Bayesian Approaches to Functional Integration of Genomic Data

JINGJING YANG, PHD

DEPARTMENT OF HUMAN GENETICS

EMORY UNIVERSITY SCHOOL OF MEDICINE

Outline

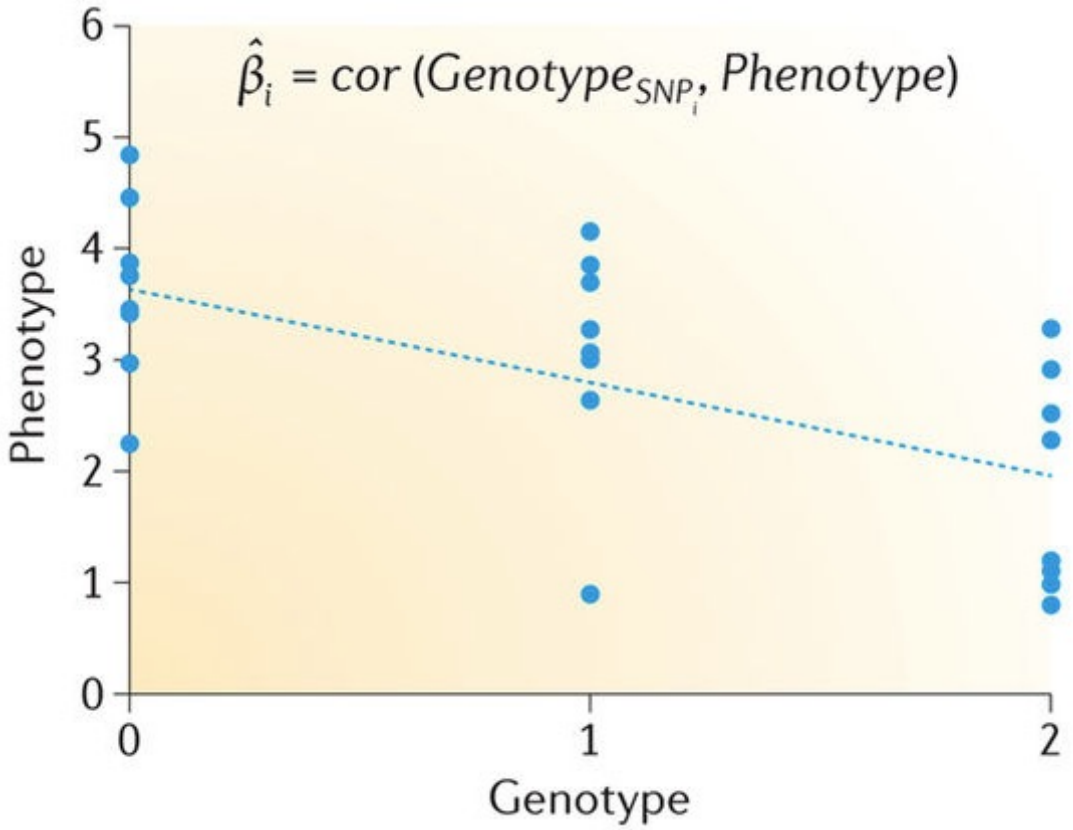
- Introduction of Genome-Wide Association Study (GWAS)
- Integrate Functional Information in GWAS
- Integrate Transcriptomics Data
- Summary and Ongoing Research

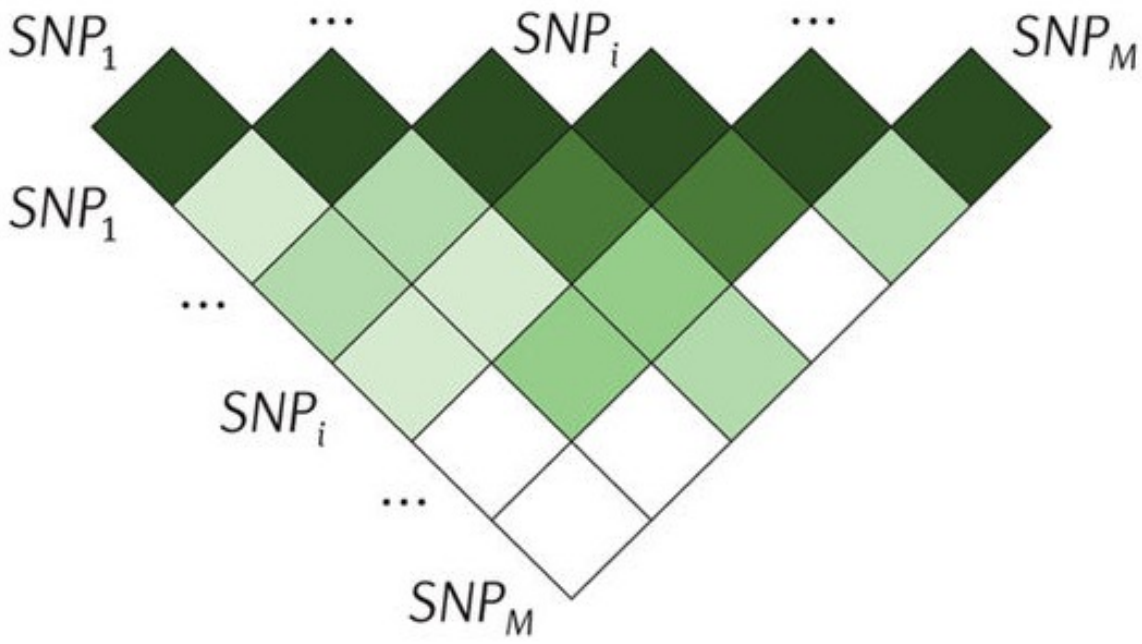
❖ Introduction of Genome-Wide Association Study (GWAS)

GWAS for Complex Traits

- Genotype data G
- Phenotype data Y (case/control or quantitative), e.g., disease status, height
- Covariates Z (age, gender, BMI, etc.)
- Standard GWAS tests if $\beta_i = 0$: $Y \sim \alpha Z + \beta_i G_i,\ G_i \in (0, 1, 2)$
- Significant P-value threshold: 5×10^{-8}
- Successfully identified > 58K unique SNP-Trait associations, based on the report on GWAS Catalog, 02/07/2018

GWAS





Pasaniuc B. & Price A. L., Nat. Rev., 2017

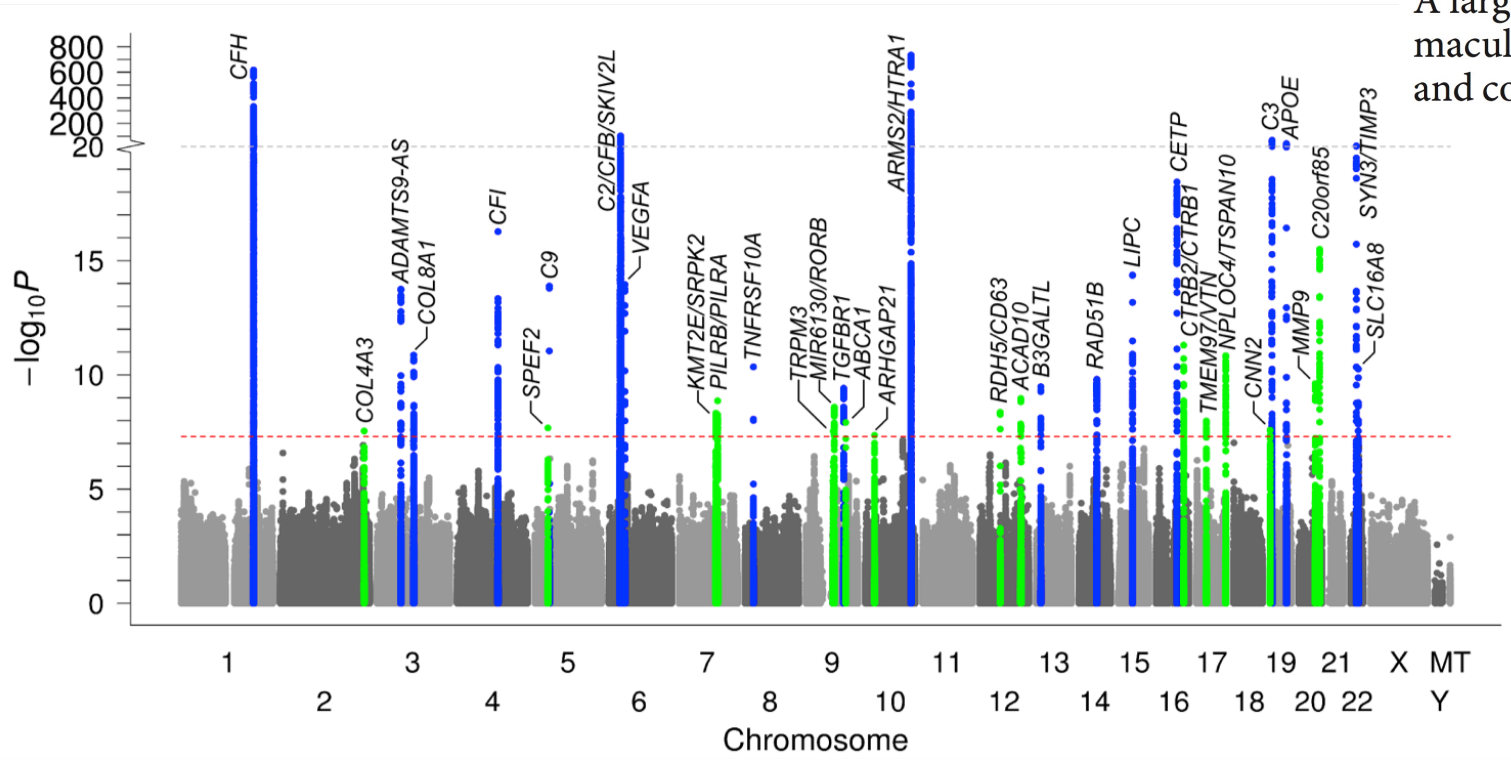
Age-related Macular Degeneration (AMD)

One of the leading causes of blindness in elderly people (ages > 60)

- Risk factors include Smoking, Diet, and Genetics
- Seddon et al. (2005) estimated Heritability 46%~71% from the US twin study

From National Eye Institute https://www.nei.nih.gov/photo/

Limitations of GWAS



A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants

Fritsche L.G. et. al. Nature Genetics, 2016.

Sample Size: 16,144 cases vs. 17,832 controls

Identified 52 independently associated variants distributed across 34 loci

Majority of the associated variants are of unknown functions …

How to fine-map functional associations?

Standard Fine-mapping Approach

Sequential Forward Selection

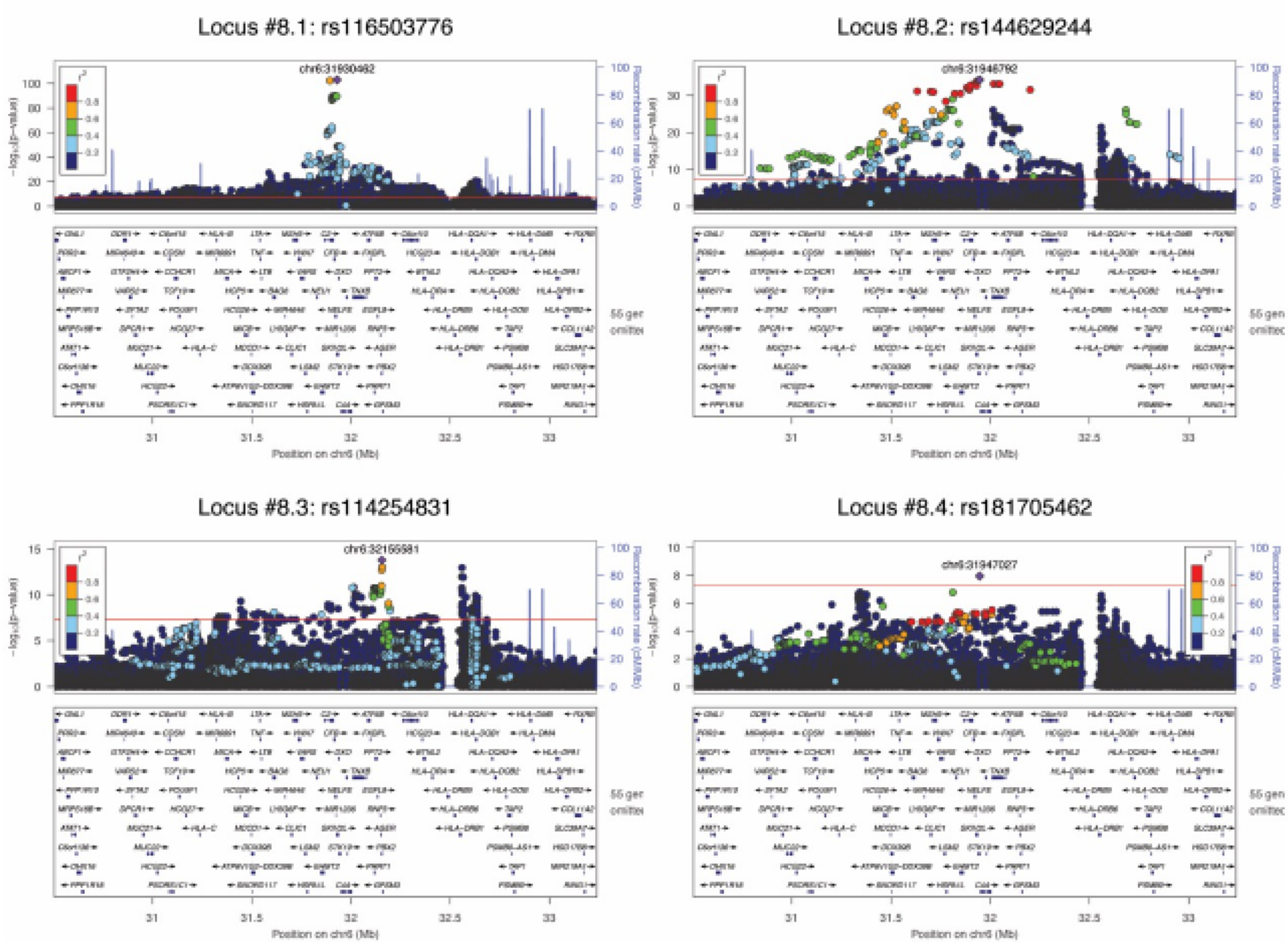
Aim: Within each region of interest, identify all statistically independent variants

1. Select variant with smallest P value ($P < 5x10^{-8}$), write into results file

2. Conduct region-wide association analysis conditioning on variants in results file

3. From the results of 2., if smallest $P < 5x10^{-8}$, select variant write into results file; otherwise stop

4. Repeat 2. and 3.



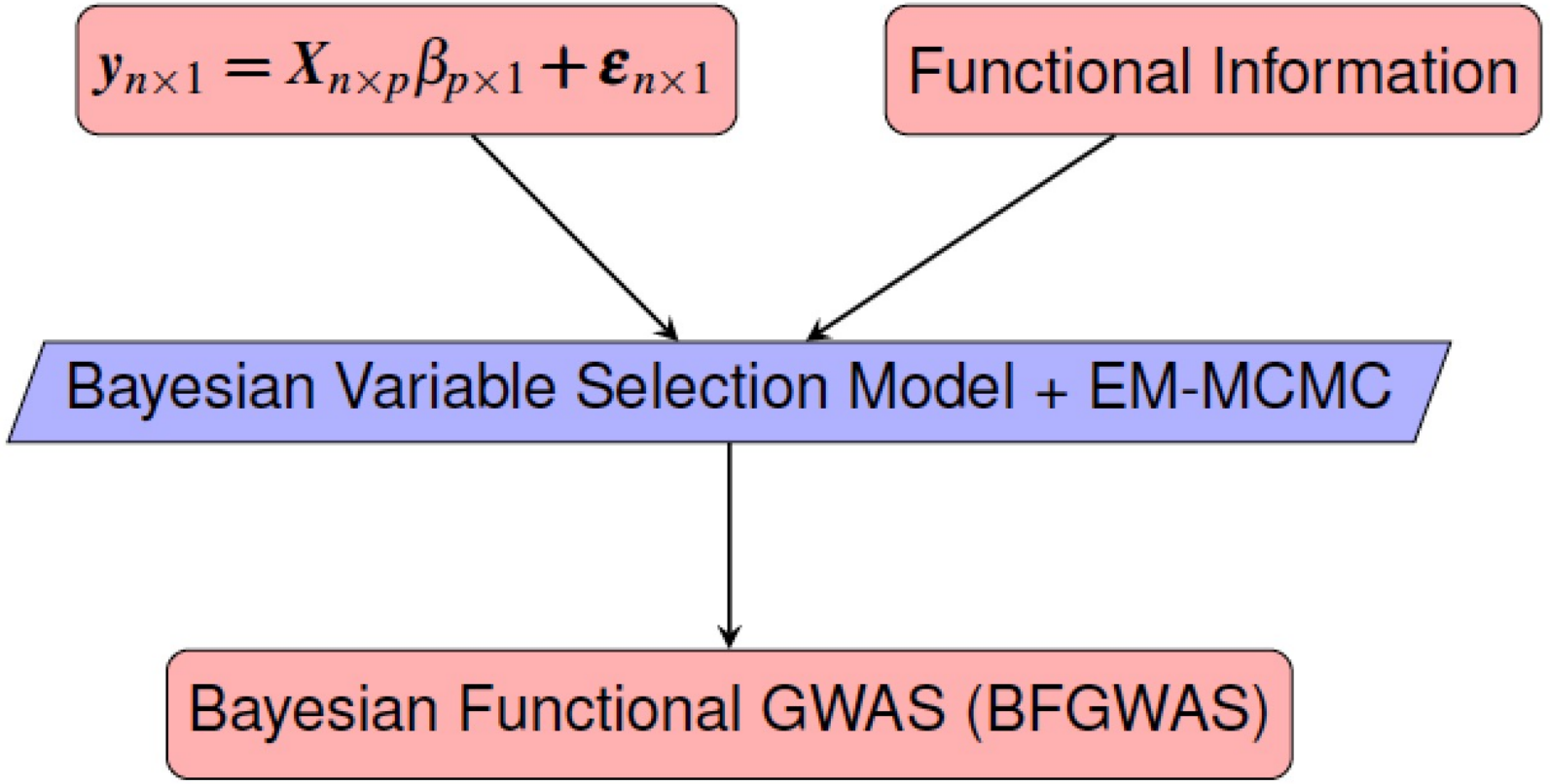
Example LocusZoom plots made by Fritsche L

Motivations

- Understand biological mechanisms for genetic association studies
- Account for linkage disequilibrium (LD) for fine-mapping “causal” candidate signals
- Integrate functional information in GWAS
- Use summary statistics for analysis convenience and computational efficiency

❖ Integrate Functional Information in GWAS

Method Diagram



Bayesian Hierarchical Model

Joint Linear Regression Model

$$\boldsymbol{y_{n\times 1} = G_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}}, \qquad \boldsymbol{\epsilon \sim MVN(0, \tau^{-1} I)}$$

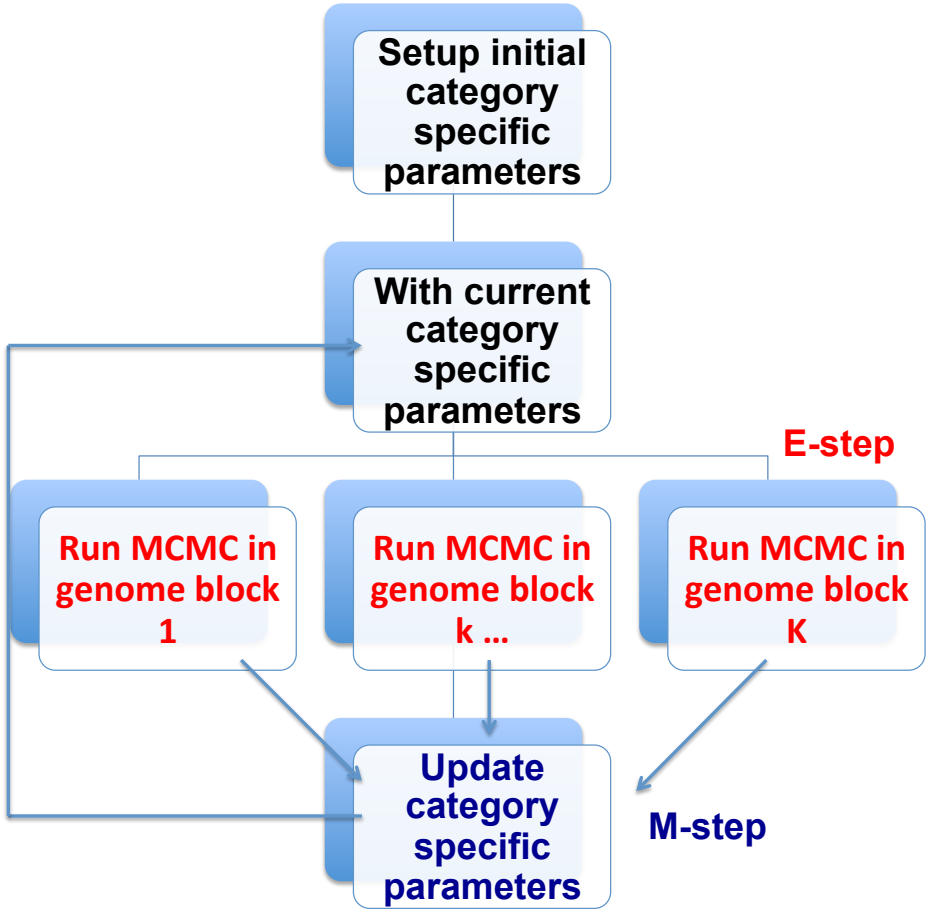
Annotate genome-wide variants into multiple non-overlapped categories

Assuming category-specific **Spike-and-Slab** prior for effect-sizes

$$\boldsymbol{\beta_{i_q} \sim \pi_q N\left(0, \tau^{-1}\sigma_q^2\right) + \left(1 - \pi_q\right)\delta_0}$$ for variant i of annotation q

Goal: estimate $\{\pi_q, \sigma_q^2, \beta_i, E[\beta_i \neq 0]\}$

EM-MCMC Algorithm



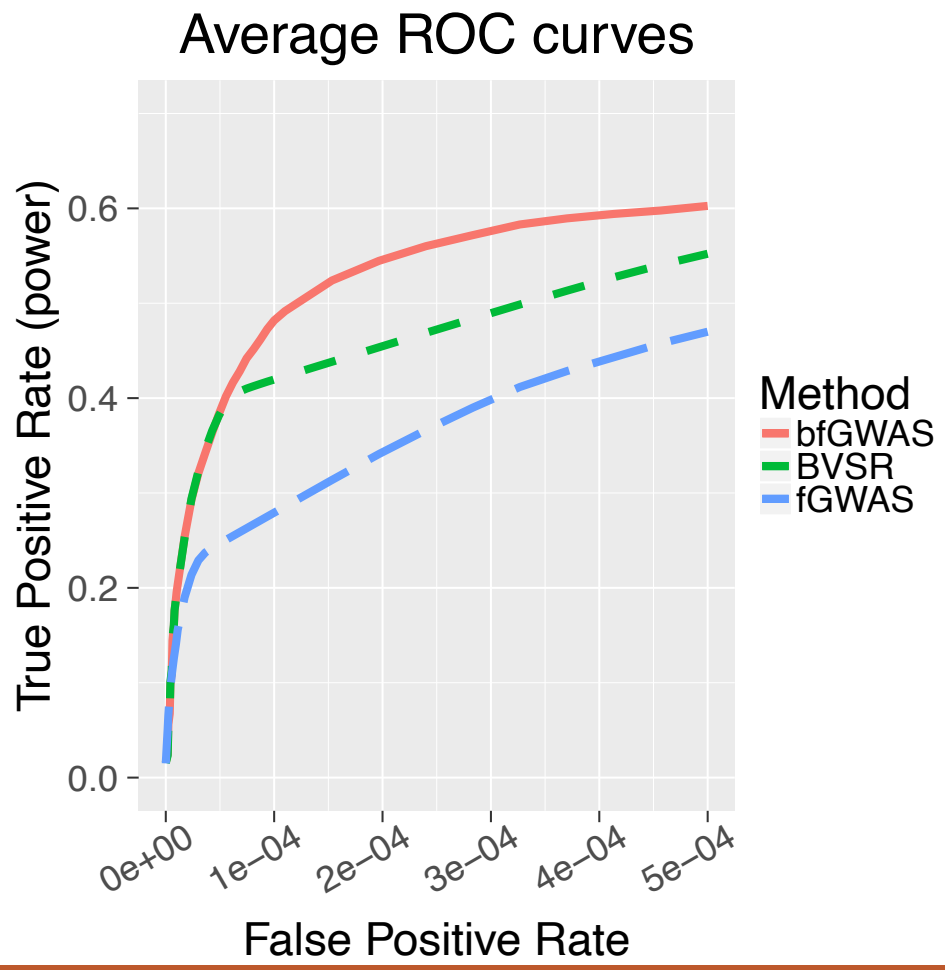
Enabled genome-wide analysis

Improved MCMC convergence rate

Simulation Study

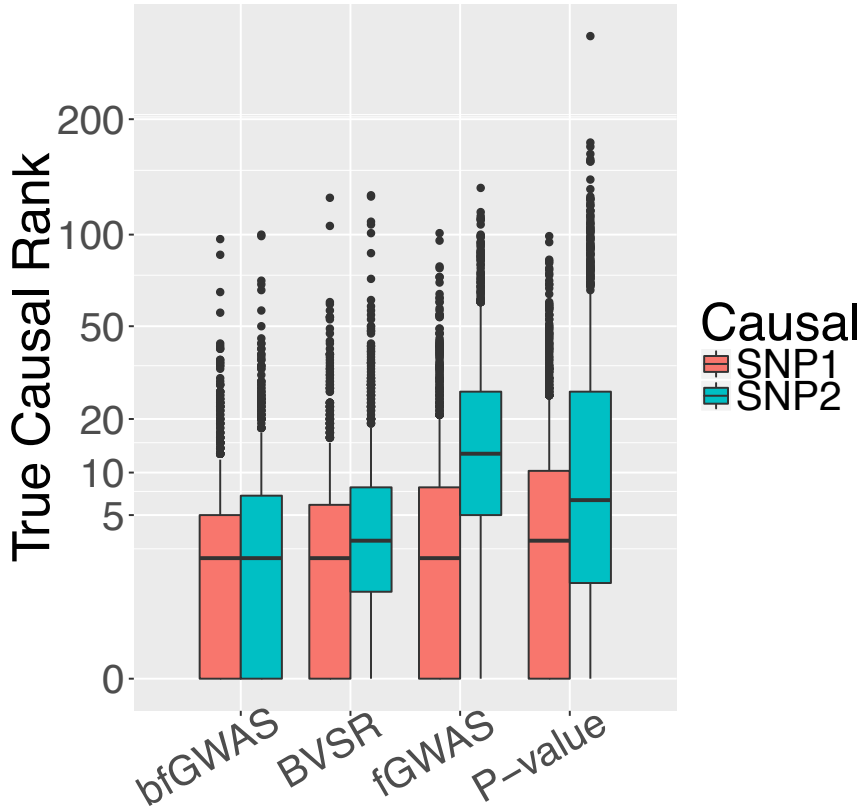
- Real genotype data from the AMD GWAS (100 x 5,000 variants)
- Two complementary annotations, “coding” and “noncoding”, following the pattern observed in the real AMD data
- Two causal SNPs in LD for 10% genome-block
- 53x enrichment for the “coding” variants
- Quantitative traits with a total 15% heritability equally explained by 20 causal SNPs

Highest Power by BFGWAS



Results of 100 repeated simulations

Highest Power to Fine-map Multiple Causals in LD



SNP1: True causal with more significant P-value

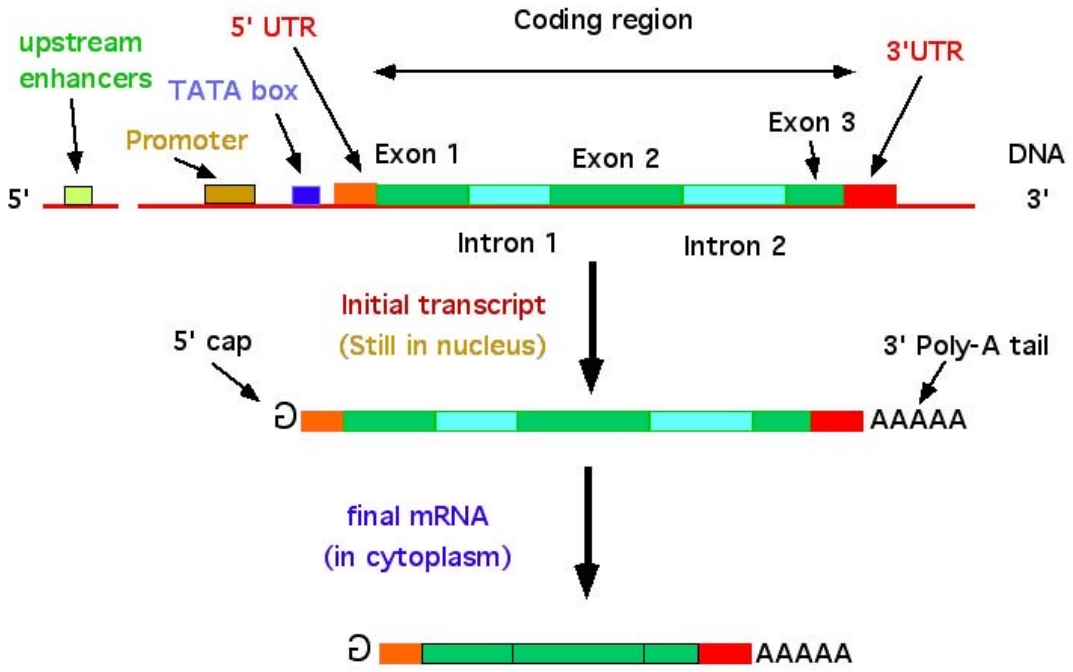
SNP2: Second true causal

Higher rank (smaller value) suggest higher power

Apply on the AMD GWAS Data

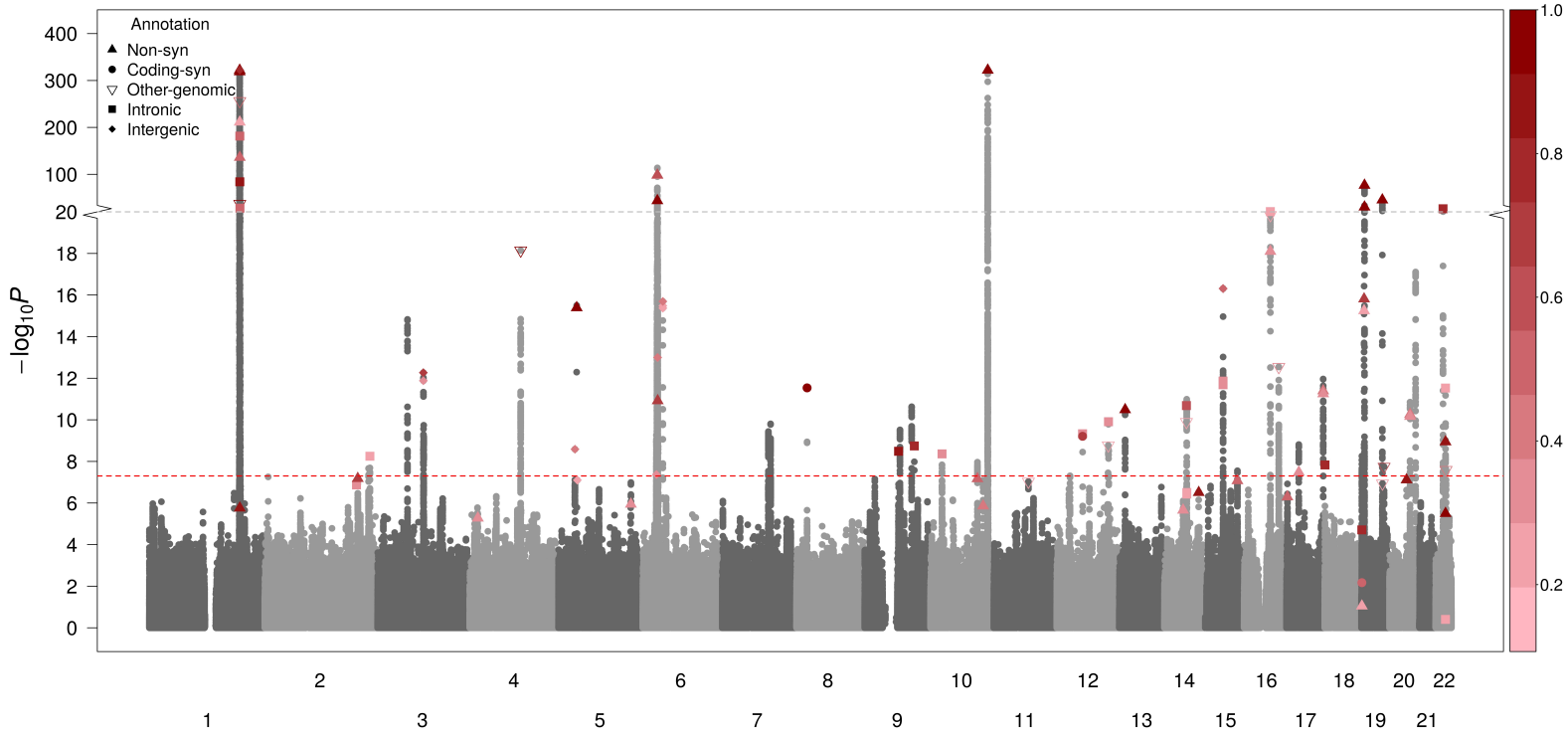
Integrate functional information annotated by SeattleSeq

- Non-synonymous (42,005)
- Synonymous (67,165)
- Intronic (3,678,235)
- Integenic (5,512,423)
- Other genomic (565,916, UTR, non-coding exons, upstream and downstream)



http://nitro.biosci.Arizona.edu

BFGWAS Results



Yang J. et.al, AJHG, 2017

Example ZoomLocus Plot

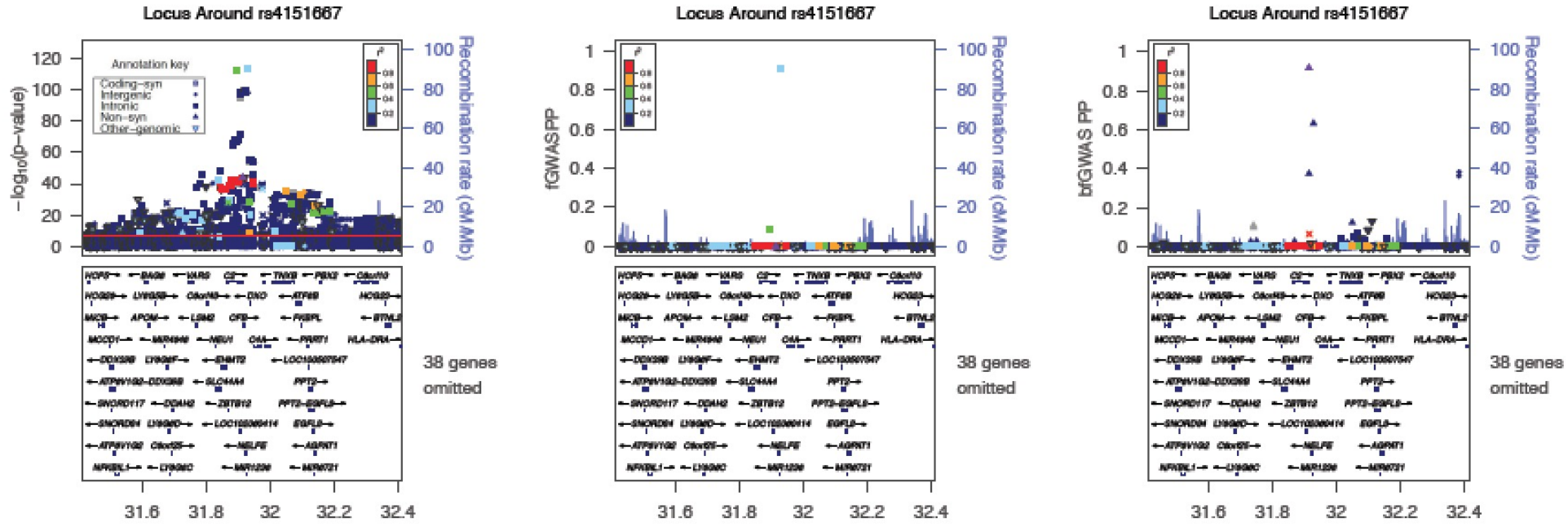


Figure 3: GWAS (left) vs. FGWAS (middle; Pickrell JK, AJHG 2014) vs. BFGWAS (right) for example locus #8.

Enrichment Results

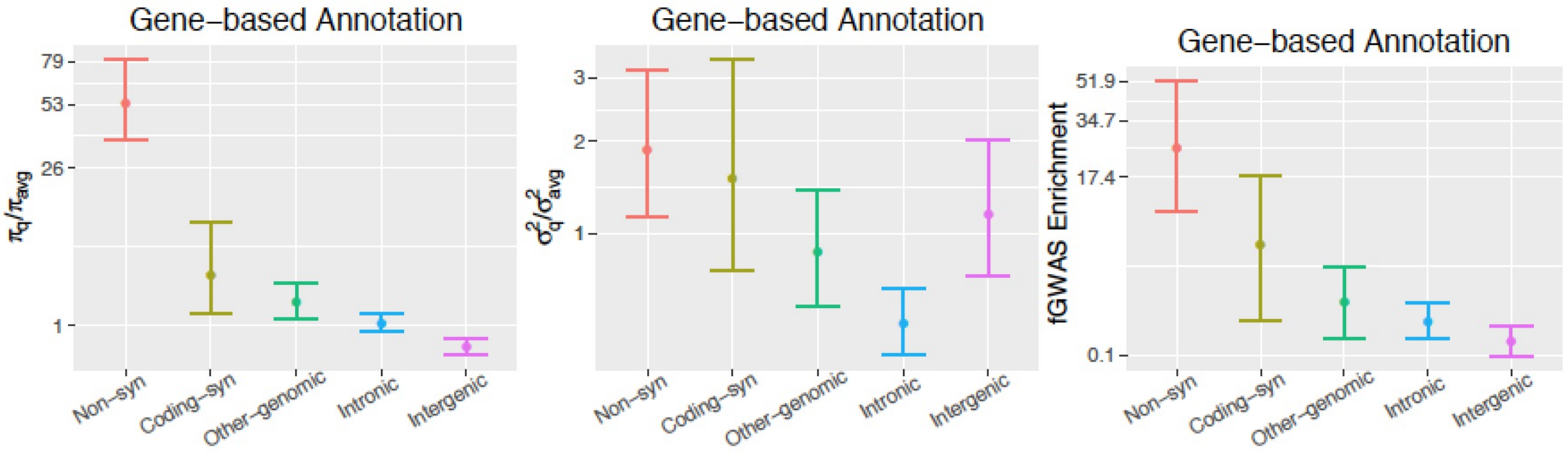


Figure 5: BFGWAS enrichment Results (left, middle) vs. FGWAS (right).

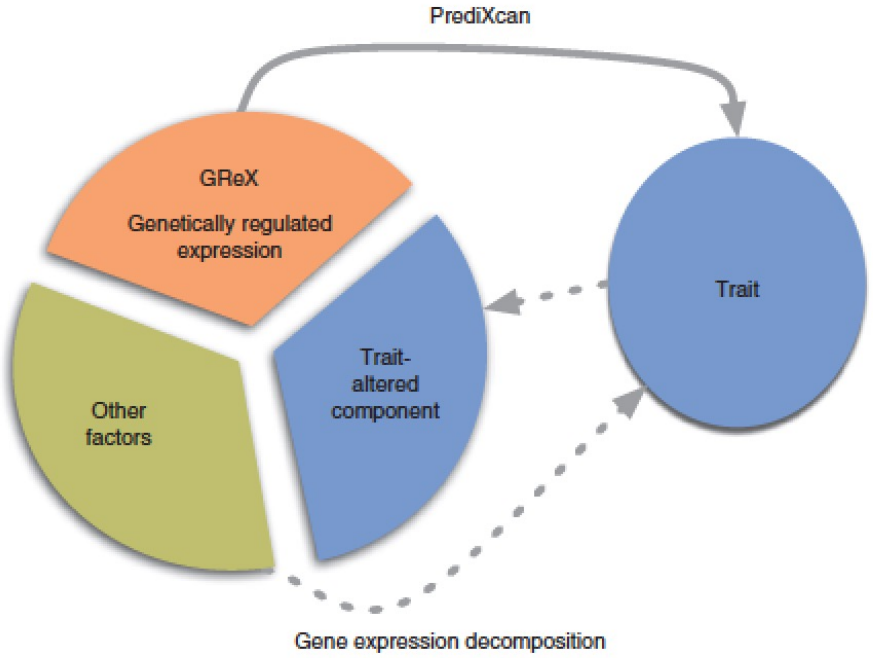
❖ Integrate Transcriptomics Data

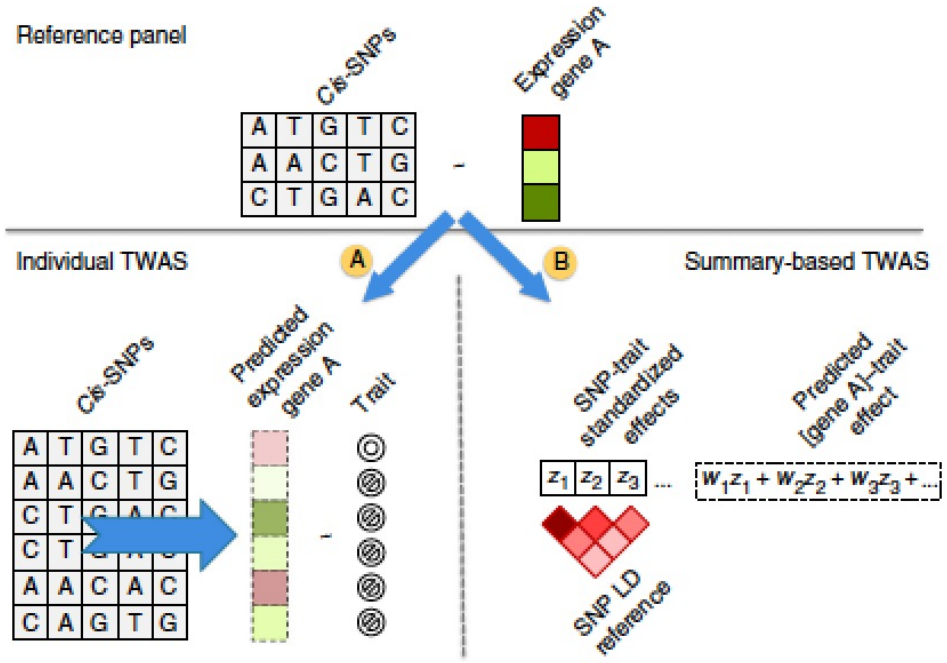
Transcriptomic Profiling

- Function element of the genome: transcriptome, comprised of different kinds of RNA molecules, e.g., mRNA, miRNA, etc.
- RNA Sequencing (RNAseq)
- Fine-map for GWAS signals that also participate in RNA regulations?
- Difficulties:
  - Tissue availability
  - Sequence Cost
  - Time cost
  - Nearly impossible for large GWAS with 10,000-1,000,000 samples
  - Efficient statistical methods for integrative analysis

Impute genetically regulated expression

- PrediXcan (Gamazon E.R. et.al., NG, 2015)
- TWAS (Gusev A. et.al., NG, 2016)



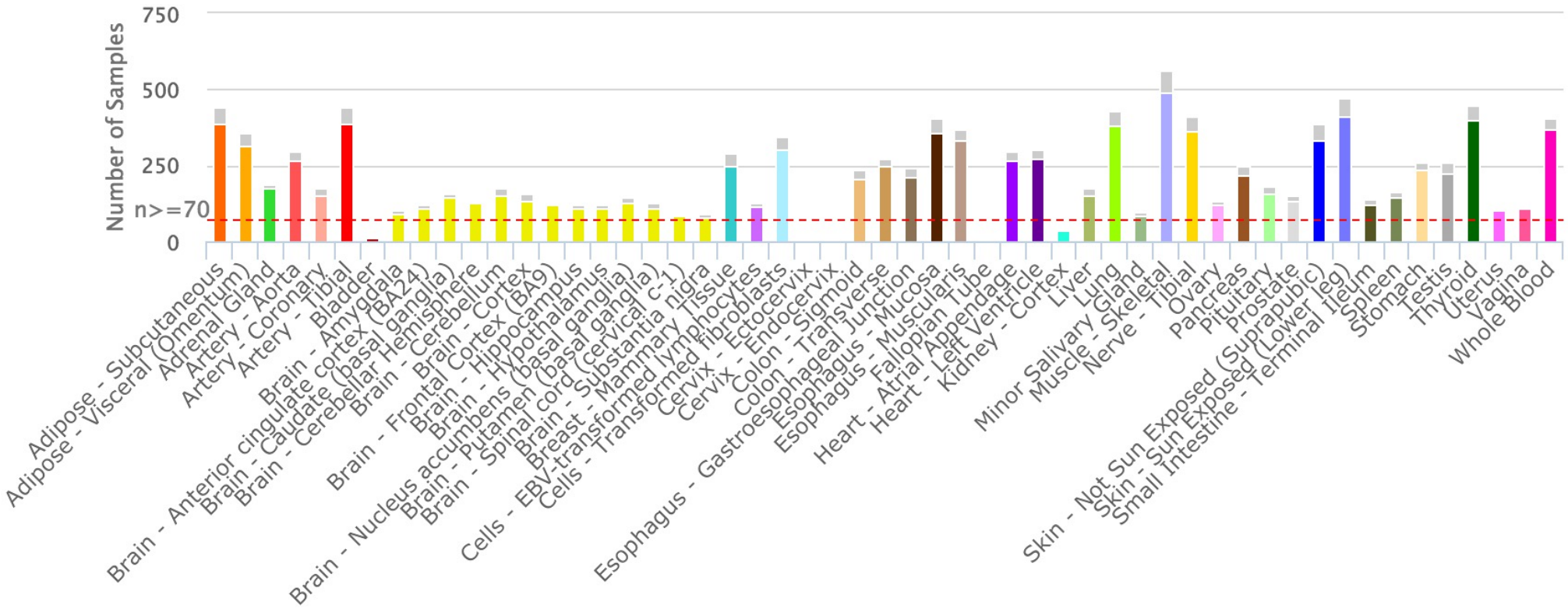


Underlying Methodology

- Gene Expression ~ cis-eQTL genotypes
- GTEx Reference Data
- Elastic Net Regression Model
- Linear Mixed Model
- Imputation quality depends on the heritability among identified cis-eQTLs

Gamazon E.R. et.al., NG, 2015

GTEx V7 Sample Counts by Tissues



GTExPortal: https://www.gtexportal.org/home/

Imputation Advantages

- Leverage large sample sizes with genotype information (increase power)
- Computational cost is much cheaper than actual experiments
- Provide small studies opportunities to study transcriptomic profiles
- Help identify functional association signals
- Help understand the underlying biology of GWAS signals

Caveats of Current Methods

- Discrepancy between the heritability among cis-eQTLs vs. genome-wide SNPs?
- Imputation accuracy?
- PrediXcan/TWAS were advocated as gene-based association methods: Traits ~ Imputed Gene Expression
- Integrate gene-expression data in GWAS?

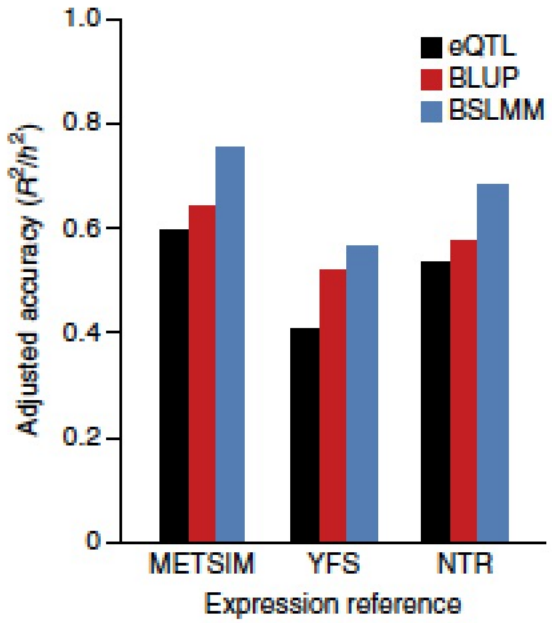


Figure 3 Accuracy of individual-level expression imputation algorithms. Adjusted accuracy was estimated using cross-validation R^2 between predicted and true expression and normalizing by corresponding *cis*-h^2_g. Bars show the mean estimate across three cohorts and three methods: eQTL, single best *cis*-eQTL in the locus; BLUP, using all SNPs in the locus; and BSLMM, using all SNPs in the locus and noninfinitesimal priors.

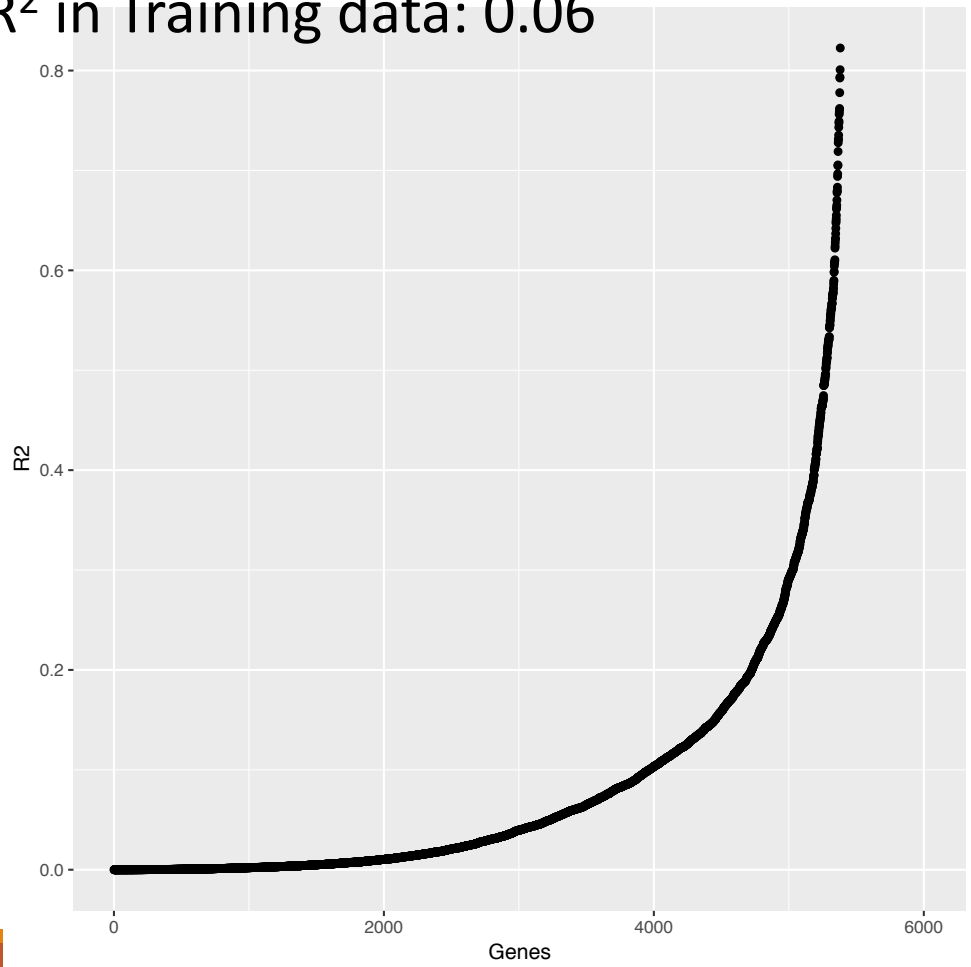
Gusev A. et.al., NG, 2016

Application on the ROS/MAP Data

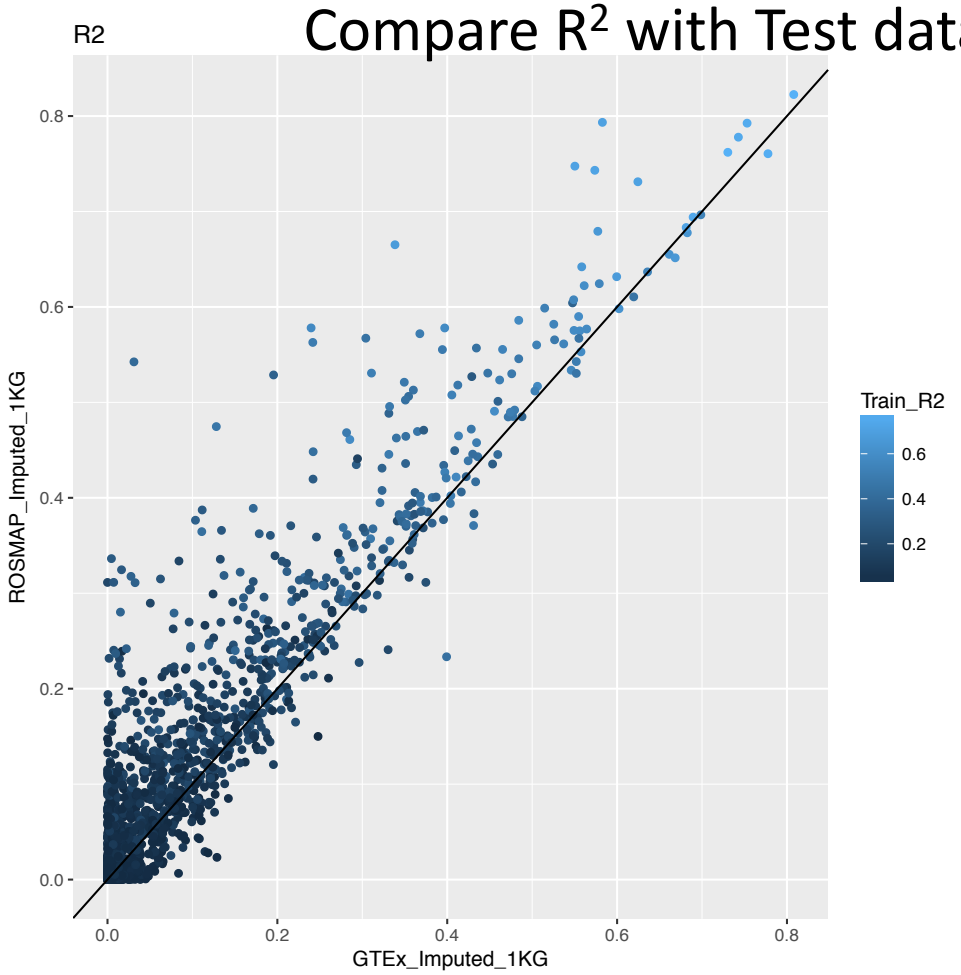
- ROS/MAP study (Bennett D. et al., 2012) for Alzheimer's Disease
- RNAseq data for 499 Samples
- Genotype data for N=2,000 (19% Alzheimer's Disease cases)
- Separated 499 Samples into a Training set (2/3 samples) and Test set (1/3 samples)
- PrediXcan model fitted using the ROS/MAP Training data vs. fitted using the GTEx Reference data

PrediXcan Imputation Results

Mean R^2 in Training data: 0.06



Compare R^2 with Test data



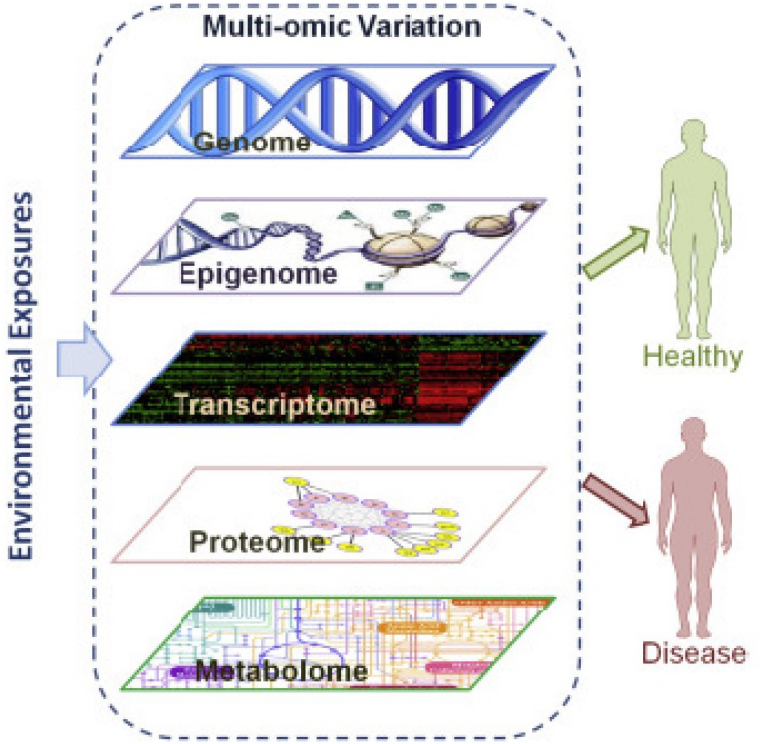
❖ Summary and Ongoing Research

Summary

- BFGWAS (freely available at https://github.com/yjingj/bfGWAS)
  - integrates functional information in GWAS while accounting for LD
  - Computational efficient EM-MCMC algorithm
  - Provides a list of risk loci and fine-mapped association candidates, as well as enrichment results
- Integrating transcriptomics data is useful but challenging
  - Consider genome-wide heritability for imputing genetically regulated gene-expression
  - Consider phenotypes and covariates in imputation
  - Advanced statistical method for integrative analysis with GWAS data

Ongoing Research

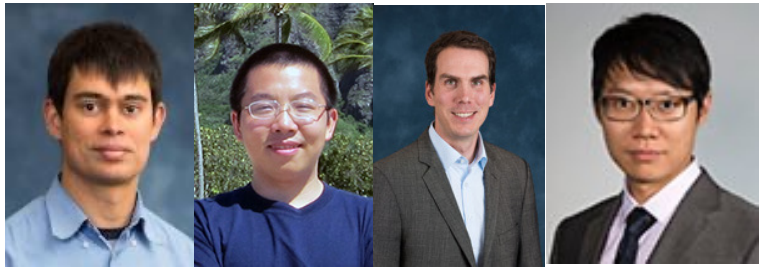
- Improve imputation accuracy for genetically regulated gene-expression
- Integrate gene-expression data into GWAS using BFGWAS framework
- Apply on the ROS/MAP data for studying Alzheimer's disease
- Extend the methodology for other omcis data, e.g., epigenomics and proteomics
- Recruiting motivated postdocs and graduate students



From Sun, Y. and Hu, Y. (2016).

Acknowledgements

University of Michigan

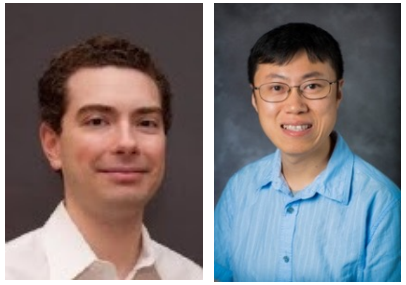


UNIVERSITY OF MICHIGAN

International AMD Genomics Consortium

International Age-related Macular Degeneration Genomics Consortium

Emory University



Dr. Aliza Wingo

EMORY UNIVERSITY SCHOOL OF MEDICINE | Department of Human Genetics

RADC Research Resource Sharing Hub

Rush Alzheimer's Disease Center (RADC)