coR-ge: Investigation of Stratified False Discovery Rate Control in Environments of Complex Correlation

Christopher B. Cole[1,2,3], Joanne Knight[1,2,4,5]

[1] Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[2] Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada
[3] Biomedical Sciences Division, Department of Biology, University of Ottawa, Ottawa, Ontario, Canada
[4] Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[5] Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

Introduction

The reproducibility of Genome Wide Association Studies is dependent upon the accurate and reliable correction for many millions of simultaneous tests; limitations of conventional methodologies have led to the exploration of alternate techniques. It has been theorized that by stratifying tests based on additional information and applying False Discovery Rate Control, more real associations can be identified.[1] The validity of this approach in a complex genetic environment has never been examined.

In order to accurately and reproducible assess claims of stratified False Discovery Rate's (sFDR) efficacy, we present **coR-ge** (**cor**rection of **ge**nomes in **R**), a software package designed for the rapid and accurate examination of multiple testing correction (MTC) methodologies in varying genomic and disease models. The source code, along with a *quick start* guide is provided for reference at http://chris1221.github.io/sFDR.

We present a brief overview of **coR-ge** along with case studies examining FDR and sFDR in environments of complex genomic correlation and disease models with varying minor allele frequencies.

Methods

The program is written in R, Python, and unix shell script and requires either an SGE or a PBS cluster system running *nix. Additional dependencies and recommendations are available on the *quick start* guide. The program is structured as follows:

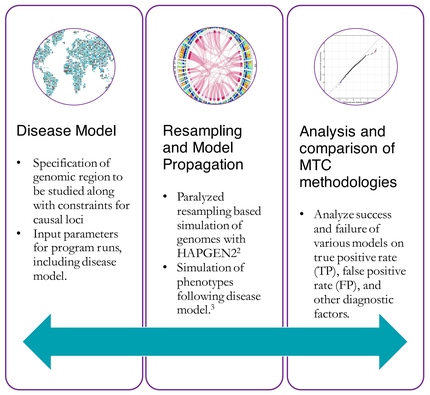

Figure 1: Logic flow behind the **coR-ge** program, source code found at http://chris1221.github.io/sFDR/

Methods (Cont.)

coR-ge takes various input parameters such as disease heritability, reference panels for LD patterns, and causal loci (either pre-specified or randomly generated with given constraints). Results can be additionally reported based on optional constraints such as linkage disequilibrium or minor allele frequencies. The program will simulate a GWAS using HAPGEN2[2] resampled genotypes and constructed phenotypes, then report various diagnostics such as True Positive rate and False Positive rate for different correction methodologies. The program maximizes the parallelization of computations in order to maximize program efficiency. **coR-ge,** when maximally parallelized runs for approximately 8 hours (n = 10,000, ~600k loci). Further development of the software will focus on documentation, user experience, and speed.

LD Structures

In order to reproducibly and accurately simulate genomes, linkage disequilibrium (LD) structures must be maintained. Figure 2. examines the similarity of pairwise R^2 values between subsequent runs of **coR-ge**.

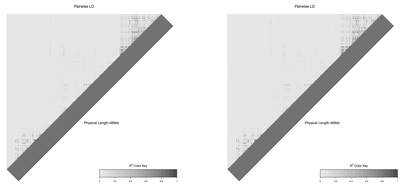

Figure 2: Comparison of linkage disequilibrium patterns between runs. 1000 Genomes CEU population build 36 used as reference sequence. Displayed are pairwise R^2 measures for each SNP in a 0.5 kb region of chromosome 1.

sFDR Methods Comparison

We present the following as a case study. Figures 3.a. and 3.b. display the differential true positive and false positive rates respectively between stratified FDR and FDR correction grouped by number of "nonsense" variants in priority strata. Displayed are the differences between the two methodologies with kernel density estimate. All simulations have been run using 1000 Genomes CEU population chromosome 1 as reference for LD patters. A disease model with 0.45 heritability and 0.55 random Gaussian noise was calculated for 100 permutations.

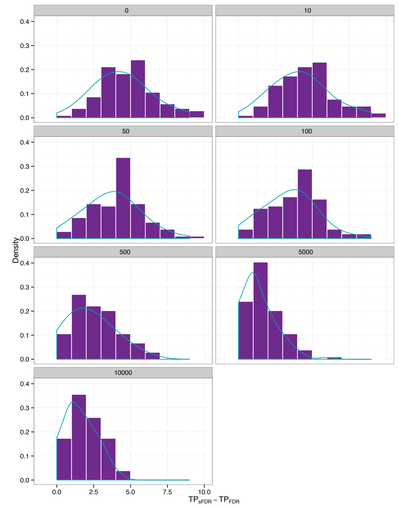

Figure 3: Improvement of true discovery rate between sFDR and FDR correction methodologies.

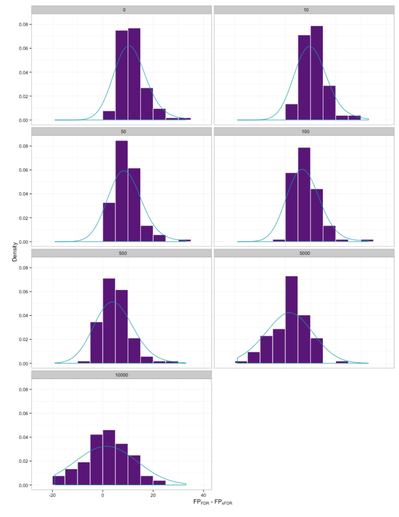

Figure 4: Decrease in false positive detection between sFDR and FDR correction methodologies.

MAF and Gene Pathways

Continuing the case study, differential effects of causal variant minor allele frequencies are examined in stratified FDR and reported grouped by number of "nonsense" variants in the priority strata. We also identify the ability of sFDR to identify variants in a realistic 10 gene pathway with differential enrichment. The inflation of stratum-specific type-1 error was found to be enrichment dependent.

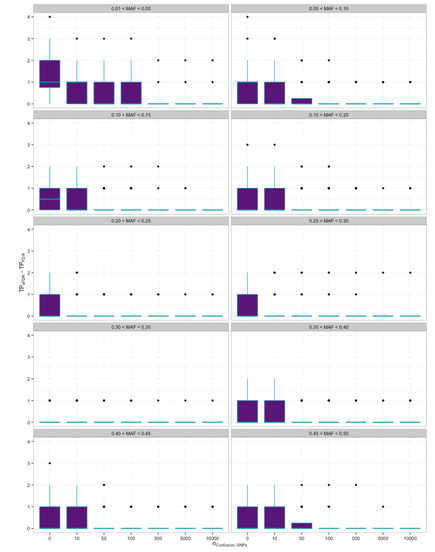

Figure 5: Differential identification of 5 causal SNPs with uniform effect facetted by minor allele frequency.

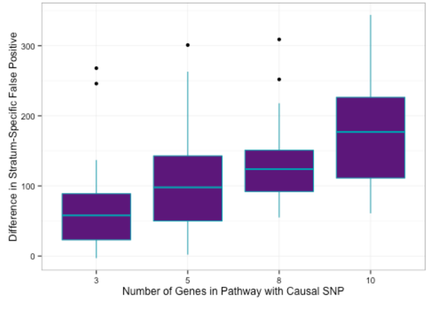

Figure 6: Inflation of stratum-specific false positive rate with increasing enrichment of gene pathway.

MAF and Gene Pathways (Cont.)

We have observed differences in the ability of sFDR to identify variants based on their minor allele frequency. sFDR identifies significantly better at lower minor allele frequencies than does aggregated FDR, though the mechanisms of this difference are unclear. Power appears to be a parabolic function of minor allele frequency, however this may be a relic of permutation and awaits validation.

When a realistic gene model is used with 10 real genes selected randomly from chromosome 1, we see more true positives with sFDR, however there is an inflation of the stratum specific type-1 error rate. This inflation appears to be a function of enrichment, and may be caused by correlation structures in the genes. This situation represents a realistic use case of sFDR.

Discussion / Conclusion

sFDR shows promise as a novel multiple testing correction methodology. In a chromosome 1 segment of the human genome, sFDR identified more true positives (paired Welch *T* test $P < 2.2 \times 10^{-16}$ for each *n*) and less false positives (paired Welch *T* test $P < 2.2 \times 10^{-16}$ for each *n*) than aggregated FDR. By examining the differential identification of true causal variants, we have observed decreased power to detect variants at low MAF, even when effect sizes are held constant. A type II ANOVA model showed an insignificant (P = 2026) difference between effect sizes in MAF groups.

In the future, the authors will utilize **coR-ge** to examine the effects of sFDR given complex disease models such as schizophrenia and cardiovascular disease. The number of false positives caused solely by correlation in stratum genes will also be examined in depth. We additionally will be working with mathematical statisticians in order to validate the heuristic trends observed in this permutation analysis.

With the continued controversy over the use and misuse of sFDR, **coR-ge** provides a computationally efficient heuristic method of determining the efficacy of different multiple testing correction methodologies.

References

1. Sun, L., Craiu, R. V, Paterson, A. D., & Bull, S. B. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, *30*(6), 519–30.
2. Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, *27*(16), 2304–2305.
3. Tune H. Pers, Juha M. Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, Sailaja Svedantam, Stefan Gustafsson, Tonu Esko, Tim Frayling, Elizabeth K. Speliotes, Genetic Investigation of ANthropometric Traits (GIANT) Con, L. F. (2015). Biological interpretation of genome-wide association studies using predicted gene functions.