

FINAL REPORT

PI: Eva Ellinghaus, PhD

Project Title: Whole exome sequencing of PSC patients with severe disease

A. Summary of the project

Primary sclerosing cholangitis (PSC) is a chronic inflammatory disease of the intra- and extrahepatic bile ducts. There is scarce knowledge on etiology and pathogenesis of PSC, and due to lack of effective therapy, PSC has become one of the leading indications for liver transplantation in Northern Europe and the US. The strong genetic component to PSC risk is highlighted by an approximately 9-39 fold increase in sibling risk relative to the general population and recent studies have identified 16 susceptibility regions in the human genome that increase the risk for developing PSC. These, however, explain only 7.3% of variance in PSC liability. Therefore, to identify further sources of heritability, e.g. rare susceptibility variants with potentially high penetrance, systematic resequencing studies need to be performed.

By using a next-generation sequencing-based approach, we sequenced the exomes, i.e. the protein-coding part of the human genome, of 64 unrelated PSC patients with severe disease to dissect the contribution of rare and low frequency coding variants to the complex phenotype of PSC. Newly identified mutations with potentially damaging effect have been validated and followed-up in a large case-control sample. Functional analysis of these mutations is aimed for and the ultimate goal is to identify potential targets for therapeutic intervention.

B. Studies and results for year 2 of the project

Variant calling, in particular calling single nucleotide variants (SNVs), was performed with GATK v1.2 according to their best practice recommendations. For annotation of called variants we applied ANNOVAR's summarize_anno.pl script. Subsequently, we searched for disease mutations in diverse databases, for example in the Human Gene Mutation Database (HGMD), and *in silico* mutation effect prediction algorithms helped to prioritize variants for a follow-up. Mutations in known artefact-prone genes have been marked as potential artefacts. SNVs have been kept only if the following quality criteria were fulfilled: stable genotype (no question mark in the 'BestQual' column), both alleles covered on the forward and reverse strand, and the SNV was not in regions of homology (no 'H' in the 'Homology' column). Final SNVs have been manually inspected in the Integrative Genomics Viewer (IGV). As additional quality control check, we used Illumina's Oncoarray 530k genotype data to determine the concordance rate between the array and the sequencing data. The concordance rate of 17.524 overlapping SNPs was 0.98 on average. After filtering (frequency of variants, damaging prediction by various tools etc.) and QC, we selected the most interesting candidates based on the number of samples affected and accumulation of different damaging variations in a gene across samples, both in GWAS regions and regions that are not known to be disease associated. Technical validation of 80 candidate mutations discovered by whole-exome sequencing was performed by means of Sanger sequencing. We identified mutations in several interesting genes like *MST1* (known PSC risk gene), *SLC9B1*, *SLC25A5*, *MUC6*, *ABCD1* and *AK2*, to name a few. At the present time, we are on the verge of performing the replication genotyping in ~1000 independent PSC cases and 2000 healthy controls, employing the cost-effective Sequenom platform, which will allow us to determine disease association and frequency of the variants.