CENTRE FOR INTEGRATIVE SEQUENCING

*i*SEQ lunch seminar series

Talk: Predicting Disease Risk from Whole Genome Data

Speaker: Bjarni J Vilhjalmsson, Bioinformatics Research Center, AU

Venue: Merete Barker Auditory, The Lakeside Theatres

Time: 5 November 2014 at 12.00 – 13.00

Abstract:

Predicting genetic disease susceptibility or differential drug responses has long been argued as an important benefit of the Human Genome project. Now, more than a decade after sequencing the first human genome, genetic risk predictions have become common in clinical settings, e.g. for breast cancer where the choice of preventative measures and treatment can have a significant impact. Current genetic tests rely primarily on single rare alleles with large effects on disease risk. For most common diseases, such rare genetic variants explain only a small amount of cases. In contrast, evidence from genome-wide association studies (GWAS) of common diseases suggest that most common diseases are highly polygenic with a large number of causal variants with small or moderate effects. Polygenic risk scores (PRS) are a promising alternative to predicting risk prediction as their accuracy is expected to increase as sample sizes in GWAS continue to grow (Dudbridge, PloS Genet 2013; Chatterjee *et al.*, Nat Genet 2013). PRS are also widely used to detect genetic signals, quantifying genetic correlations and understanding the genetic architecture. The standard PRS involves applying a P-value threshold to association statistics and prune variants that are in linkage disequilibrium (LD), preferentially retaining the more significant variant (LD-clumping). However, this approach discards information and limits the predictive accuracy. To address this problem, we propose a Bayesian polygenic risk score that estimates LD from a reference panel and re-weights the effect estimates obtained from GWAS summary statistics. We show that the resulting prediction, which we call LDpred, is the best unbiased prediction when model assumptions hold. LDpred is both computationally efficient and yields well-calibrated predictions. In simulations using real and simulated genotypes we found LDpred to outperform to previously proposed approaches that use GWAS summary statistics as training, especially for larger training sample sizes. We applied LDpred to WTCCC diseases and observed improved prediction R^2 for all but one disease, and for autoimmune diseases that improvement was substantial (e.g. the R^2 on the observed scale improved from 28.3% to 35.5% for T1D). We applied LDpred to five diseases for which we had GWAS summary statistics (as training data) and an independent validation dataset. The Nagelkerke prediction R^2 improved from 20.1% to 25.2 for schizophrenia (SCZ), 9.3% to 12.0% for multiple sclerosis, 4.37% to 5.19% for breast cancer, 3.30% to 3.65% for type-2 diabetes, and 1.36 % to 1.7% for coronary artery disease.

Refreshments:

Sandwiches will be provided. Therefore, please email Anne Hedemand (anne@biomed.au.dk) no later than 4 November 2014, if you would like to participate.