrects for population stratification. By applying this method we have identified four novel genome-wide significant genetic variants that are associated with PSC, as well as several interesting candidates just

below the significance level.

Currently, we are running several gene-based analyses using Variant Association Tools (Wang *et al*.). The idea of these tools is to combine the P-values of SNPs that are located within one gene, and that the combination of several weaker signals becomes significant after combining them. Hopefully, this will increase our list of newly discovered genes that are associated with PSC.

Currently, we are busy setting up the replication phase of our findings. For replication we will use samples from the UK, US and from multiple European centers. This has been formally approved by the IPSCSG (International PSC Study Group). Genotyping will be performed using the Sequenom technology. We will start designing a Sequenomplex and the genotyping will be performed at Christian-Albrechts-University in Kiel, Germany.

In the near future, we plan to present and publish the results and we will name PSC Partners Seeking a Cure in our acknowledgments. We are very thankful to get the opportunity to perform this study.

Yang J, Lee SH, Goddard ME and Visscher PM (2011). REML analysis and GCTA Software, GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. Jan 88(1): 76-82.

Gao Wang, Bo Peng and Suzanne M. Leal (2014) Variant Association Tools for Quality Control and Analysis of Large-Scale Sequence and Genotyping Array Data, Am J Hum Genet 94 (5): 770–83.