PGRNseq Irinotecan: Association Analysis Update

Data

Sample size:

256 samples. 229 that have a non-missing, non-zero Neutrophil count phenotype

Phenotypes:

Neutrophil Count Nadir (1000 cells/mcL)
log Neutrophil Count Nadir

Covariates:

Sex
Site of ascertainment
Dosage regimen

Exome chip:

253 samples
237,431 autosomal SNPs
Unknown platform of origin

PGRNseq:

253 samples
Consensus variants from 3 variant callers (GATK Unified Genotyper, Atlas-SNP2, FreeBayes)
8,739 SNPs genome wide

Processing:

PCA + model-based clustering to identify genetically homogenous set of samples. The procedure is simple: clean exome chip data; run a PCA on the resulting genotype matrix; cluster samples in the first two principal components; subset to samples in the cluster that contains self-reported europeans.

Here we use a model-based clustering technique (implemented in the R package mclust). Essentially, we model the data as a mixture of gaussians. Clustering in two dimensions (i.e. the top two PCs) means that we fit a mixture of bivariate gaussians. Since an arbitrary number of gaussians will perfectly fit any observed data, the method introduces a penalty on the number of parameters in the model in the maximum likelihood estimation. The result is a mixture model that best describes the data given the smallest number of gaussians.

Interestingly, this technique splits the European cluster as well. Looking at some plots, this split is partially explained by differing sites of ascertainment. Much like population structure due to ancestral events, collecting or genotyping samples separately can drive subtle differences variation across the genome.

References:

C. Fraley, A. E. Raftery, T. B. Murphy and L. Scrucca (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington.
C. Fraley and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611:631.

Plots:

Phenotype distributions
PCA + clustering

Single Marker Methods

Statistics:

Simple and multiple linear regression

Tools:

GenABEL maximum likelihood GWA R package

Filters:

MAF cutoff 0.05
HWE FDR cutoff 0.2
Genotyping rate cutoff 0.90

Models:

SNP-only
Sex, site, and dose as categorical covariates

Data:

182 genetically european samples
8,525 markers from PGRNseq

Results:

Flat QQ-plots for simple and full models
Failure to recapitulate known signal: UGT1A1*93: rs10929302

Plots:

QQ-plot for simple and multiple linear regression GWAS

Gene-based Methods

Statistics:

Sequence Kernel Association Test - Optimized (SKAT-O):

This is a popular gene-based association method. The idea is to combine burden and variance component tests into a single, general test.

There are several burden methods that have been applied to association studies. SKAT-O incorporates the Weighted Counting Burden Test (WBT). We model the mean of the phenotype as a linear combination of covariates and genetic effects over p SNPs:

$$Y_i = X_i A + G_i \beta$$

We assume that each β_j is a function of its MAF. So, we can write (1) as:

$$Y_i = X_i A + \beta_0 \sum_{i}^{p} w_j g_i j$$

Where w_j is the MAF of the jth variant. In doing this, we assume: that all variants are causal and all variants have the same magnitude and direction of effect.

For the variance component test, SKAT-O uses the original SKAT method. Briefly, we assume that each β_j follows a distribution with mean zero and variance $w_j^2\psi$. Then, we are able to test whether $\beta = 0$ (i.e. $\psi = 0$). The interpretation is that all variants considered have no variance of effect.

SKAT-O is a linear combination of the burden and variance component tests above:

$$Q_{SKAT-O} = (1-\rho)Q_{SKAT} + \rho Q_{burden}$$

This introduces a parameter ρ that determines the contribution of each statistic. In practice, this parameter is estimated by doing a grid search, and choosing the value that minimizes the resulting p-value.

References:

Key reference: Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies PMID: 22863193
Sequence kernel association tests for the combined effect of rare and common variants PMID: 23684009
Optimal tests for rare variant effects in sequencing association studies PMID: 22699862

Bioinformatics:

Gene information drawn from UCSC hg19 known genes list. Using “VariantAnnotation”, “TxDb.Hsapiens.UCSC.hg19.knownGene” and “org.Hs.eg.db” R packages.

This set of annotations maps onto a much larger set of genes than there should be. The PGRNseq platform is described as interrogating 84 genes, and we are testing

Tools:

SKAT and VariantAnnotation R packages

Filters:

HWE FDR cutoff 0.2
Genotyping rate cutoff 0.90

Models:

SNP-only
Sex, site, and dose as categorical covariates

Data:

182 genetically european samples
8,733 markers from PGRNseq

Results:

Top hits:
SULT1A2
COX10
DBH
CES2

Plots:

Table top results (include: p-value, gene, and # of SNPs)
QQ-plots for simple and covariate models