Reading Response

Susanna Makela

April 3, 2017

I thought this would be an easy read since the technical report itself is only four pages long. Of course, I should've known that means that the supplementary information, where all of the statistical and mathematical details are, would be ten times as long (literally). I'm not very familiar with genetics and the statistical methods that have been developed for problems in that field, so it was a bit challenging to wade through all of the abbreviations and notation. It seems that the problem boils down to estimating a binomial probability using the observed genotype for person j and SNP i, x_{ij}. That probability is modeled on the logit scale as being a linear function of the observed trait (outcome) y_j and the allele frequency of SNP i for person j, π_{ij}. This model is different from past models because it reverses the position of genotype x_{ij} and the trait y_j. It also incorporates information on population structure and non-genetic contributions to a given trait (e.g. lifestyle and environmental factors) in the term π_{ij}, which is assumed to be a function of an unobserved latent variable $\mathbf{z}$ capturing these factors. The authors have previously developed a method for estimating π_{ij} using a low-rank approximation to $\text{logit}(\pi_{ij})$. This is a key component of their method because it enables the model to "account for non-genetic effects confounded with population structure without the need to directly observe or model them"; as the authors explain, the dimension of the latent variable $\mathbf{z}$, d, is chosen to best satisfy the assumption that $x_i \mid \pi_i(\mathbf{z}) \sim Binomial(2, \pi_i(\mathbf{z}))$.

The authors apply their method to the Northern Finland Birth Cohort data, which, as a Finn, I'm happy to see. They identify three new loci not found with other methods – in particular, they identify a locus that is statistically significantly associated with height.