coR-ge

Investigation of Stratified False Discovery Rate Control in Environments of Complex Correlation

Christopher B. Cole

June 18th 2015

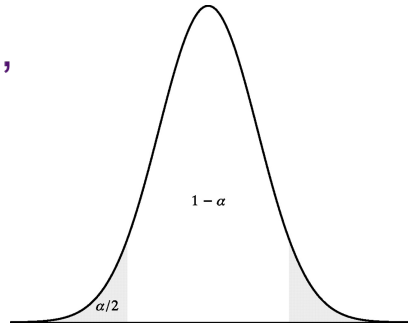
Harvey Stancer Research Day

Genome Wide Association Studies

- Agnostic search of the genome for significant associations
- Hypothesis free
- Many millions of tests
- Moving towards hypothesis driven GWAS (GWAS-HD)
  - Single SNP association with re-prioritization based on biological hypothesis
  - Stratified FDR

Multiple Testing Correction Primer

- If H_0 is true, $P(|Z| < \phi(\alpha)) = 1 - \alpha$
  - i.e. if "significant" P value is 0.05, given H_0 is true, chance of correctly failing to reject is 0.95
- If A and B are independent, $P(A\ and\ B) = P(A)P(B)$
- Probability of all true H_0 in correct region, for n Independent tests, is 0.95^n
- Increased number of Type 1 Error



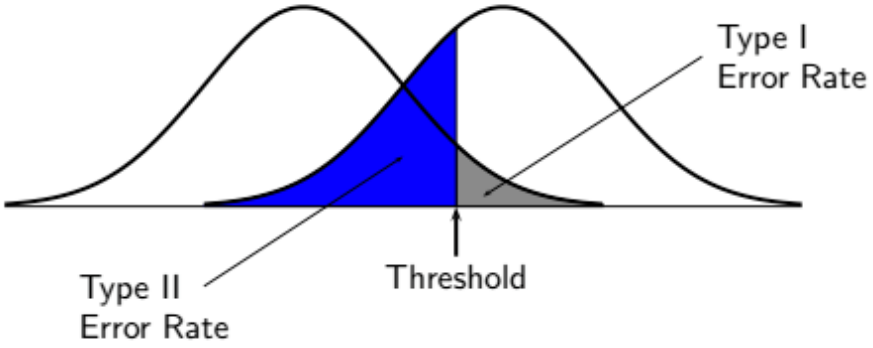
Methods

Multiple Testing Correction Primer

- Significance Level (α) is 0.05
- 3 false tests
  - E.g. Number of jelly beans eaten influences musical ability
  - Probability of each false test being "not significant"
    = $P(Null\ Region) \cdot P(Null\ Region) \cdot P(Null\ Region)$
    = $0.95 \cdot 0.95 \cdot 0.95$
    = 0.857375
- 10 false tests
  - $= 0.95^{10} = 0.5987$

Methods

Multiple Testing Correction Primer



- Type I Error: Reject H_0 when H_0 is true (False Positive)
- Type II Error: Fail to reject H_0 when H_0 is false (False Negative)

Methods

Multiple Testing Correction Primer

What does this really mean?
- True positive: Causal variant is there, and is significant
- False positive: Not a causal variant, and is significant

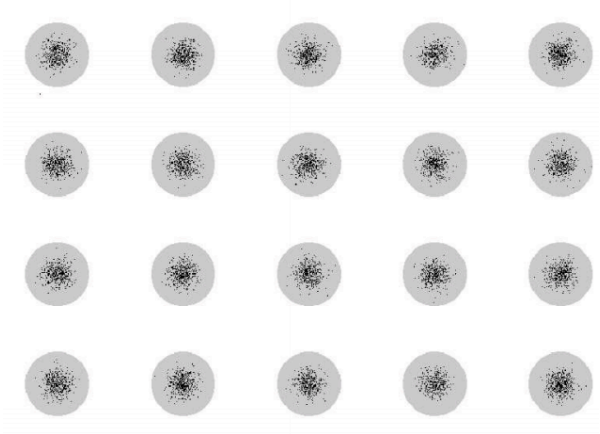
Why do we care?
- More TP = More loci that can be replicated.
- Power to detect TP > Cost of detecting FP.

Multiple Testing Correction Primer

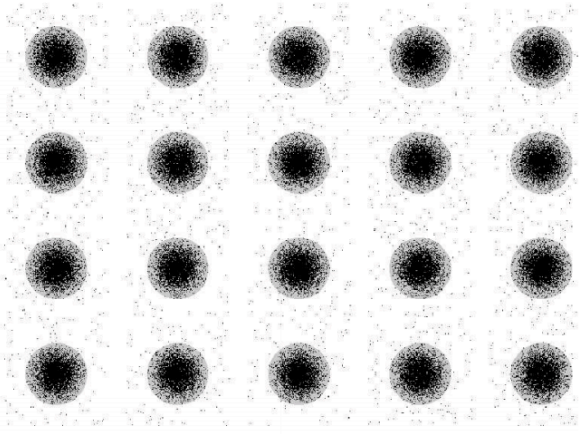
- As ↑ n_{tests}, ↑ False positives
- Control methods
  - Bonferroni
    - Chance of even one false positive $\leq \alpha$
    - $P(FP > 0) \leq \alpha$
  - FDR
    - Proportion of discoveries which are false $\leq \alpha$
    - $FDR \equiv E\left[\frac{FP}{TP}\right] \leq \alpha$

Methods

Multiple Testing Correction Primer



Family-wise Error Rate (Bonferroni)
Chance of even one false positive $\leq \alpha$

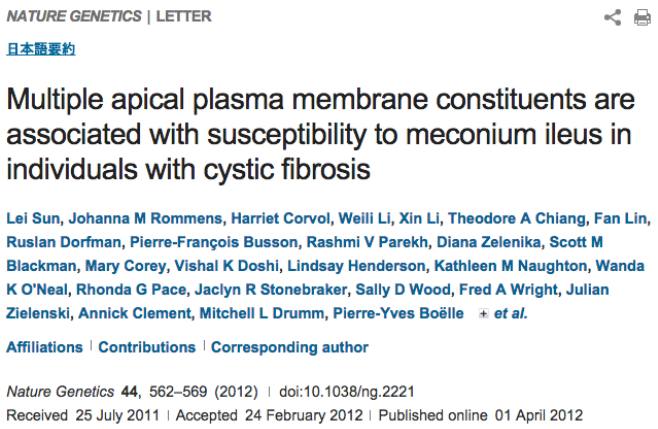


False Discovery Rate Control
Proportion of FP $\left(\frac{FP}{TP}\right) \leq \alpha$

Analysis

Stratified FDR

- Sun et. al 2006
- Control FDR separately in different groups
- According to prior information
  - E.g. Cystic Fibrosis Genes
  - Meconium Ileus
- Lots of questions
  - Effect of correlation etc.



NATURE GENETICS | LETTER

日本語要約

Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis

Lei Sun, Johanna M Rommens, Harriet Corvol, Weili Li, Xin Li, Theodore A Chiang, Fan Lin, Ruslan Dorfman, Pierre-François Busson, Rashmi V Parekh, Diana Zelenika, Scott M Blackman, Mary Corey, Vishal K Doshi, Lindsay Henderson, Kathleen M Naughton, Wanda K O'Neal, Rhonda G Pace, Jaclyn R Stonebraker, Sally D Wood, Fred A Wright, Julian Zielenski, Annick Clement, Mitchell L Drumm, Pierre-Yves Boëlle + *et al.*

Affiliations | Contributions | Corresponding author

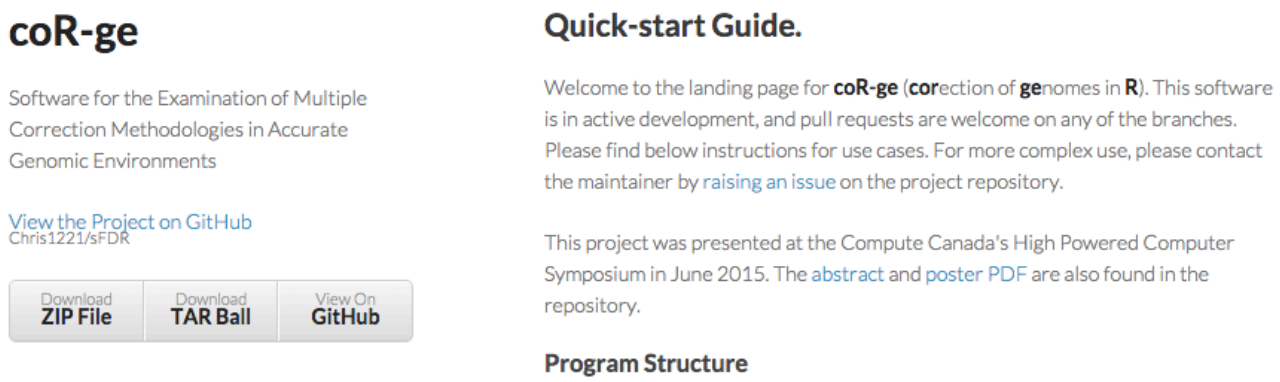
Nature Genetics **44**, 562–569 (2012) | doi:10.1038/ng.2221
Received 25 July 2011 | Accepted 24 February 2012 | Published online 01 April 2012

Introducing: **coR-ge**

- Software for the Examination of Multiple Correction Methodologies in Accurate Genomic Environments
- Permutation testing of correction methodologies
- Different environments
- Open source: http://chris1221.github.io/coR-ge/
- ~ 5000 lines of code, fully parallelised on SGE (HPCVL) and PBS (SCINET) clusters.

Analysis

Introducing: **coR-ge**



coR-ge

Software for the Examination of Multiple Correction Methodologies in Accurate Genomic Environments

View the Project on GitHub
Chris1221/sFDR

Download ZIP File | Download TAR Ball | View On GitHub

Quick-start Guide.

Welcome to the landing page for **coR-ge** (**cor**ection of **ge**nomes in **R**). This software is in active development, and pull requests are welcome on any of the branches. Please find below instructions for use cases. For more complex use, please contact the maintainer by raising an issue on the project repository.

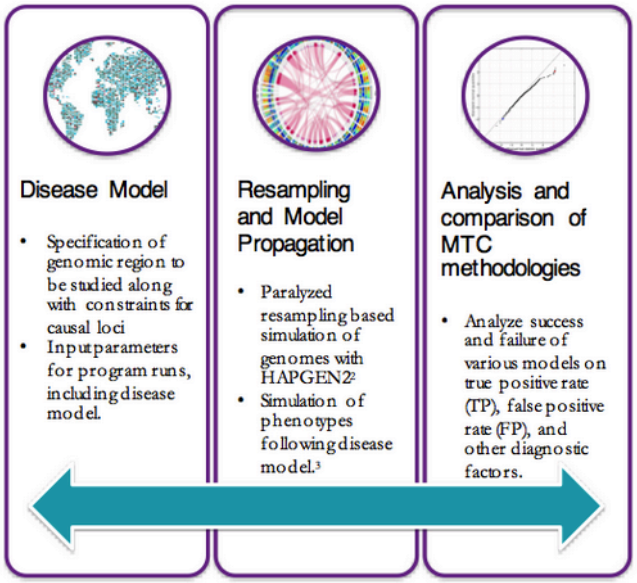
This project was presented at the Compute Canada's High Powered Computer Symposium in June 2015. The abstract and poster PDF are also found in the repository.

Program Structure

- http://chris1221.github.io/coR-ge/

Analysis

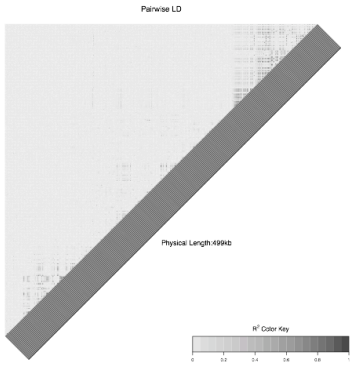
Introducing: **coR-ge**

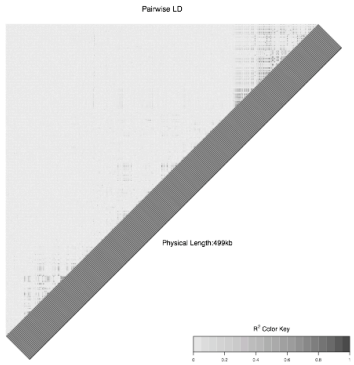


Analysis

Introducing: **coR-ge**

- Maintain LD structures
  - Pairwise R^2
- Specify Disease Model
  - Heritability
  - Causal loci
  - Phenotype Generation

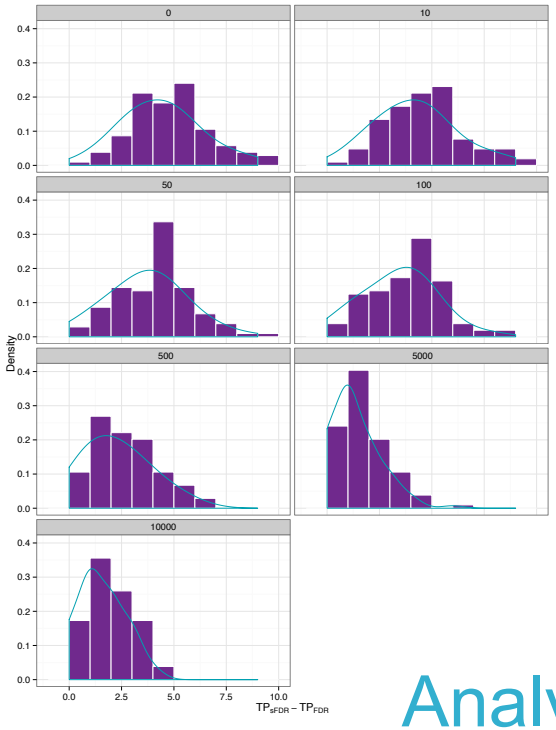




Analysis

Introducing: **coR-ge**

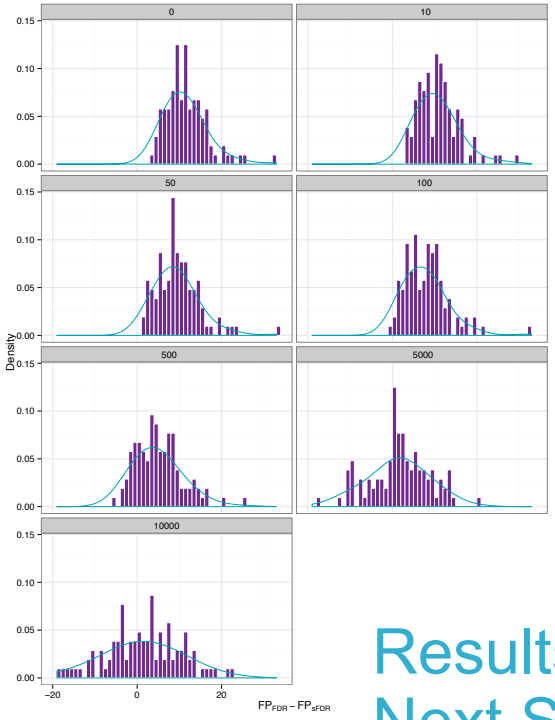
- Heritability: 0.45
- 50 Causal SNPs randomly distributed
- Normal phenotype with $Z(0,0.55)$ noise
- Δ True Positives between FDR and sFDR
- Grouped by number of “confusion” SNPs



Analysis

Introducing: **coR-ge**

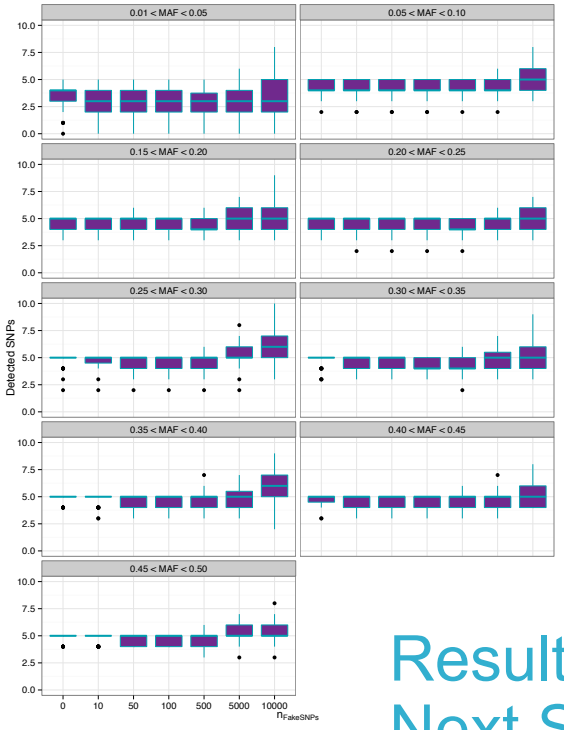
- Heritability: 0.45
- 50 Causal SNPs randomly distributed
- Normal phenotype with $Z(0,0.55)$ noise
- Δ False Positives between FDR and sFDR
- Grouped by number of “confusion” SNPs



Results & Next Steps

Introducing: **coR-ge**

- Heritability: 0.45
- 5 causal loci in each MAF group
- Normal phenotype with $Z(0,0.55)$ noise
- Effect sizes evenly
  - Type II ANOVA for difference P = 0.2026
- Δ True positives between FDR and sFDR
- Grouped by number of “confusion” SNPs
- **Grouped by MAF category**



Results & Next Steps

Introducing: **coR-ge**

- Heritability: 0.45
- 10 gene system
- 1 causal loci in each gene
- Δ True positives between FDR and sFDR
- Δ False positives between FDR and sFDR

Results &
Next Steps

Summary

- Introduced a software suite for the comparison of multiple testing correction methodologies
- Showed preliminary evidence of sFDR > FDR
  - More TP
  - Less FP
  - When real gene group used, more TP and more FP.

Results & Next Steps

Next steps:

- Improved user documentation
- Realistic examinations of correlation structures
- Working with mathematical statisticians to prove these heuristic trends

Results &
Next Steps

Acknowledgements

- Dr. Joanne Knight
- Sarah Gagliano / Sejal Patel
- Hartmut Schemider
- OxStatGen / Hadley Wickham

NSERC
CRSNG

Institute of Medical Science
UNIVERSITY OF TORONTO

Campbell Family
Mental Health Research Institute

Results & Next Steps

camh
Centre for Addiction and Mental Health