Rare variant association tests in presence of heterogenity between cases

Ozvan Bocher[1], Gaëlle Marenne[1], Hervé Perdry[2], Emmanuelle Génin[1]

[1] Univ. Brest, EFS, UMR 1078 [2] Univ. Paris Sud / Paris-Saclay, U1018

The wide availability of Next Generation Sequence data opens new opportunities to discover rare variants associated to rare or common diseases. Rare variants association tests have been proposed specifically to this end. To obtain a higher statistical power, they test association between a whole genomic region (usually encompassing one gene) and the disease. These tests can be broadly classified in two categories: burden tests (CAST, WSS) and variance tests (SKAT).

However, even these tests may have an unsatisfactory power, due to the limited size of case samples and to the large number of genes to be tested when an agnostic approach is used. In this context, incorporating information on clinical heterogeneity among cases, e.g. differences in disease presentation, severity or age at onset, is an appealing way to build association tests with a higher sensitivity.

We propose extensions of both burden and variance tests to the situation where the cases are divided in such subgroups. The burden tests are extended by means of a multinomial logistic regression. A geometrical interpretation of the SKAT test for binary phenotypes is used to construct a natural extension of this test to our setting. The power of these tests is investigated under various simulation scenarios.

Burden tests

Burden tests attribute to each individual a *genetic burden*, and compare the burdens of cases and controls. For $C > 2$ groups of individuals we use a multinomial regression:

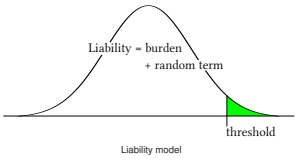
$$\log \frac{P(Y = c)}{P(Y = 1)} = X\alpha_c + B\beta_c \quad (c = 0, \ldots, C-1)$$

where X is a matrix of covariates and B is the vector of burdens.

Hereafter we consider the WSS burden, a weighted sum of rare variants within a gene giving the highest weights to the rarest variants (Madsen & Browning, 2009).

Simulations

To assess power at $\alpha = 2.5 \cdot 10^{-6}$, we simulated datad under a liability model using 3384 LAMTOR3 haplotypes from UK10K TWINS (147 SNPs on 16 Mb, among which 114 with maf < 0.01)



Liability model

- "burden" is computed from a subset of "causal" rare variants
- effect size measured by h^2 = prop. of liability variance due to burden
- threshold depends on disease prevalence (fixed to 0.001 in all our simulations)

We simulate a group of 200 controls and two groups of 100 cases three scenarios:

- SVSR (same variants same risks): the two groups of cases are identical
- DG (different genes): the first groups of cases is identical to controls
- DVSG (different variants in the same gene): all groups are distincts with h^2 in Cases 2 = $4 \times h^2$ in Cases 1

The data are then analyzed in 4 different ways:

- Controls vs All Cases: compares controls with cases pooled together
- Controls vs Cases 1: compare controls with first group of cases
- Controls vs Cases 2: compare controls with second group of cases
- Three groups: compare the three groups using extended test

SKAT

SKAT (Wu et al 2012) is built from the mixed model

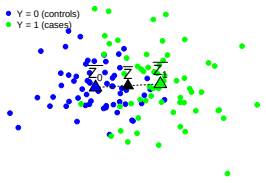
$$\text{logit } P(Y = 1) = X\beta + Zu$$

where Z is a matrix of weighted genotypes, $u \sim MVN(0, \tau I)$ (Wu et al, 2012).

- estimate $\hat{\beta}$ in the null model ($\tau = 0$), let $\hat{\pi} = \text{logit}^{-1}\left(X\hat{\beta}\right)$
- compute the (score) test statistics for $H_0 : \tau = 0$;

$$Q = (Y - \hat{\pi})' ZZ' (Y - \hat{\pi}).$$

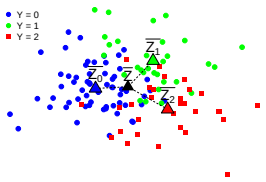
In the absence of covariates, there's a geometric interpretation of Q:



$$Q = n_0^2 \left\|\overline{Z}_0 - \overline{Z}\right\|^2 = n_1^2 \left\|\overline{Z}_1 - \overline{Z}\right\|^2$$

where $\overline{Z}$ (resp. $\overline{Z}_0, \overline{Z}_1$) is the center of mass of genotypes of all individuals (resp. of controls, of cases). For several groups we propose to use

$$R = \sum_c n_c \left\|\overline{Z}_c - \overline{Z}\right\|^2 = \sum_c \frac{1}{n_c} \left(1_{Y=c} - \widehat{\pi}_c\right) ZZ' \left(1_{Y=c} - \widehat{\pi}_c\right)$$

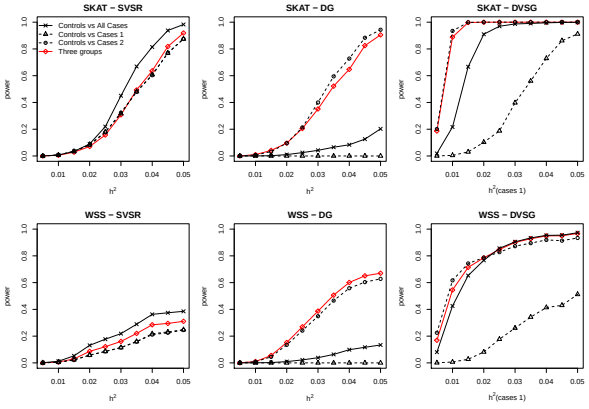


The p-value are estimated by permutations. For small p-values, we use SKAT "small sample" procedure:

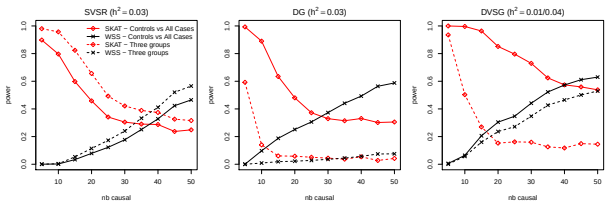
- Estimate statistics' moments 1, 2 and 4 from **50 000** permutations
- Approximate the distribution by a scaled chi-square

Results

Simulations with **30** causal variants (among **114**), h^2 up to 0.05



Simulations with up to **50** causal variants (among **114**), h^2 fixed



Conclusion

- There's a large gain of power when there's heterogeneity between the two groups of cases
- The loss of power when cases are homogenous is moderate (is it possible to find a compromise) ?
- h^2 being fixed, SKAT loses power when there are many causal variants with small effects.This is due to the fact that SKAT does not take into account the presence of singletons.
- We plan to extend SKAT-O along the same lines
- Published article with more on Burden tests: Bocher et al 2019 doi.org/10.1002/gepi.22210
- All tests are implemented in our R package Ravages http://github.com/genostats/Ravages/