Rare-Variant Association Testing for Exome Data

The Sequence Kernel Association Test

Sequence Kernel Association Test (SKAT)

- Gene-level (or SNP set) analysis approach
- Tests an association between SNP sets and continuous or discrete phenotypes
- Bypasses the problem of different tagging SNPs being associated with outcomes of interest across ethnic groups

SKAT Main Effects Model

$$Y_i = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{X}_i + \boldsymbol{\beta}' \boldsymbol{G}_i + \epsilon_i$$

- $\boldsymbol{Y_i}$ = outcome for subject i
- $\boldsymbol{\alpha_0}$ = intercept term
- $\boldsymbol{X_i}$ = vector of non-genetic covariates
- $\boldsymbol{G_i}$ = vector of genotypes
- $\boldsymbol{\varepsilon_i}$ = error term; follows any distribution with mean 0 and variance σ^2

- Assume each β_j, j=1,…,p, follows an arbitrary distribution with mean 0 and variance $\boldsymbol{w_j \tau}$
  - Where the weights (w_j) are specified by the user

SKAT basics

- Testing
  H_0: $\beta = 0$ is equivalent to testing H_0: $\tau = 0$

- The score test for variance component in the corresponding mixed model is of the form:

$$Q_\rho = (1 - \rho)\ Q_s + \rho Q_B$$

  - where ρ is the parameter of the unified test, Q_S is a test statistic of SKAT, and Q_B is a score test statistic of weighted burden test

Kernel

- There are pre-specified 6 types of kernels:
  - "linear“
  - "linear.weighted"
  - "IBS“
  - "IBS.weighted“
  - "quadratic"
  - "2wayIX"
- You can use one of them or you can give your own kernel matrix as a parameter.

Default Kernel

- The default kernel is the weighted linear kernel
- The kernel matrix for the weighted linear kernel is

$$\boldsymbol{K = GWWG}$$

  - Where **G** is the n x p matrix of genotype data and **W** is the p x p diagonal matrix of the weights corresponding to each variant.

Q Statistic

$$Q_\rho = (1 - \rho)\, Q_s + \rho Q_B$$

- The Q statistic has a mixture of chi-squared distribution under the null hypotheses that can be evaluated explicitly and used as a reference distribution to compute the p-values.

Weights

- The matrix W is a diagonal matrix that contains the weights of the *p* variants
- Good choices of weights can improve power
- Weights are pre-specified
- If weight j is large, then that variant makes a large contribution to the Q statistic
- Upweighting a causal variant (which is expected to have a large effect) can improve the power
- We don’t know which variants are causal and thus we don’t always know which weights to use

Weights

- SKAT authors suggest using

$$\sqrt{w_j} = Beta(MAF_j; \alpha_1, \alpha_2)$$

- Beta PDF: $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$

Where B denotes the Beta function, alpha and beta (in our weight equation, alpha-1 and alpha-2) are shape parameters and the function is evaluated when x = MAF_j

Results

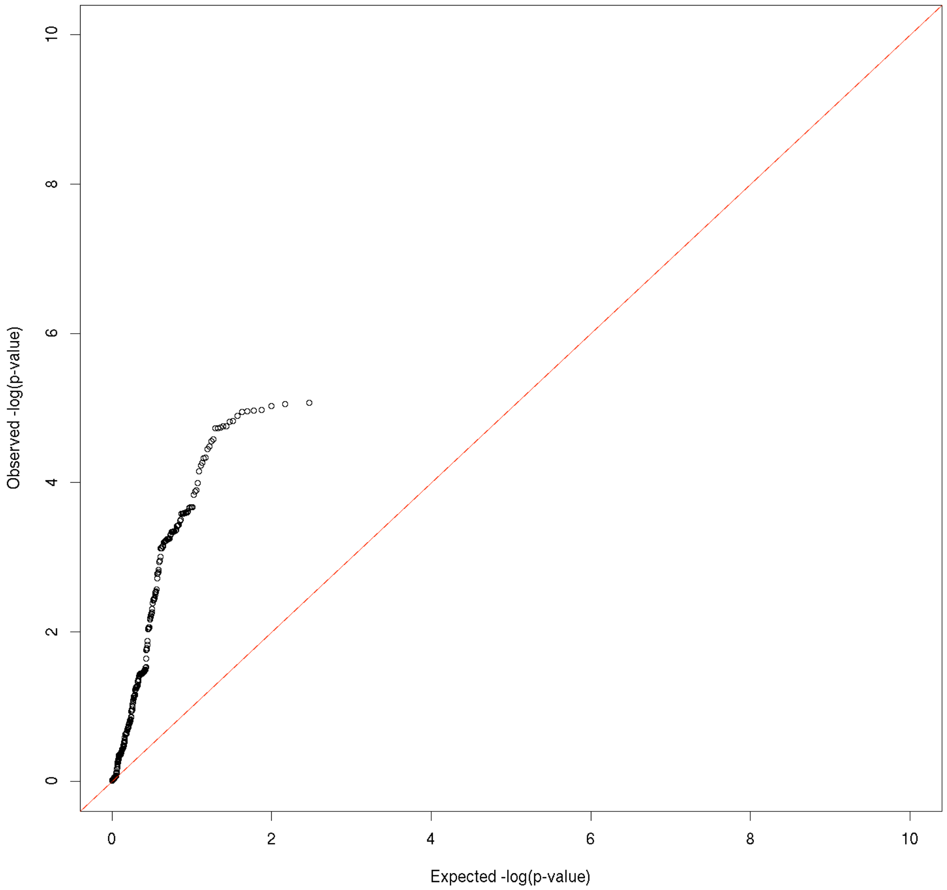
- The Q statistic and associated p-value will tell us if the SNP set is associated with the outcome.
- H_0: $\tau = 0$, assesses whether there is any variance in the SNP set (B_js) from the mean of 0 in any (+/-) direction

Example

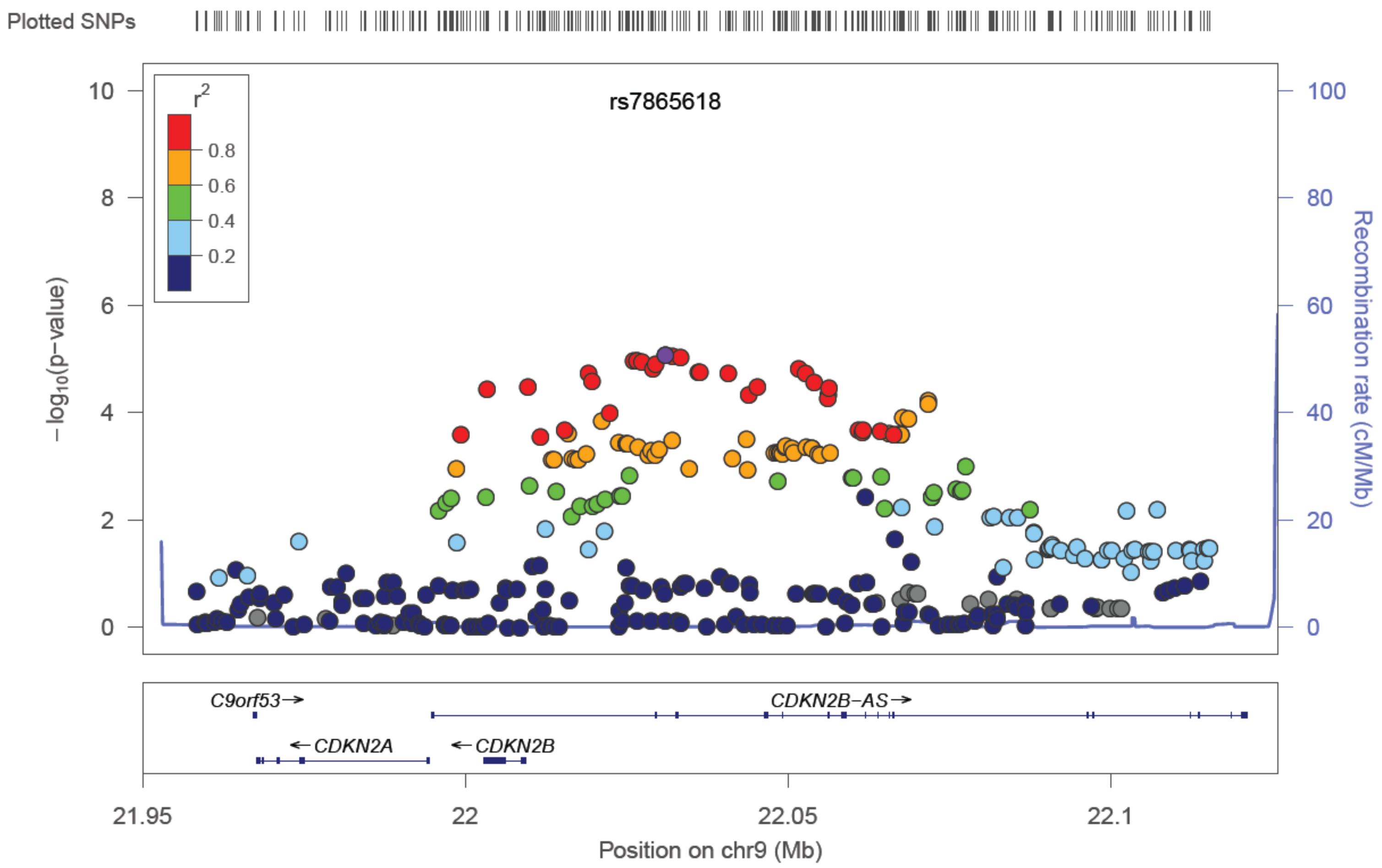
- The association between SNPs in the Chr9p21 region and the Gene Expression

- **Predictor**: SNPs in the Chr9p21 region (297 SNPs)
- **Outcome**: gene expression of the genes across the whole genome
- Single SNP Association Test
  - Gene expression=single SNP + random(family) (297 tests)
- Sequence Kernel (SNP set) association (SKAT)
  - Gene expression (after familiar adjustment)= All SNPs (1 test)

Single SNP Association Analysis Between CDKN2BAS and SNPs in the Chr9p21



Single SNP Association Analysis Between CDKN2BAS and SNPs in the Chr9p21



CDKN2BAS (SKAT p=0.429)

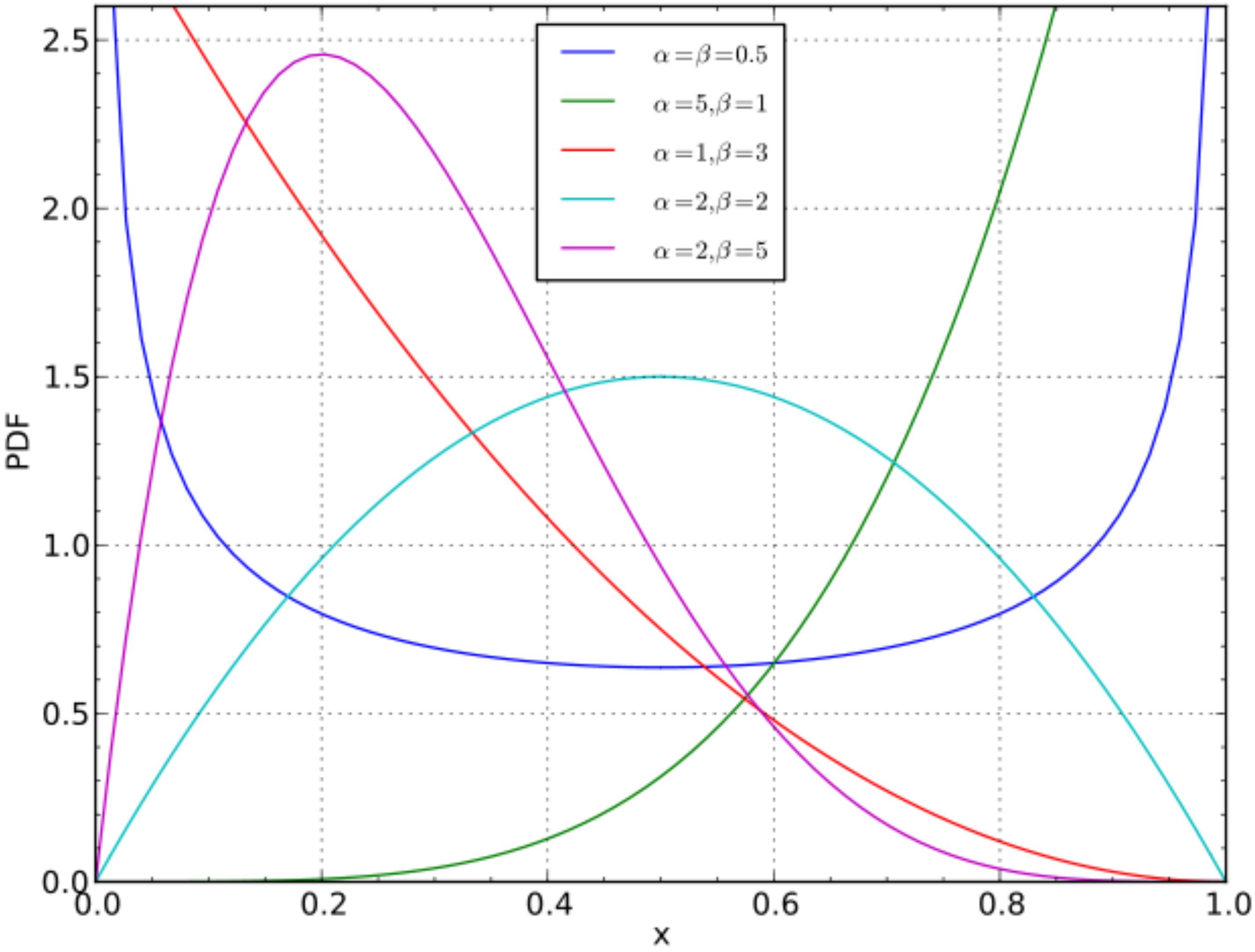
SNP	MAF	B_SNP_P_CDKN2BAS
rs7865618	0.446317	8.58E-06
rs634537	0.44901	8.93E-06
rs2157719	0.495097	9.34E-06
rs613312	0.439054	1.07E-05
rs615552	0.453232	1.09E-05
rs543830	0.438996	1.11E-05
rs599452	0.438958	1.13E-05
rs564398	0.440362	1.28E-05
rs679038	0.439493	1.51E-05
rs944801	0.459316	1.52E-05

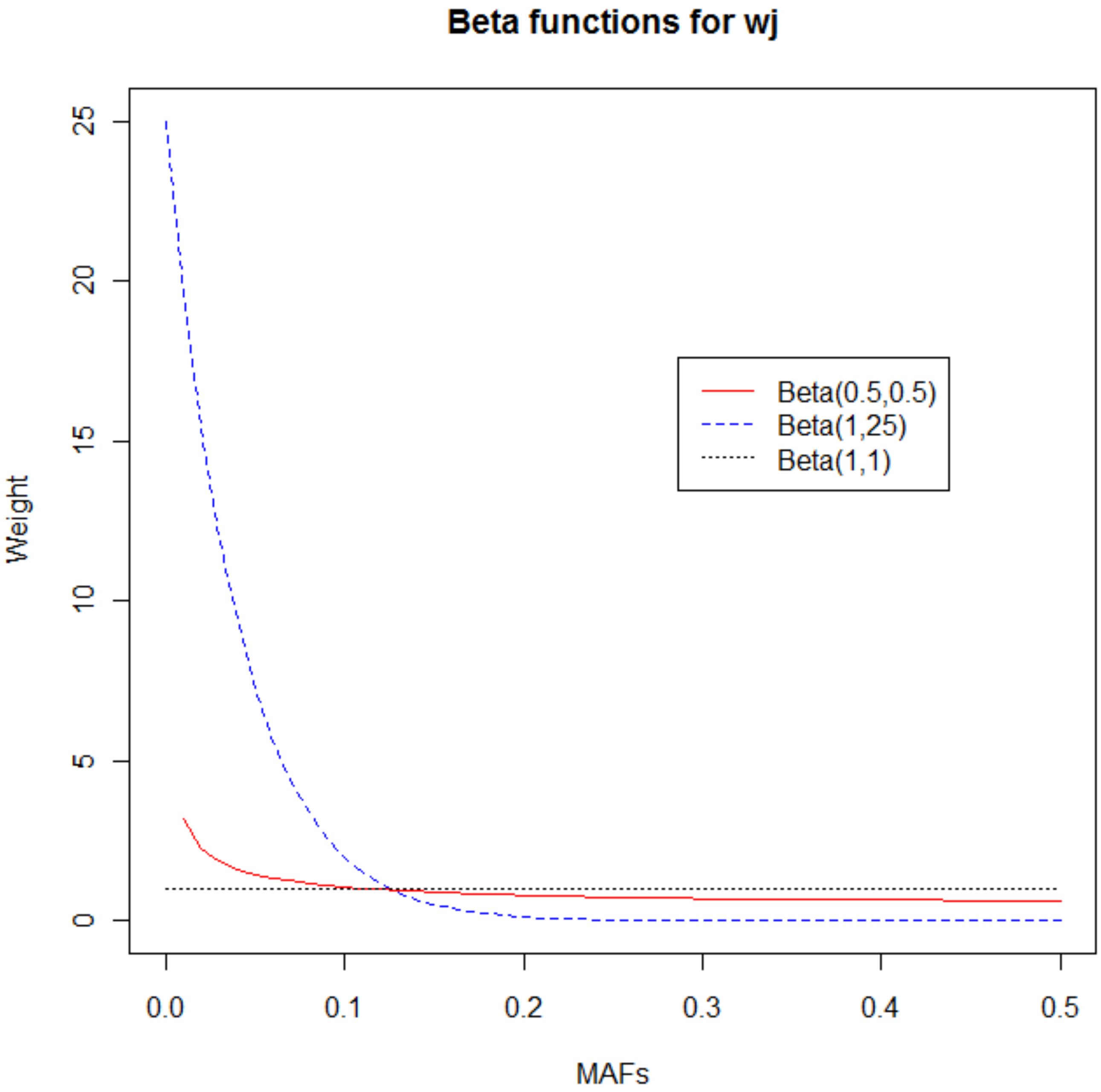
Specifying weights

- Beta (1, 25)
  - – Up regulate rare variants and down regulate common variants
- Beta (1, 1)
  - – Equal weights to all variants
- Beta (0.5, 0.5)
  - – Madsen & Browning weight

$$\sqrt{w_j} = 1 / \sqrt{MAF_j(1 - MAF_j)}$$

Beta distributions





Beta functions for wj
Weight
0
5
10
15
20
25
MAFs
0.0
0.1
0.2
0.3
0.4
0.5
Beta(0.5,0.5)
Beta(1,25)
Beta(1,1)

Applying Different Weight to CDKN2BAS

	Beta (1, 25)	Beta (1, 1)	Beta (0.5, 0.5)
SKAT p value	0.429	0.0020	0.0021

Example p-values
Adjusting Weight Changes Results Dramatically:
Top Results with Weight beta (0.5, 0.5)

		Pvalue	Pvalue	Pvalue_
Transcript	N	beta 1 25	beta 1 1	beta 0.5 0.5
ENST00000301908	801	0.036968216	0.000178972	0.000130954
ENST00000370551	801	0.684927084	0.000178319	0.000217497
ENST00000412318	801	0.112609677	0.000302452	0.00026976
ENST00000497037	801	0.241666486	0.000345966	0.00034696

Conclusion

- Choosing appropriate weight is very important in SKAT
- Beta (1, 25) gives very little weight, if any, to the common variants
- Beta (1, 1) has very little power picking up signals from rare variants
- Beta (0.5, 0.5) can pick up signals from both common and rare variants, but suffers from lower power.

Next Steps

- We’ve been working with Shawn Lee to test his SKAT programs for:
  - Gene-environment kernels (currently unweighted)
  - Gene-gene kernels (currently unweighted)
  - Meta-analysis subroutines (Meta-SKAT)
  - Modifications for family data