

# Statistical tests for Rare Variants Data Rare Variants in Human Genetic Diseases: Comparison of Association Statistical Tests

Lise Bellanger, Elodie Persyn, Floriane Simonet, Richard Redon, Jean-Jacques Schott, Solena Le Scouarnec, Matilde Karakachoff, Christian Dina

### ▶ To cite this version:

Lise Bellanger, Elodie Persyn, Floriane Simonet, Richard Redon, Jean-Jacques Schott, et al.. Statistical tests for Rare Variants Data Rare Variants in Human Genetic Diseases: Comparison of Association Statistical Tests. International Biometric Conference, Jul 2014, Florence, Italy. hal-01160576

HAL Id: hal-01160576

https://hal.science/hal-01160576

Submitted on 5 Jun 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## **Rare Variants in Human Genetic Diseases: Comparison of Association Statistical Tests**



Lise BELLANGER<sup>1</sup>, Elodie PERSYN<sup>2</sup>, Floriane SIMONET<sup>2</sup>, Richard REDON<sup>2</sup>, Jean-Jacques SCHOTT<sup>2</sup>, Solena LE SCOUARNEC<sup>2</sup>, Matilde KARAKACHOFF<sup>2</sup>, Christian DINA<sup>2</sup>

<sup>1</sup> Laboratoire de Mathématiques Jean Leray (LMJ) UMR CNRS 6629, University of Nantes, France. E-mail: lise.bellanger@univ-nantes.fr <sup>2</sup> Institut du thorax, Inserm UMR 1087 / CNRS UMR 6291, University of Nantes, France

### **Background and Objectives**

Genome-wide association studies (GWAS) aim to identify genetic loci of susceptibility of complex diseases. These studies require:

sampling of individuals from 2 types of individuals: cases and controls



genotyping of individuals for a dense set of genetics variants (Single Nucleotide Variants (SNVs)) across the GWASs allow to identify common genetic variants (CVs) associated with many common diseases. But these variants do not fully explain the heritability (i.e. proportion of phe ole to genetic var



Exome sequencing studies: Exons of genes are sequenced for cases and controls.

Minor Allele Frequency (MAF): frequency at which the least common allele occurs in a given population.

Rare variant (RV): mutation present in less than 1-5% of the population. (SNV with MAF≤ 1-5%).

Most popular statistical tests for GWAS based on testing single SNP are powerless due to

### rare + low effects

Development of many statistical tests specifically targeting RVs to detect association between a set of variants (e.g. located on a gene) and a disease.

### **Our Objectives:**

Which? Comparative evaluation of the existing tests for the identification of disease-associated RVs in order to identify the most suitable ones in the context of associated RVs.

Why? There exists a lot of tests but none is uniformly most powerful. Adaptive tests are needed!

How ? With simulated data and real data

### Material and Methods (\*) Also named burden tests. Statistical tests for Rare Variants CMC genotype matrix Notations SNVs ∈ {1, ..., N} individuals ∈ {1, ..., P} SNVs Cohort Allelic Sum Test compares the Combined Multivariate Weighted-Sum Statistic is analogous proportion of cases that present at j ∈ {1, ..., P} SNVs X= genotype matrix (N x P) for **one gen** Collapsing test modifies the CAST the Mann-Whitney-Wilcoxon to improve its performance when least one RV with that of controls $X_{ij} \in \{0,1,2\}$ : genotype of individual rank test, with weights on RVs both RVs and CVs are present. for the SNV j $Y_i \in \{0,1\}$ : phenotype of subject *i* Pooled Variable Threshold method is based on intuition that there exists an Adaptive Sum test introduces data Test (unknown) threshold T for which MAF threshold weights Refs Poo Permut adaptive weights to differentiate between to +/variants with MAF $\leq T$ are substantially beneficial, neutral or deleterious variants. more likely to be functional. ves СМС yes fixed no no poss.(\*) 2 ВОМР WSS yes no no yes yes aSUM no Burden Or Mutation Position test is an hybrid VT ves variable no no ves likelihood approach that combines a burden test C-Alpha test KBAC no yes yes with a test of the position distribution of variants. C-Alpho no fixed ves poss. ves C-alpha statistics tests the presence SKAT no no yes poss. poss. of a mixture of risk, protective or Kernel-based adaptive Clustering method poss SKAT-O nο groups mutation patterns across the variants nο poss. 9 BOMP hybrid no ves poss. yes $SK\Delta T$ and assigns each mutation pattern a kernelbased weight adaptively determined by data. SKAT-O Summary of properties of the tests for RVs to be compared Whether pooling over variants, using a MAF threshold to define RVs, sensitive to association directions (+/-), whether possible use of weights Generalization of C-alpha test, Sequence Kernel Association Test uses a multiple regression approach with introduction of kernel. The optimal test (SKAT-O) is a generalized family of SKAT requiring permutations for p-value calculations and references for more details tests that incorporates a correlation structure of variants effects. Data Genotypes and phenotypes generated as in (11), **6 main scenarios**, sub-divided into several scenarios, 500 independent replicates, test significant level $\alpha=5\%$ , Cardiac arrhythmias Population sampling design Targeted sequencing design 500 permutations for permutation-based methods:

**Evaluation** 

- sample size 500 cases and 500 controls:
- 8 causal RVs, p∈ {0,4,8,16,32} non causal RVs;  $\rho=0$  independent or  $\rho=0.9$  in Linkage Disequilibrium (LD).

Variant category	RV	LFV (Low Frequency Variant)	CV (Common Variant)
MAF	[0,001; 0,01]	]0,01; 0,05]	]0,05; 0,1]

Scenario 1: Independent RVs: no LD between RVs.

Scenario 2: RVs in LD: all RVs (both causal and non-causal) correlated

nario 3: No LD between causal RVs and non-causal RVs: causal RVs correlated, non-causal RVs correlated.

Scenario 4: Independent RVs and LFVs (non-causal): independent RVs and 4 non-causal LFVs Scenario 5: Independent RVs and LFVs: independent RVs and 4 LFVs (3 non-causal and 1 causal)

nario 6: Independent RVs, LFVs and CVs. independent RVs, 4 non-causal LFVs and 2 non-causal CVs.

Type I error rate analysis (OR=1 in the logistic regression model generating phenotypes: null case);

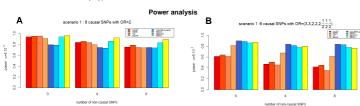
→ Empirical power analysis (OR≠1: non-null cases)

## 167 Controls 163 targeted genes 167 Cases Brugada syndrome: a rare heritable arrhythmia syndrome associated with an increased risk of sudden cardiac death (12). Calcific Aortic Valve Brugada syndrome patients (BrS) Stenosis patients (CAVS) Pipeline of data creation and analysis

After preprocessing, 135 matrices (one per studied gene) 1 to 280 variants in column and the 334 patients in line

### Results

### Scenario 1: Satisfactory Type I error rates for all tests



A: 8 deleterious causal RVs with common OR=2

- Ideal situation for pooled association tests. Their performance deteriorates when the number of non-causal RVs increases.
   KBAC is the most powerful.
- - > Pooled association tests CAST, WSS and VT suffer
  - C-Alpha, SKAT, SKAT-O and KBAC perform similarly and are most powerful

135 genes x tests for RVs - MAF threshold defining RVs = 0,01

Correlation PCA on -log<sub>10</sub>(p-values) matrix

- The first dimension remains linked to
   all studied burden tests and some non-burden tests: SKAT-O, KBAC and BOMP.
- The second dimension appears as the dimension of two non-burden tests:
- C-Alpha and its generalization SKAT.

### **Conclusion and Perspectives**

- > Simulated data: simulations are time-consuming even with the use of parallel computing! Perspectives: tests comparison for all scenarios and also for a population genetics model.
- > Real sequencing data: detection of differences between tests in a practical set-up. It reveals difficulties linked to preprocessing steps and also to the experimental design (choice of controls). Perspectives: recommendations on the use of RVs tests and strategy to detect RVs on real data.
- > The definition of a more powerful test depends on the unknown truth (of the association pattern): improvements to existing tests like adaptive tests are needed!

## References 1. Morgenthaler and Thilly (2007) Mutat. Res. 615:

 Li and Leal (2008) Am. J. Hum. Genet. 83: 311– 3. Madsen and Browning (2009). PLoS Genet. 5(2):

e1000384 4. Han and Pan (2010) *Hum. Hered.* **70**: 42-54. . Price et al. (2010) Am. J. Hum. Genet. 86: 832-

7. Liu and Leal (2010) PLoS Genet. 6. e1001156 N. Neale et al. (2011) PLoS Genet. 7, e1001322.
 Wu et al. (2011) Am. J. Hum. Genet. 89: 82–93.
 Lee et al. (2012) Am. J. Hum. Genet. 91: 224– 11. Basu and Pan (2011) Genet. Epidemiol. 35: 606-

12. Crotti et al. (2012) JACC. 60(15): 1410-1418.

