Scalable Bayesian Method for Functional Genome-wide Association Studies

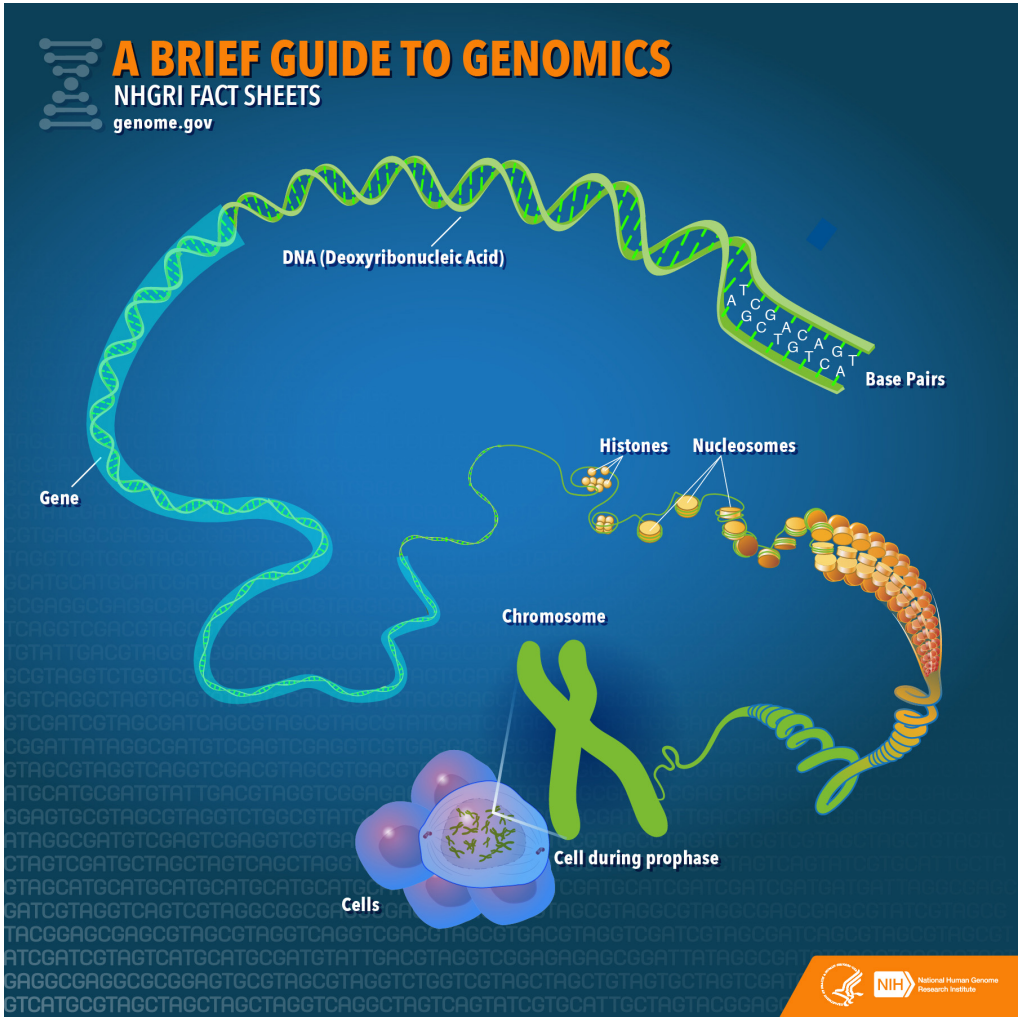
Jingjing Yang

Department of Human Genetics
Emory University School of Medicine

Outline

A Brief Guide to Genomics

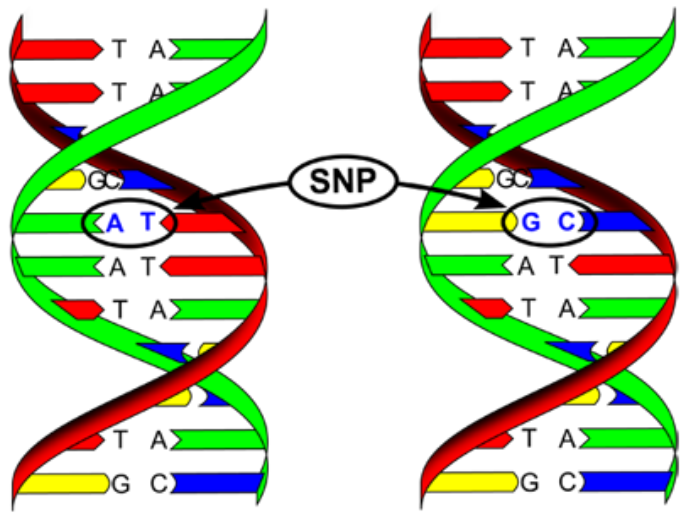
- Deoxyribonucleic acid (DNA) molecules are made of a double helix
- Each DNA strand is made of four nucleotides — Adenine (A), Thymine (T), Guanine (G), and Cytosine (C)
- The Microarray or Sequencing technology allows us to identify the nucleotide type (A, T, G, or C) along the DNA chain



NHGRI

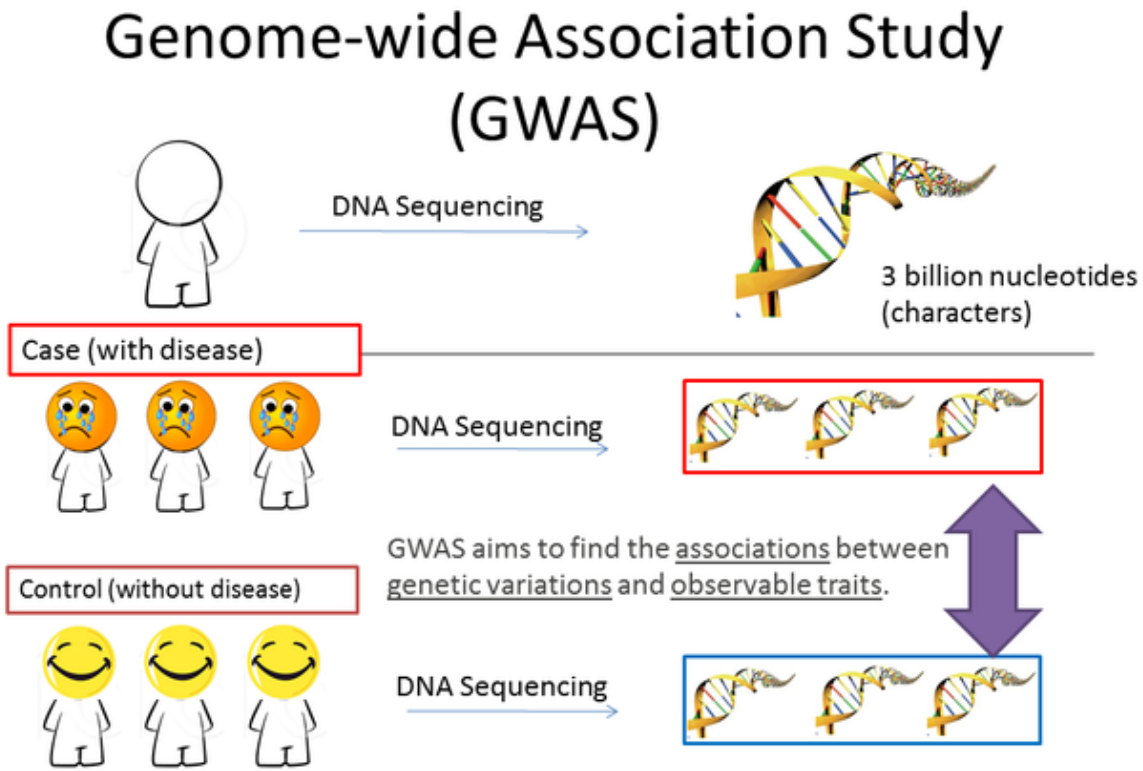
Single Nucleotide Polymorphism (SNP)

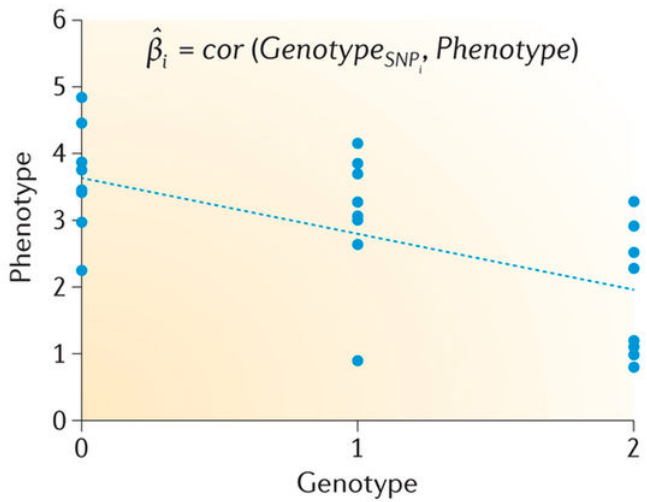
- Most common type of genetic variation
- Represent a difference in a single DNA building block (A-T, G-C)
- For example, a SNP T/C may replace T with C, resulting possible genotypes TT, TC, CC in the population
- The number of the minor nucleotide type (i.e., minor allele) in the population (0, 1, 2) will be used as the genotype data



tubascan.eu.

GWAS





From `Quora.com` and Pasaniuc B & Price AL, Nat. Rev. 2017

Standard GWAS Method

Consider the phenotype vector (Y) and genotype data vector (X_i) for the SNP i

- Logistic regression model $E[logit(Y)] = X_i\beta_i$ for case-control studies
- Linear regression model $Y = X_i\beta_i + \varepsilon_i$ for quantitative phenotypes
- Testing $H_0 : \beta_i = 0$
- Significance threshold **P-value** $\leq 5 \times 10^{-8}$, accounting for genome-wide multiple independent tests

Current GWAS Status

Limitations of Standard GWAS

- Identified significant SNPs are often located in non-coding DNA regions
- $\sim$1.2% of total DNA are known as coding regions
- Underlying biological mechanisms are often unknown

Classification	Approximate percentages[a]	Approximate numbers[a]
Intronic	40	1,047
Intergenic	32	838
Within non-coding sequence of a gene	10	262
Upstream	8	210
Downstream	4	105
Non-synonymous coding	3	79
3′ untranslated region	~1	26
Synonymous coding	~1	26
5′ untranslated region		
Regulatory region		
Nonsense-mediated decay transcript		
Unknown	~1	26
Splice site		
Gained stop codon		
Frameshift in a coding sequence		

GWAS Catalogue Signals as of December 2010. Freedman M.L. Nature Genetics, 2011.

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/`.

Standard GWAS of AMD

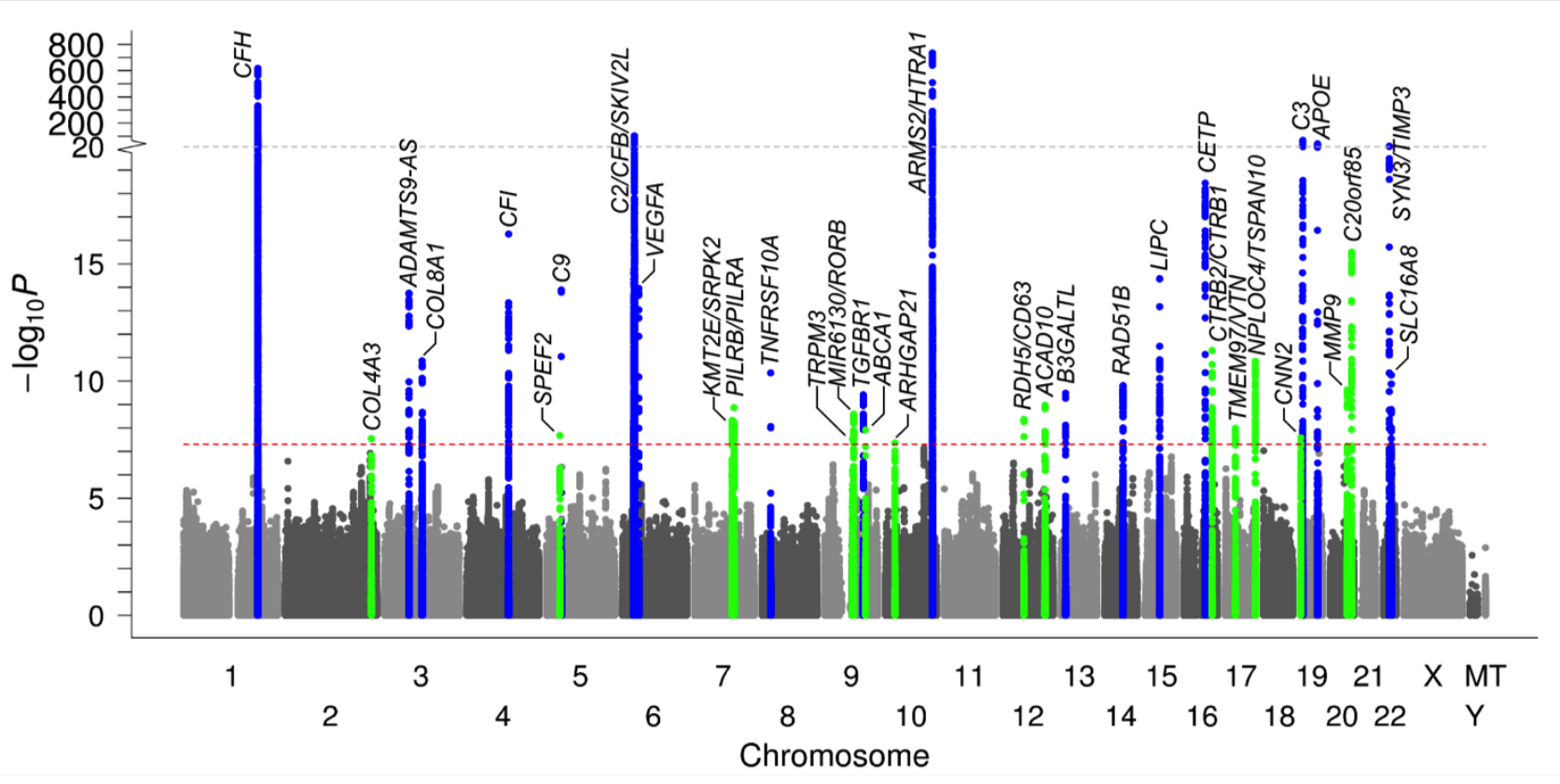
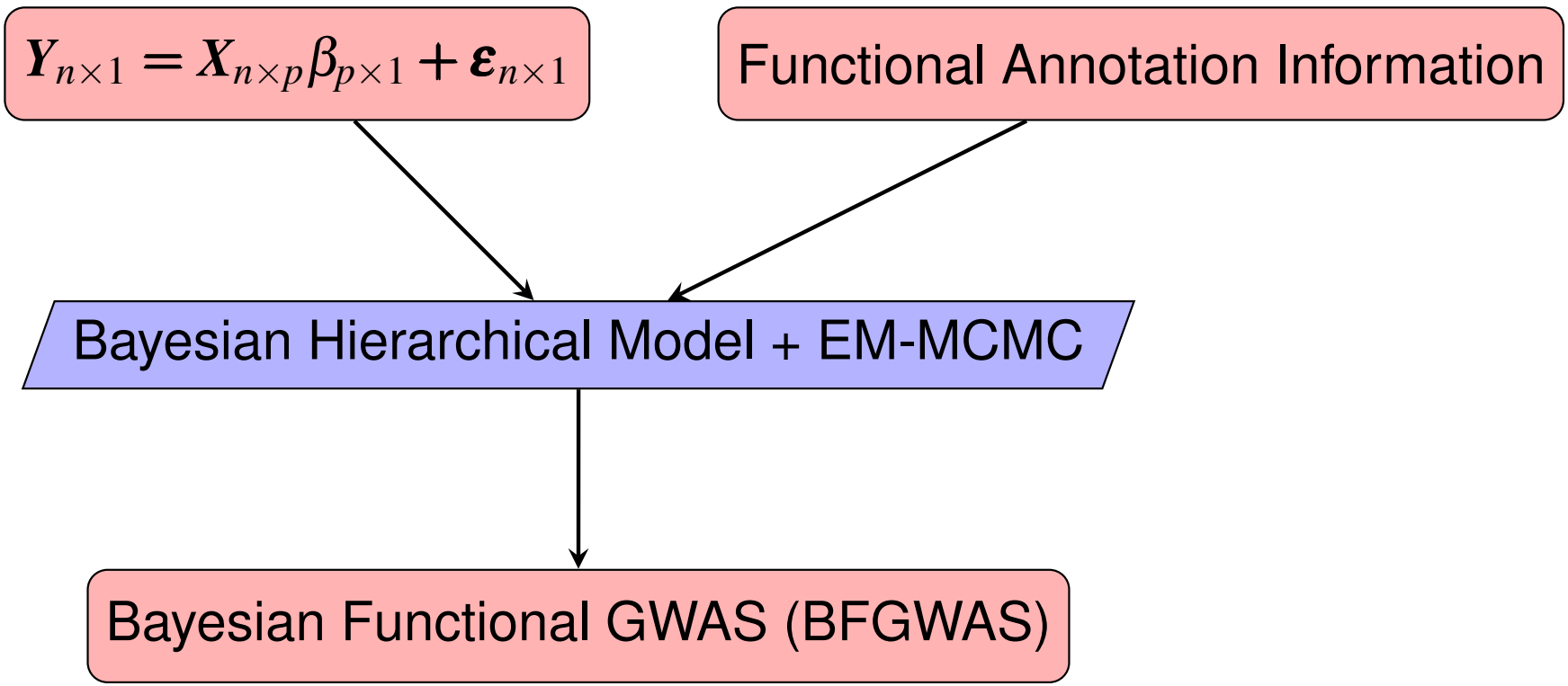


Figure 1: Majority of the associated variants are of unknown biological functions (Fritsche LG et al., 2016).

Motivations

- Understand biological mechanisms for genetic association studies
- Account for linkage disequilibrium (LD, nonrandom correlation among SNPs), for fine-mapping “causal” candidate signals
- Account for known functional annotations in GWAS to prioritize functional SNPs
- Derive scalable computation algorithm for genome-wide genotype data

Method Diagram



Bayesian Hierarchical Model

Joint linear regression model

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \varepsilon \sim MN(0, \tau^{-1}I). \tag{1}$$

Prior:

- $\beta_{i_q} \sim \pi_q N(0, \tau^{-1}\sigma_q^2) + (1-\pi_q)\delta_0$, for variants of annotation q
- Introduce a latent indicator vector $\boldsymbol{\gamma}_{p\times 1}$, equivalently

$$\gamma_{i_q} \sim Bernoulli(\pi_q),\ \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \sim \delta_0(\cdot),\ \boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim MVN_{|\boldsymbol{\gamma}|}(0, \tau^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}})$$

Parameters of Interest

- Category-specific (Enrichment parameters):
  - $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$: Causal probability per annotation
  - $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_Q^2)$: Effect-size variance for associated variants per annotation
- SNP-specific (Association evidence):
  - β_i: Genetic effect-size
  - $E[\gamma_i]$: Bayesian posterior inclusion probability (Bayesian PP), i.e., probability of being an associated SNP
- Region-level (Association evidence):
  - Regional-PP: Regional posterior inclusion probability, i.e., probability of being a risk locus

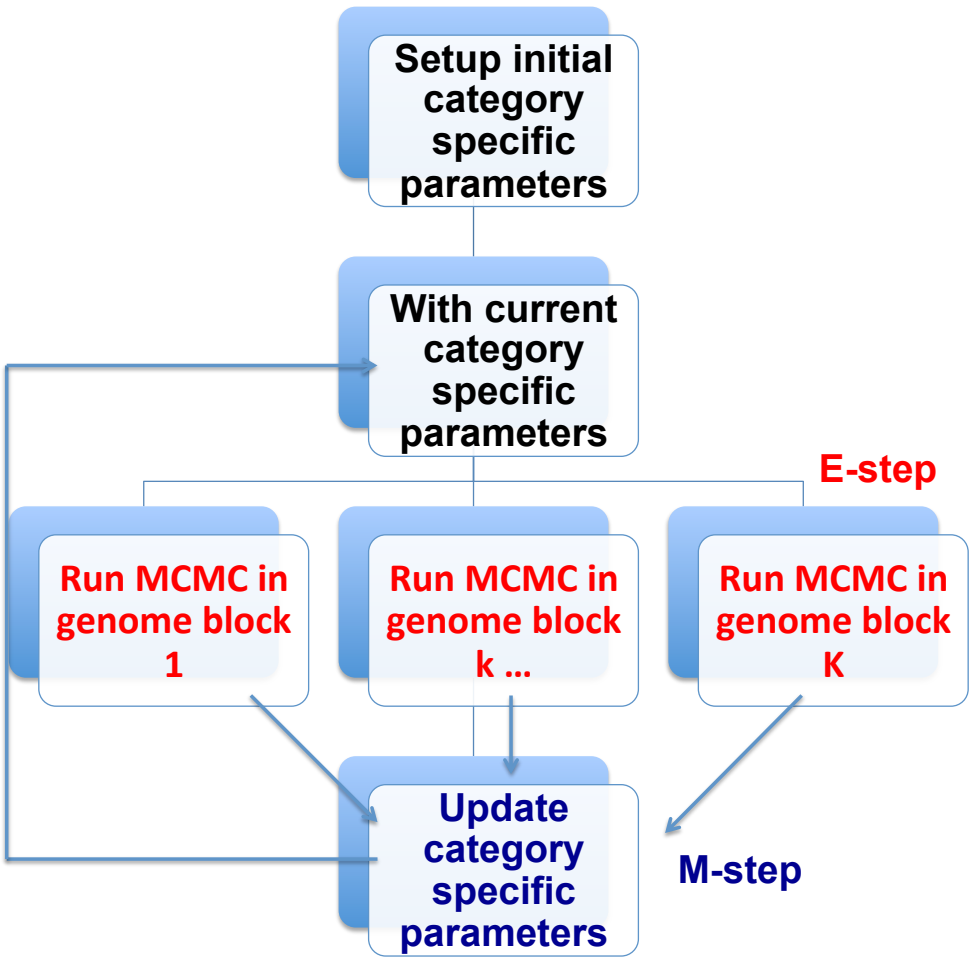
Bayesian Hierarchical Model

- Hierarchical priors
  - $\pi_q \sim Beta(a_q, b_q)$;
  - $\sigma_q^2 \sim InverseGamma(k_1, k_2)$;
  - $\tau \sim Gamma(k_3, k_4)$
- The joint posterior distribution

$$P(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \tau | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{A}) \propto \tag{2}$$
$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau) P(\boldsymbol{\beta}|\boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, \tau) P(\boldsymbol{\gamma}|\boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\sigma}^2) P(\tau),$$

- Product of Likelihood and Priors
- Challenges of Standard MCMC: memory usage and convergence rate

EM-MCMC Algorithm



Enabled genome-wide analysis

Improved MCMC convergence rate

MCMC Algorithm

Given category-specific parameters (π_q, σ_q^2) and residual variance τ^{-1}:

- Propose a new indicator vector γ
- Calculate conditional posterior likelihood

$$P(\gamma|Y,X) \propto |\Omega|^{-1/2} \exp\left\{\frac{\tau}{2} \boldsymbol{Y}^{\boldsymbol{T}} \boldsymbol{X}_{|\gamma|} V_\gamma \Omega^{-1} \boldsymbol{X}_{|\gamma|}^{\boldsymbol{T}} \boldsymbol{Y}\right\}, \Omega = V_{|\gamma|} \boldsymbol{X}_{|\gamma|}^{\boldsymbol{T}} \boldsymbol{X}_{|\gamma|} + I$$

- Apply Metropolis-Hastings algorithm
- If accepted, update effect-size estimates:

$$\widehat{\beta}_{|\gamma|} = \left[X_{|\gamma|}^T X_{|\gamma|} + V_\gamma^{-1}\right]^{-1} X_{|\gamma|}^T Y$$

- Summary statistics $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}, \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y})$ can be used here to save computational cost

Summary Statistics from Standard GWAS and LD

Assume both phenotype vector Y and genotype vector X_i are centered:

- Under the single variant model $Y = X_i\beta_i + \varepsilon$

$$\widehat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y$$

- Any element of $\boldsymbol{X^T Y}$ can be approximated by $\widehat{\beta}_i (X_i^T X_i)$
- LD coefficient (i.e., correlation) between X_i and X_j:

$$r_{ij} = \frac{X_i^T X_j}{\sqrt{(X_i^T X_i)(X_j^T X_j)}}$$

- $[\boldsymbol{X^T X}]_{ij}$ can be approximated by $\widehat{r_{ij}} \left(\sqrt{(X_i^T X_i)(X_j^T X_j)} \right)$
- $X_i^T X_i \approx 2nf_i(1-f_i)$ with minor allele frequency (MAF) f_i

Using summary statistics saves up to 90% computation time for MCMC with comparable results

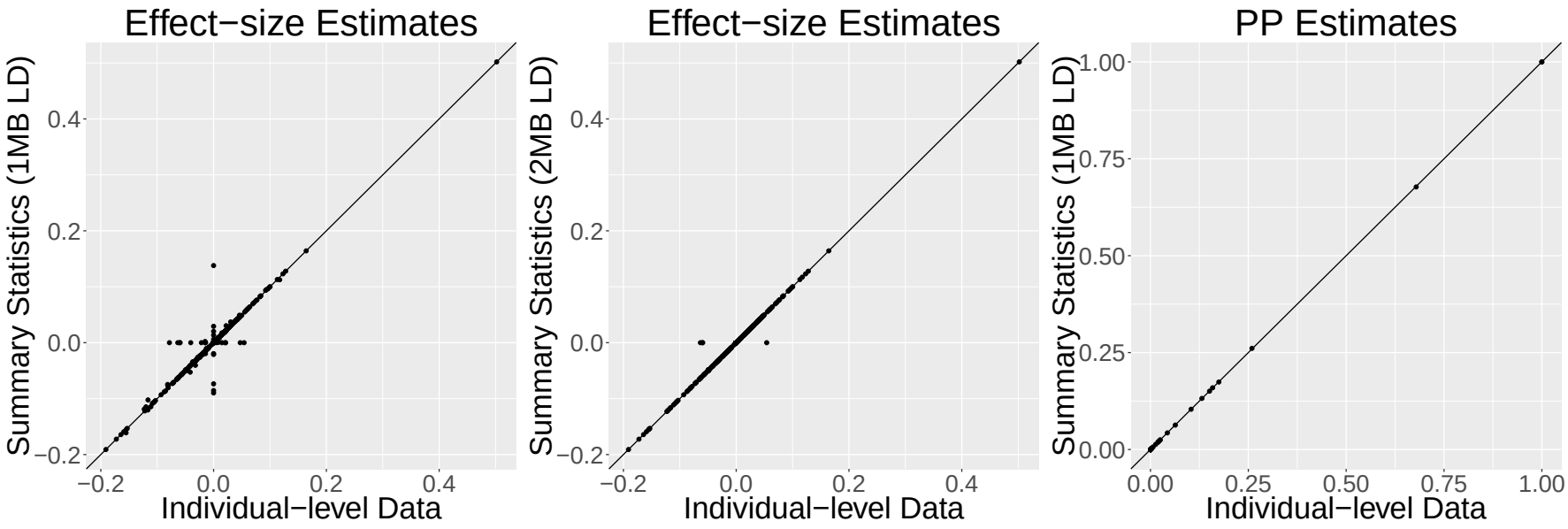


Figure 2: Using Summary Statistics vs. Individual-level Data.

EM Updates

MAPs (maximum a posteriori estimates):

Let $\widehat{\gamma_{jq}} = E[\gamma_{jq}]$

- Causal probability per annotation

$$\widehat{\pi_q} = \frac{\sum_{j_q=1}^{m_q} \widehat{\gamma_{j_q}} + a_q - 1}{m_q + a_q + b_q - 2}$$

- Effect-size variance per annotation

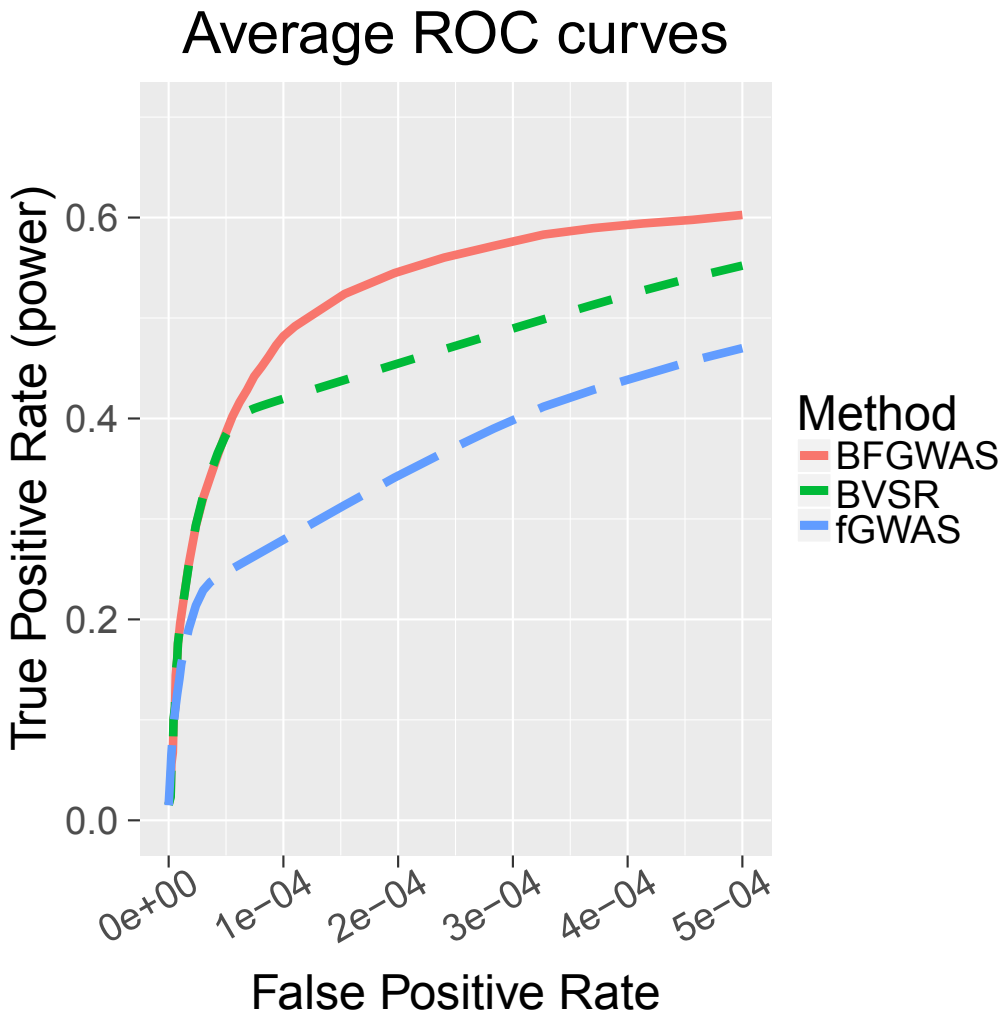
$$\widehat{\sigma_q^2} = \frac{\tau \sum_{j_q=1}^{m_q} (\widehat{\gamma_{j_q}} \widehat{\beta_{j_q}^2}) + 2k_2}{\sum_{j_q=1}^{m_q} \widehat{\gamma_{j_q}} + 2(k_1 + 1)}$$

Simulation Setup

- 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

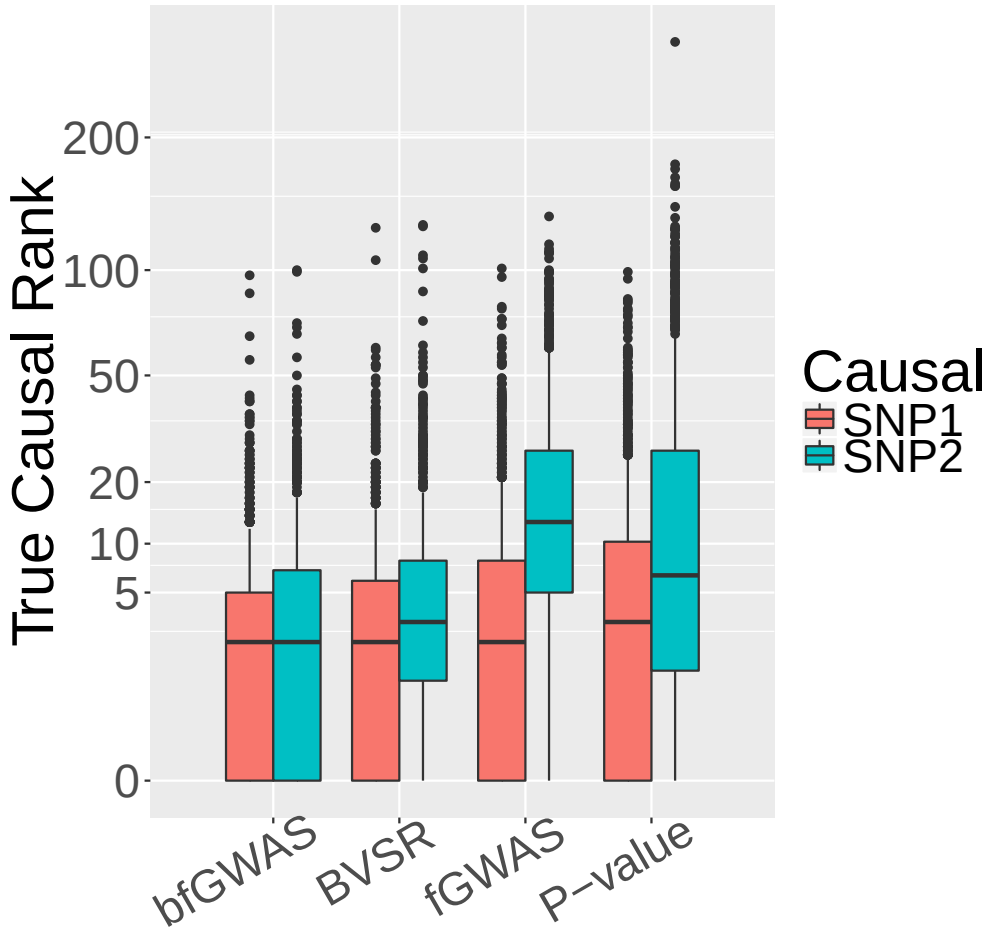


Highest Power to Discover Multiple Causals

SNP1: True causal with more significant P-value

SNP2: Second true causal

Higher ranks (smaller values) suggest higher power



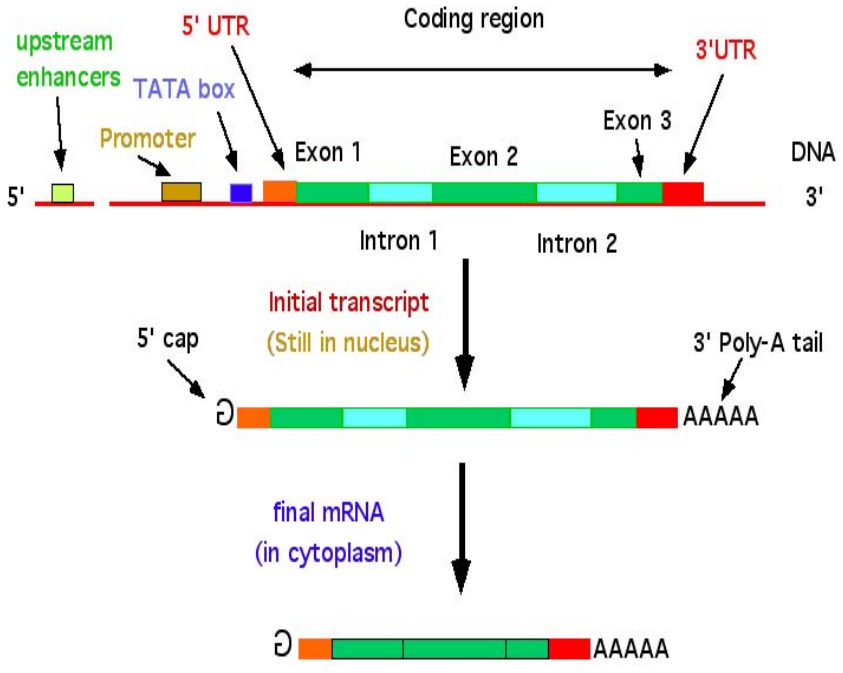
International AMD Genomics Consortium Data

- $\sim$10M low-frequency and common variants (MAF>0.5%)
- $\sim$ 16K cases and $\sim$18K controls (unrelated European)
- Phenotypes adjusted for age, gender, DNA source, and first 2 principal components
- GWAS results with gene-based annotations

Gene-based Annotations

Annotated by SeattleSeq:

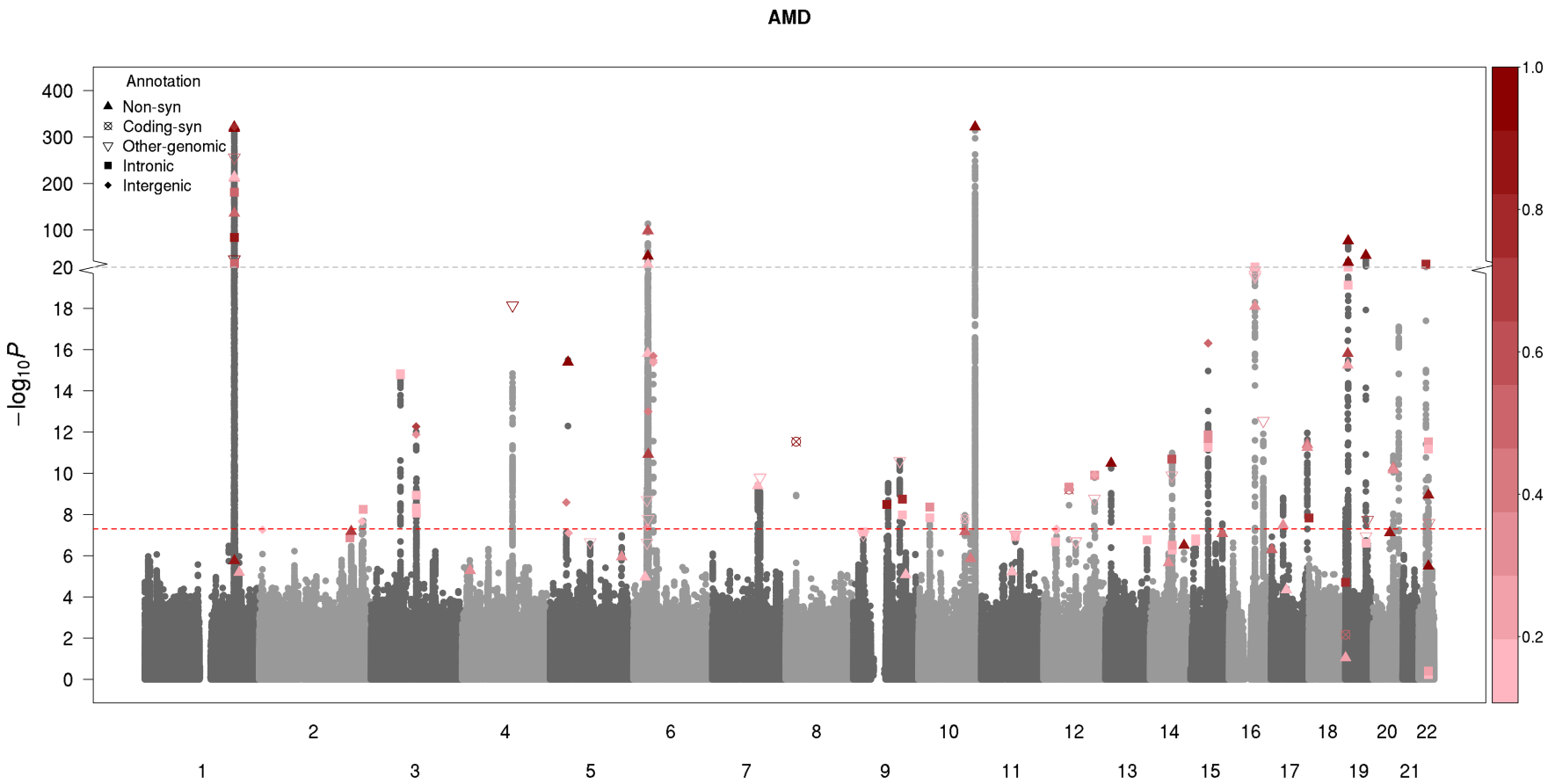
- Non-synonymous (42,005)
- Synonymous (67,165)
- Intronic (3,679,235)
- Intergenic (5,512,423)
- Other genomic (565,916, UTR, non-coding exons, upstream and downstream)



http://nitro.biosci.arizona.edu/

BFGWAS Results with Gene-based Annotations

Colored variants with Bayesian PPs > 0.1068 ($\sim$p-value < 5×10^{-8}).



BFGWAS Results with Gene-based Annotations

By Bayesian PP >0.1068, our method identified 150 variants with association evidence

	Non-syn	Coding-syn	Intronic	Intergenic	Other-genomic
Associations	47	4	54	18	27
Enrichment	72x	4x	0.9x	0.2x	3x

By Regional-PP > 0.95, our method identified 5 potentially novel loci, in addition to 32 known loci (Fritsche LG et al., 2016)

5 Potentially Novel Loci

Annotation	SNP/Gene	Previous Associations
Missense	*rs7562391/PPIL3*	
Missense	*rs61751507/CPN1*	Age-related Hearing Impairment (Fransen E et al., 2015)
Missense	*rs2232613/LBP*	Encodes Lipid Transfer Protein (Masson D et al., 2009)
Downstream	*rs114348558/ZNRD1-AS1*	Lipid Metabolisms (Kettunen J et al., 2012)
Splice	*rs6496562/ABHD2*	Coronary Artery Disease (Nikpay M et al., 2015)

- Known AMD risk loci *CETP, APOE, and LIPC* are also associated with Lipid Metabolisms and Coronary Artery Disease (Kettunen J et al., 2012, Nikpay M et al., 2015)
- Known AMD risk loci *CETP* is part of the Lipid Transfer Protein family (Masson D et al., 2009)

LocusZoom plots around the **Non-synonymous** SNP *rs4151667* (purple triangle).

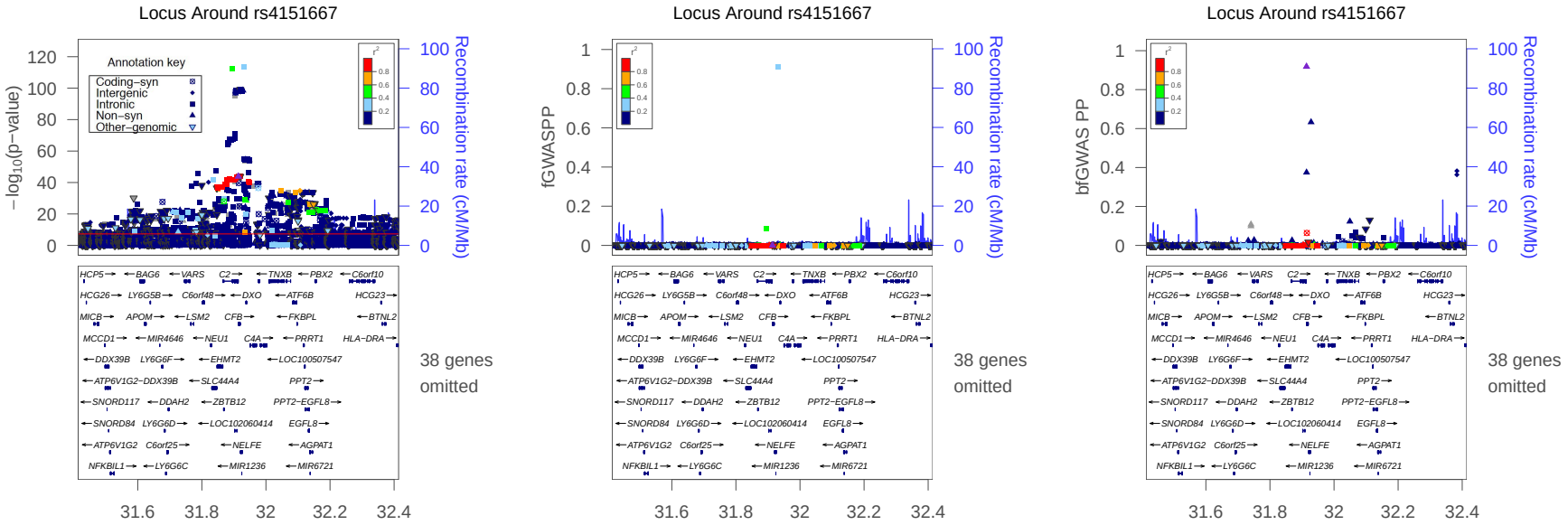


Figure 3: GWAS (left) vs. FGWAS (middle; Pickrell JK, AJHG 2014) vs. BFGWAS (right) for example locus #8.

Model Comparison

- Model1: top 2 SNPs (Intronic) by sequential forward selection
- Model2: top 2 SNPs (Non-synonymous) by BFGWAS

	Model1	Model2	Difference
AIC	95,857.36	**95,752.63**	104.73
BIC	95,891.1	**95,786.36**	104.74
−Log-likelihood	47,924.68	**47,872.31**	52.37

Haplotype Analysis

Haplotype with lead SNP *rs116503776* from standard GWAS and top 2 SNPs *rs4151667, rs115270436* by BFGWAS

rs116503776 *SKIV2L*	*rs4151667* *CFB*	*rs115270436* *SKIV2L*	Freq	OddsRatio	P-value
A	A	G	0.3%	0.364	8.9×10^{-11}
A	T	G	6.6%	0.522	1.5×10^{-86}
A	A	A	3.2%	0.561	5.0×10^{-36}
A	T	A	1.7%	**1.102**	$\mathbf{9.2 \times 10^{-2}}$
G	T	A	**87.8%**	-	Reference

Haplotype analysis by Fritsche LG et al. (2016) also found *rs116503776/SKIV2L* tags two previously identified **Non-synonymous** SNPs *rs4151667/CFB, rs641153/CFB*.

Example Locus *C3*

LocusZoom plots around the known **Non-synonymous** SNP *rs147859257* (purple triangle).

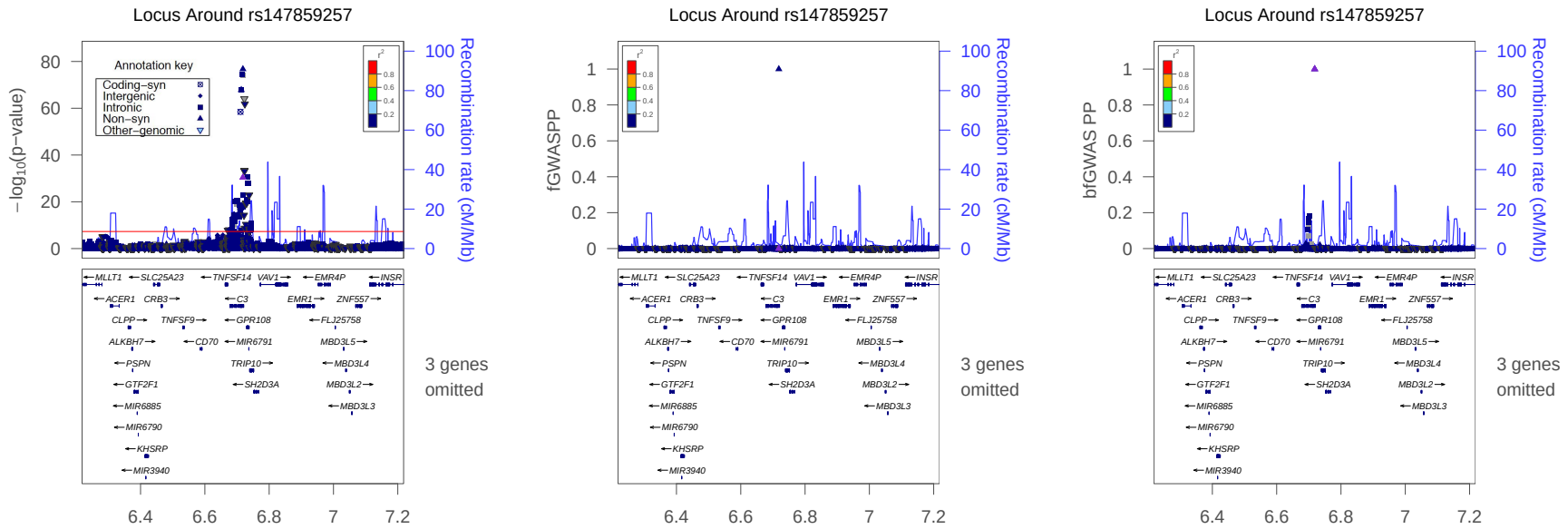


Figure 4: GWAS (left) vs. FGWAS (middle; Pickrell JK, AJHG 2014) vs. BFGWAS (right).

Enrichment Results

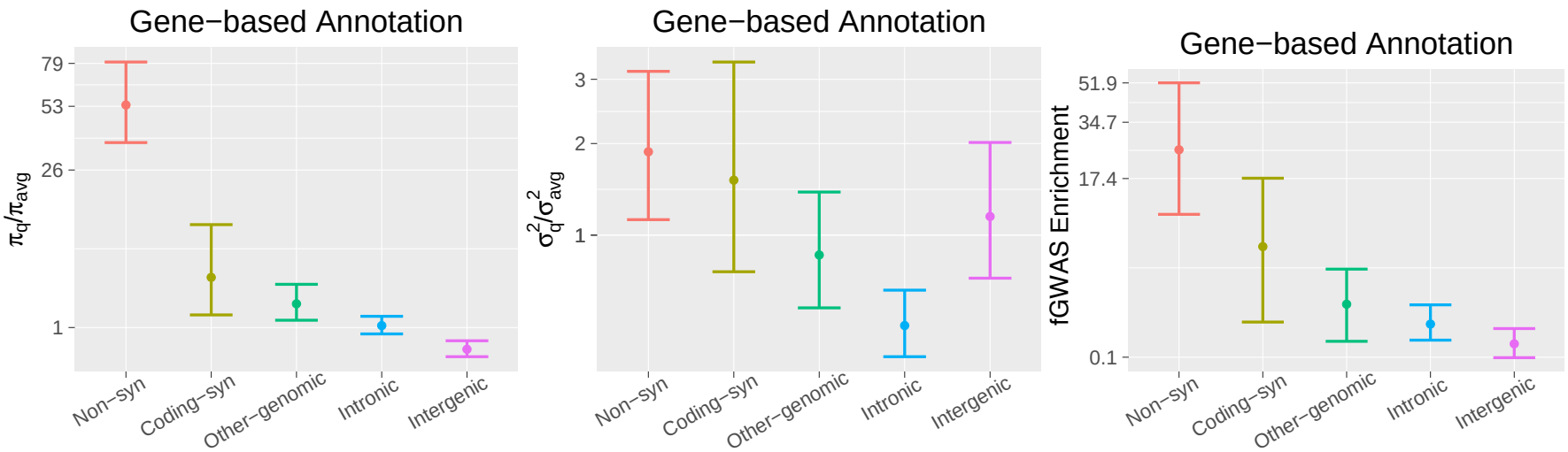


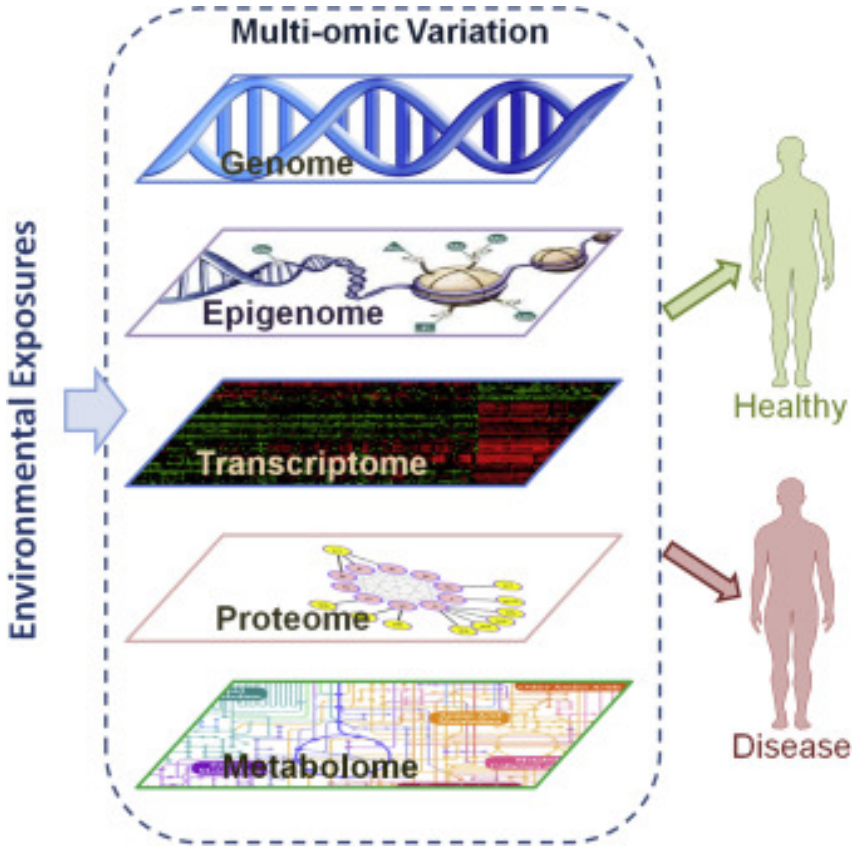
Figure 5: BFGWAS enrichment Results (left, middle) vs. FGWAS (right).

Summary

- **BFGWAS** integrates functional annotations in GWAS while accounting for LD
- Computationally efficient due to the scalable EM-MCMC algorithm and using summary statistics: $(\widehat{\beta}_i, \widehat{r_{ij}}, f_i)$
- Provides a list of risk loci and fine-mapped association candidates, as well as enrichment results
- Software **BFGWAS** is freely available at `https://github.com/yjingj/bfGWAS_SS`
- Method paper is available at `http://www.cell.com/ajhg/abstract/S0002-9297(17)30324-5`

Ongoing Research Topics

- Extend BFGWAS for multiple functional annotations
- Integrate gene expression (transcriptomic) data in GWAS
- Study longitudinal and image type “quantitative” phenotypes



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

Acknowledgments

- University of Michigan
  - Gonçalo Abecasis
  - Lars Fritsche
  - Xiang Zhou
- International AMD Genomics Consortium, `http://eaglep.case.edu/iamdgc_web/`

EMORY
UNIVERSITY
SCHOOL OF
MEDICINE