Genome-wide Association Studies

BIOS 770

02/10/2022

Jingjing Yang (jingjing.yang@emory.edu)

Outline

- Linear Mixed Model (LMM)
- Heritability Estimation by REML
- Fine-map GWAS Results
  - Conditional Analysis
  - Bayesian Method (FINEMAP)
- Multivariate GWAS

How to Address Population Stratification?

- Meta-analysis
- Account for top genotype Principal Components in GWAS
- Adjust false positives by Genomic Control Factor
- Check GWAS results by QQ plot
- Linear Mixed Model

Linear Mixed Model (LMM)

- Accounts for population stratification and relatedness
- Consider the following standard linear mixed model:

$$y_{n\times 1} = W\alpha + x\beta + Z_{n\times m}u_{m\times 1} + \epsilon$$
$$u_{m\times 1} \sim MVN_m(0, \lambda\tau^{-1}K)$$
$$\epsilon \sim MVN(0, \tau^{-1}I_n)$$

  - $y_{n\times 1}$ denotes the phenotype vector;
  - x denotes the genotype vector of the test SNP;
  - W denotes the confounding covariates: age, sex, top PCs, etc.;
  - $u_{m\times 1}$ denotes the random effect size vector with variance-covariance matrix $\lambda\tau^{-1}K$; taking $\mathrm{m} = n, Z = I_n$ for population based GWAS;
    - K is a known $m\times m$ relatedness matrix
    - I_n is an n$\times n$ identity matrix
  - ϵ denotes the error vector with variance-covariance matrix $\tau^{-1}I_n$.

Linear Mixed Model (LMM)

- Efficient statistical inference algorithm used by Genome-wide Efficient Mixed-Model Association (GEMMA) (X. Zhou & M. Stephens, Nature Genetics, 2012).
  - Obtain maximum-likelihood estimates (MLEs)
  - Obtain restricted/residual maximum-likelihood (REML) estimates
  - Calculate exact test statistics

Log-likelihood and Log-REstricted Likelihood Functions

$$l(\lambda,\tau,\boldsymbol{\alpha},\beta)=\frac{n}{2}\log(\tau)-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log|\mathbf{H}| -\frac{1}{2}\tau(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{x}\beta)^T\mathbf{H}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{x}\beta) \quad (1)$$

and

$$l_r(\lambda,\tau)=\frac{n-c-1}{2}\log(\tau)-\frac{n-c-1}{2}\log(2\pi)+\frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^T(\mathbf{W},\mathbf{x})\right| -\frac{1}{2}\log|\mathbf{H}|-\frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right|-\frac{1}{2}\tau\mathbf{y}^T\mathbf{P}_x\mathbf{y} \quad (2)$$

where $\mathbf{G}=\mathbf{ZKZ}^T$, $\mathbf{H}=\lambda\mathbf{G}+\mathbf{I}_n$ and $\mathbf{P}_X=\mathbf{H}^{-1}-\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})((\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x}))^{-1}(\mathbf{W},\mathbf{x})^T\mathrm{H}^{-1}$.

- MLE $\hat{\alpha},\hat{\beta}$, and REML $\hat{\tau}$ can be easily obtained if λ is known.
- MLE of $\hat{\alpha},\hat{\beta}$ do not depend on $\hat{\tau}$.
- REML $\hat{\tau}$ is an unbiased estimator for residual variance.

If λ is Known

If λ is known, the log-likelihood is maximized at:

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\beta} \end{pmatrix} = ((\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x}))^{-1}(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}\mathbf{y},$$

$$\hat{\tau} = \frac{n}{(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\alpha}}-\mathbf{x}\hat{\beta})^T\mathbf{H}^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\alpha}}-\mathbf{x}\hat{\beta})} = \frac{n}{\mathbf{y}^T\mathbf{P}_x\mathbf{y}}.$$

The last equation uses the property $\mathbf{P}_x\mathbf{H}\mathbf{P}_x = \mathbf{P}_x$. This can be derived by noticing $\mathbf{P}_x = \mathbf{M}_x(\mathbf{M}_x\mathbf{H}\mathbf{M}_x)^-\mathbf{M}_x$, where $\mathbf{M}_x = \mathbf{I}_n - (\mathbf{W},\mathbf{x})((\mathbf{W},\mathbf{x})^T(\mathbf{W},\mathbf{x}))^{-1}(\mathbf{W},\mathbf{x})^T$ and $-$ denotes generalized inverse.

Similarly, the log-restricted likelihood is maximized at

$$\hat{\tau} = \frac{n-c-1}{\mathbf{y}^T\mathbf{P}_x\mathbf{y}}.$$

Therefore, finding MLE and REML estimates is equivalent to optimizing the following functions with respect to λ:

$$l(\lambda) = \frac{n}{2}\log(\frac{n}{2\pi}) - \frac{n}{2} - \frac{1}{2}\log|\mathbf{H}| - \frac{n}{2}\log(\mathbf{y}^T\mathbf{P}_x\mathbf{y}),$$

$$\begin{aligned} l_r(\lambda) = {} & \frac{n-c-1}{2}\log(\frac{n-c-1}{2\pi}) - \frac{n-c-1}{2} + \frac{1}{2}\log|(\mathbf{W},\mathbf{x})^T(\mathbf{W},\mathbf{x})| \\ & - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log|(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})| - \frac{n-c-1}{2}\log(\mathbf{y}^T\mathbf{P}_x\mathbf{y}). \end{aligned}$$

Optimizing log-likelihood and log-REstricted likelihood functions with respect to λ

$$l(\lambda) = \frac{n}{2}\log\left(\frac{n}{2\pi}\right) - \frac{n}{2} - \frac{1}{2}\log|\mathbf{H}| - \frac{n}{2}\log\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right) \tag{3}$$

$$\begin{aligned} l_r(\lambda) = {} & \frac{n-c-1}{2}\log\left(\frac{n-c-1}{2\pi}\right) - \frac{n-c-1}{2} + \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^T(\mathbf{W},\mathbf{x})\right| \\ & -\frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log\left|(\mathbf{W},\mathbf{x})^T\mathbf{H}^{-1}(\mathbf{W},\mathbf{x})\right| - \frac{n-c-1}{2}\log\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right) \end{aligned} \tag{4}$$

$$\frac{\partial l(\lambda)}{\partial\lambda} = -\frac{1}{2}\mathrm{trace}\left(\mathbf{H}^{-1}\mathbf{G}\right) + \frac{n}{2}\frac{\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}}{\mathbf{y}^T\mathbf{P}_x\mathbf{y}} \tag{5}$$

$$\frac{\partial^2 l(\lambda)}{\partial\lambda^2} = \frac{1}{2}\mathrm{trace}\left(\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\mathbf{G}\right) - \frac{n}{2}\frac{2\left(\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}\right)\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right) - \left(\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}\right)^2}{\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right)^2} \tag{6}$$

Optimizing log-likelihood and log-REstricted likelihood functions with respect to λ

$$\frac{\partial l_r(\lambda)}{\partial \lambda} = -\frac{1}{2}\text{trace}\left(\mathbf{P}_x\mathbf{G}\right) + \frac{n-c-1}{2}\frac{\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}}{\mathbf{y}^T\mathbf{P}_x\mathbf{y}} \tag{7}$$

$$\frac{\partial^2 l_r(\lambda)}{\partial \lambda^2} = \frac{1}{2}\text{trace}\left(\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\right) - \frac{n-c-1}{2}\frac{2\left(\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}\right)\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right) - \left(\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y}\right)^2}{\left(\mathbf{y}^T\mathbf{P}_x\mathbf{y}\right)^2} \tag{8}$$

Efficient computation matters

- Use Brent's method to provide an initial value
- Estimate λ by Newton-Raphson's method
- Simplify trace terms and vector-matrix-vector product terms
- Use the recursion properties of the trace terms and vector-matrix-vector product terms

Table 1 Performance of different methods for GWAS with the linear mixed model

Methods		Time complexity[a]	Computing time	
			HDL-C[b]	Crohn's disease[c]
Exact methods	GEMMA	$O(mn^2 + cn^2 + pn^2 + pt_2c^2n)$	33 min	3.3 h
	EMMA	$O(mn^2 + pmn^2 + pt_2n)$	~9 d	~27 years
	FaST-LMM[d]	$O(mn^2 + cn^2 + pn^2 + pt_1c^2n)$	6.8 h	6.2 h
Approximate methods	EMMAX	$O(mn^2 + t_2n + pn^2)$	44 min	6.4 h
	GRAMMAR	$O(mn^2 + t_2n + pn)$	1.6 min	12 min

All computing was performed on a single core of an Intel Xeon L5420 2.50 GHz CPU. The time for the EMMA method is projected from a selection of 10,000 and 100 genetic markers in the HMDP and WTCCC data sets, respectively. Note that EMMA was implemented in R, whereas others were implemented in C. A C implementation of EMMA could be a few times faster. p, the number of genetic markers; n, the number of individuals; m, the number of strains (equal to n for human studies); c, the number of covariates (fixed effects) in addition to the genotypes. t_1 and t_2 are the number of optimization iterations required for Brent's method (super-linear rate of convergence) and the Newton-Raphson method (quadratic rate of convergence), respectively. Note that t_2 is expected to be smaller than t_1.
[a]Complexities are given assuming the usual genome-wide relatedness matrix, which has rank n. In the current implementation of various methods except EMMA, the first terms are actually n^3, but it would in principle be straightforward to convert them to mn^2. [b]$m = 99$, $n = 681$, and $p = 1{,}885{,}197$. [c]$m = n = 4{,}686$, and $p = 442{,}001$. [d]These results are for the algorithm in FaST-LMM that uses the standard full-rank relatedness matrix, which produces P values that are identical to those generated in GEMMA and EMMA.

Test Statistics and P-values

To test the null hypothesis $\beta = 0$, we obtain the likelihood ratio test statistic with MLE estimates and the Wald test statistic with the REML estimate as suggested[2,1]:

$$D_{lrt} = 2 \log \frac{l_1(\hat{\lambda}_1)}{l_0(\hat{\lambda}_0)},$$

$$F_{Wald} = \frac{\hat{\beta}^2}{V(\hat{\beta})}.$$

where l_1 and l_0 are the likelihood functions for the null and the alternative models, respectively; $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are the MLE estimates for the null and the alternative models, respectively; $\hat{\beta} = (\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{x})^{-1}(\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{y})$ is the estimate for β obtained using the REML estimate $\hat{\lambda}_r$ in the alternative model; and $V(\hat{\beta}) = (n-c-1)^{-1}(\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{x})^{-1}(\mathbf{y}^T \mathbf{P}_x(\hat{\lambda}_r)\mathbf{y})$ is the variance for $\hat{\beta}$. Under the null hypothesis the likelihood ratio test statistic D_{lrt} and the Wald test statistics F_{Wald} come from a $\chi^2(1)$ and a $F(1, n-c-1)$ distribution respectively, and p values can be calculated accordingly.

SNP Heritability

- SNP heritability (i.e., narrow sense heritability): the proportion of total phenotype variation explained by additive genetic effects
  - Estimated using GWAS significant SNPs
  - Estimated using SNPs with GWAS p-values < 0.05
  - Estimated using genome-wide genotypes
- Missing heritability: Big gap between SNP heritability estimated based on the standard linear regression model and the broad sense heritability

Published: 20 June 2010

Common SNPs explain a large proportion of the heritability for human height

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard & Peter M Visscher

Nature Genetics **42**, 565–569 (2010) | Cite this article

REML for SNP Heritability

- Assume LMM under the infinitesimal genetic architecture (all SNPs contributed an equal small amount to the heritability) :

$$y_{n\times 1} = \mu + u_{n\times 1} + \epsilon$$
$$u_{n\times 1} \sim MVN_n\left(0, \sigma_g^2 K_{n\times n}\right)$$
$$\epsilon \sim MVN_n(0, \sigma_\epsilon^2 I_{n\times n})$$

- SNP heritability: $h^2 = \frac{\sigma_g^2}{Var(y)}$
- Unbiased REML estimate for σ_g^2 would give us the estimated heritability $\widehat{h^2} = \frac{\widehat{\sigma_g^2}}{var(y)}$

Table 1 Estimation of phenotypic variance explained from genetic relationships among unrelated individuals by restricted maximum likelihood (Jian Yang et. al. Nature Genetics, 2010).

		No. SNPs	$L(H_0)$[a]	$L(H_1)$[b]	LRT[c]	σ_g^2 (s.e.)	σ_e^2 (s.e.)	σ_p^2 (s.e.)	h^2[d] (s.e.)
295K SNPs	Raw	294,831	−1950.89	−1936.12	29.53	0.445 (0.084)	0.546 (0.082)	0.991 (0.023)	0.449 (0.083)
	Adj.[e]	294,831	−1950.89	−1936.12	29.53	0.532 (0.101)	0.458 (0.098)	0.991 (0.023)	0.537 (0.100)
295K/516K SNPs[f]	Raw	294,831/516,345	−1950.89	−1935.94	29.89	0.449 (0.085)	0.536 (0.083)	0.986 (0.022)	0.456 (0.085)
	Adj.	294,831/516,345	−1950.89	−1935.87	30.04	0.536 (0.101)	0.449 (0.099)	0.985 (0.022)	0.544 (0.101)

[a]log-likelihood under the null hypothesis that $\sigma_g^2=0$.

[b]log-likelihood under the alternative hypothesis that $\sigma_g^2 \neq 0$;

[c]log-likelihood ratio test statistic, $LRT = 2[L(H_1) - L(H_0)]$.

[d]Estimate of variance explained by all SNPs, with its s.e. given in the parentheses.

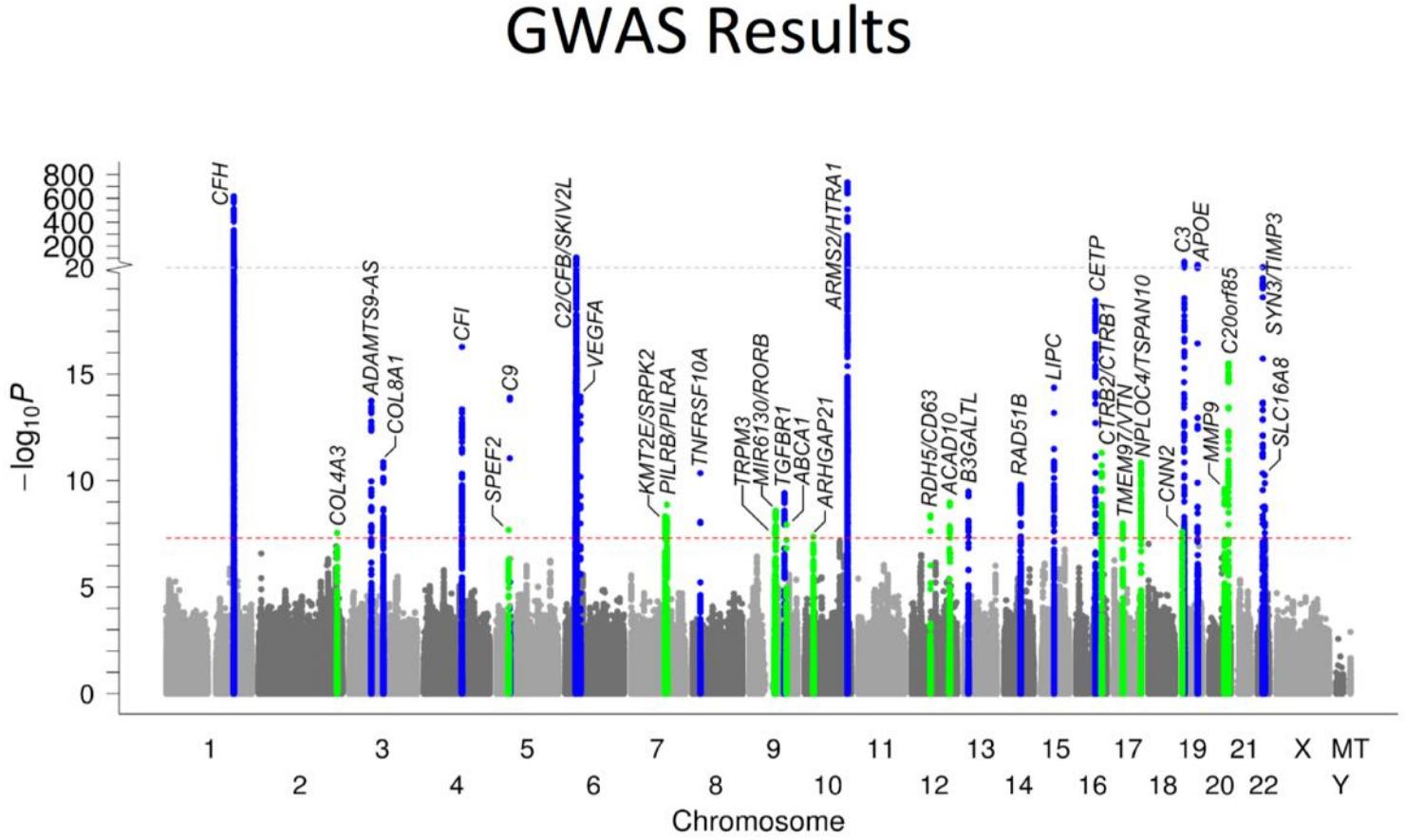
[e]Raw estimate of genetic relationship adjusted for prediction error with equation (9) (assuming $c = 0$).

[f]The genetic relationships are estimated from 1,318 individuals with 516,345 SNPs, and the other 2,607 individuals with 294,831 SNPs. See Online Methods for definitions of notations.

Missing Heritability

- Most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests.
- Where to “find” the missing heritability?
  - Incomplete linkage disequilibrium between causal variants and genotyped SNPs?
    - Genotype imputation
    - Whole genome sequencing
  - Modeling LD of genome-wide variants?
    - Multivariate linear regression model
  - Rare variants?

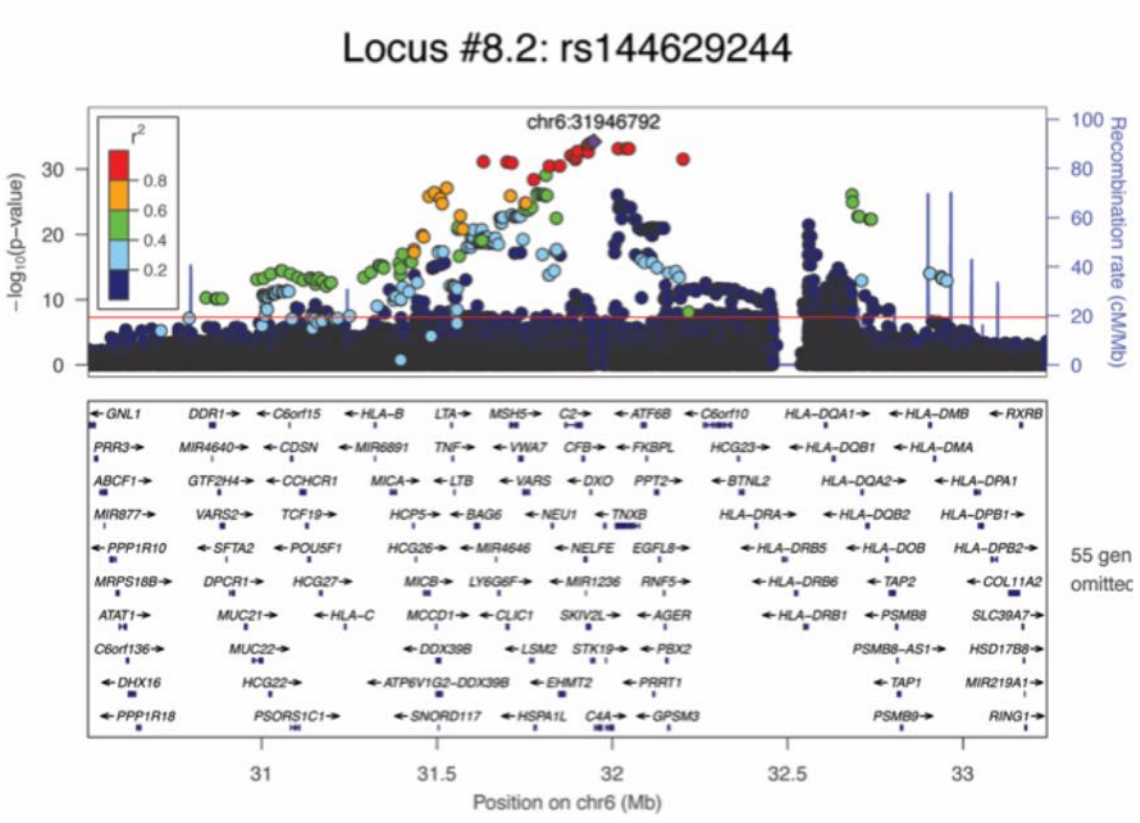
Fine-mapping GWAS Results



18 known AMD loci and **16 novel AMD loci**

Visualize GWAS Loci by Locus Zoom Plot

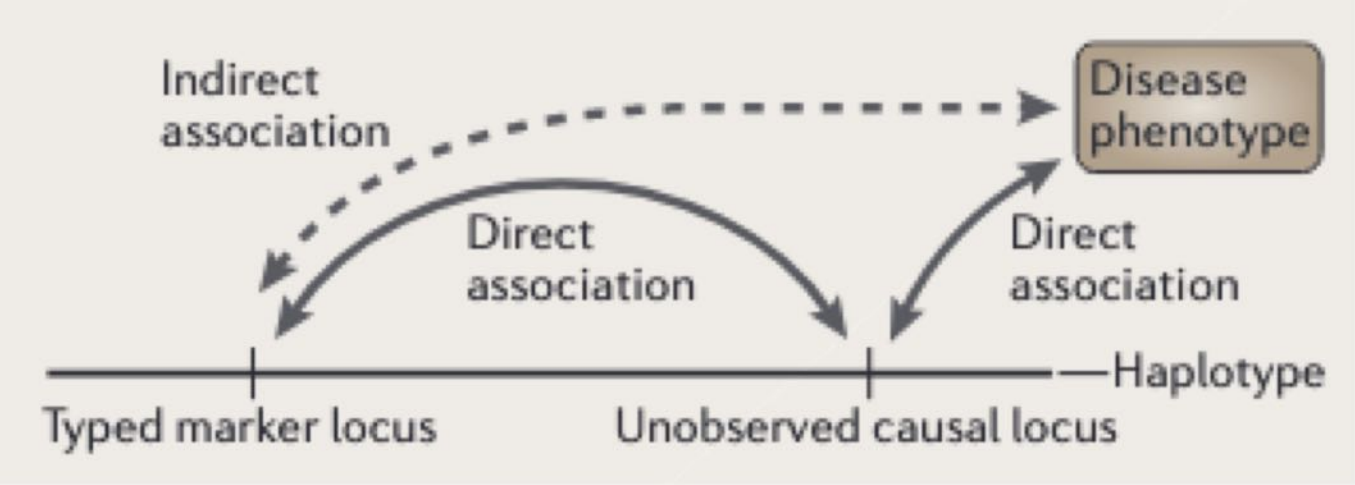
– Zoom into the peak region with gene annotations
– Visualize r^2 between the specified significant (purple diamond) signal and its neighbor SNPs
– Visualize recombination rate



Fritsche L.G. et al. Nat Genet, 2016.

Why LD is Important for Association Studies?

- Hypothesis: SNPs in strong LD with disease variant are good proxies for disease variant



Balding, 2006

- If testing (unobservable) disease variant for association would yield chi-squared statistic X^2, testing variant in LD yields r^2X^2 (useful for meta-analysis)

Fine-mapping GWAS Results

- Hypothesis: Only a small amount of genetic variants (dozens or hundreds vs. millions) would be true causal variants
- Most significant GWAS signals, i.e., significant SNPs, are located in non-coding regions
- All SNPs in LD (i.e., highly correlated) with the nearby most significant GWAS signal are likely to be tested with significant p-values
- Fine-map GWAS results: pinpointing potential true causal SNPs (true biological molecular mechanisms) from all SNPs that are in LD

Fine-map GWAS Results

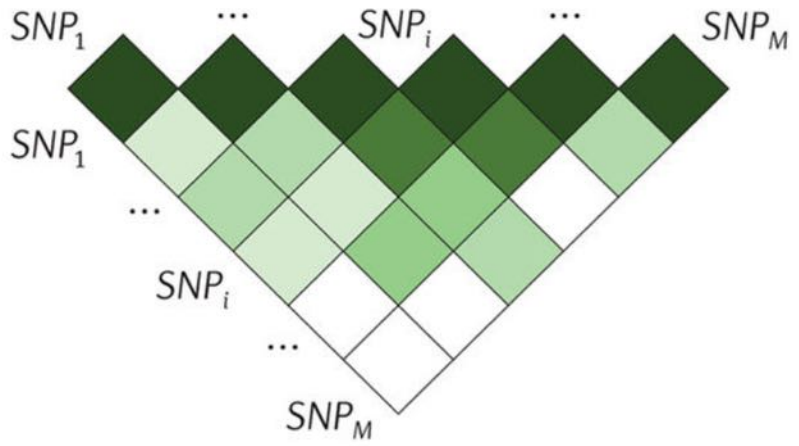
- Conditional analysis
- Fine-mapping using GWAS summary statistics and accounting for LD
- Conducted per risk locus with significant GWAS signals (region, e.g., +-5KB)

Conditional Analysis

Sequential Forward Selection

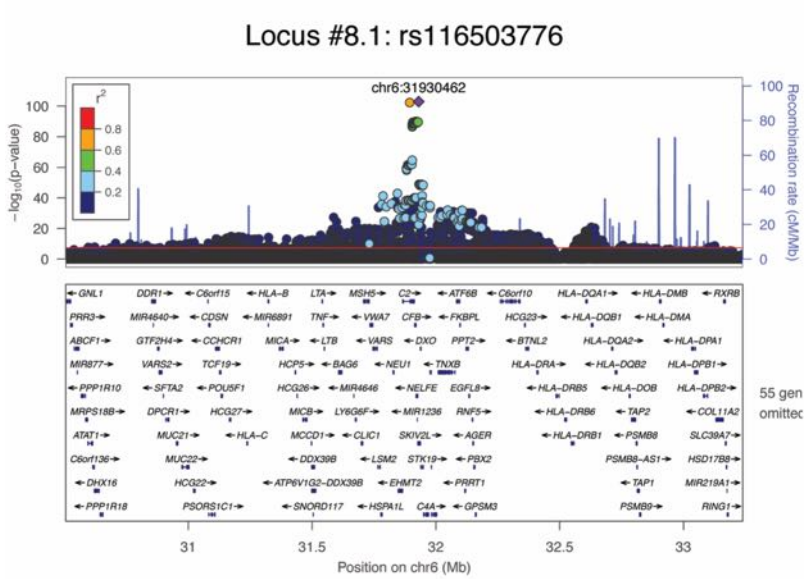
Aim: Within each region of interest, identify all statistically independent variants

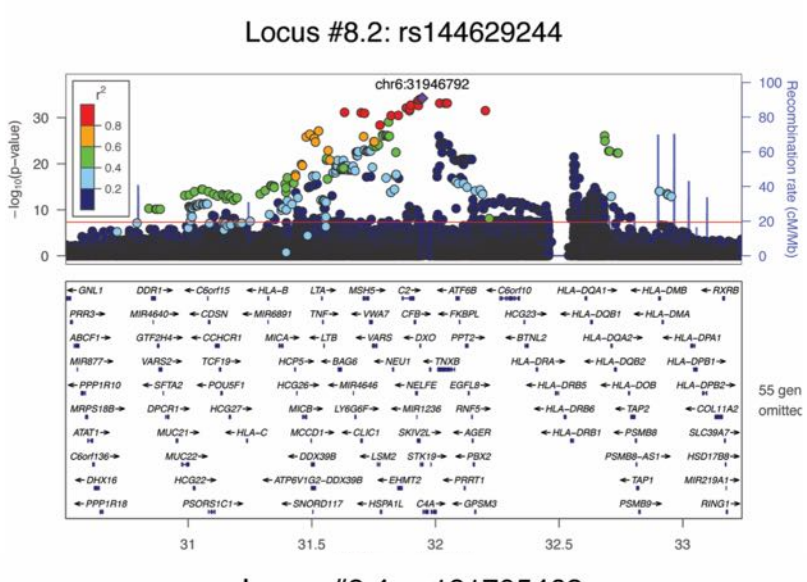
1. Select variant with smallest P value ($P < 5x10^{-8}$), write into results file
2. Conduct region-wide association analysis conditioning on variants in results file
3. From the results of 2., if smallest $P < 5x10^{-8}$, select variant write into results file; otherwise stop
4. Repeat 2. and 3.

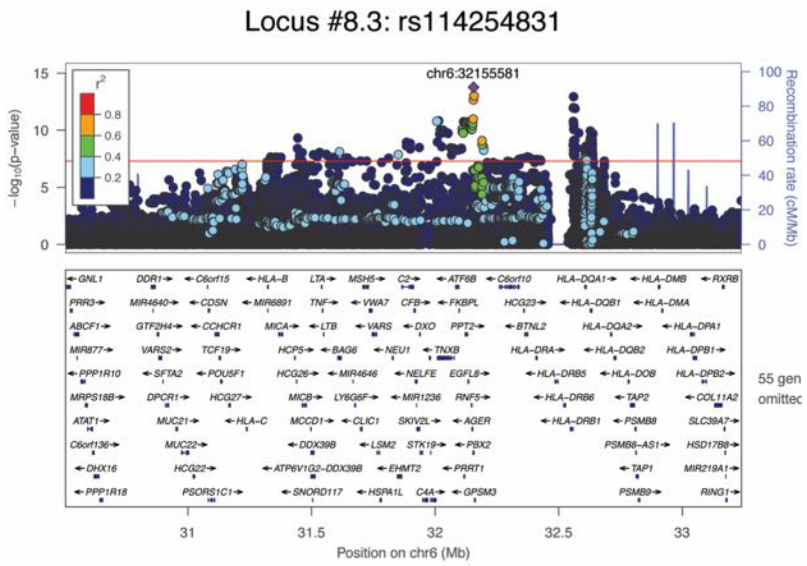


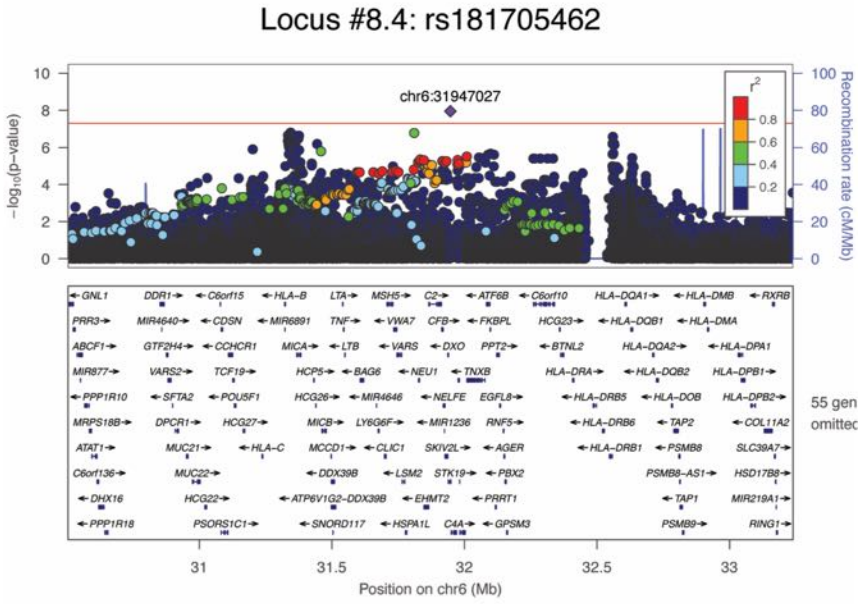
Fritsche L and Pasaniuc B and Price AL, Nat. Rev. 2017

Conditional Analysis









Conditional Analysis

- Informative about the number of complementary sources of association signals within the region
- Fails to provide probabilistic measures of causality for individual variants
- Not accounting for functional annotations (i.e., biological functions) of SNPs

Bayesian Method for Fine-mapping

- Existing methods/tools using the same Bayesian framework:
  - PAINTOR (Kichaev et al., 2014, Kichaev and Pasaniuc, 2015)
  - CAVIAR (Hormozdiari et al., 2014)
  - CAVIARBF (Chen et al., 2015)
  - FINEMAP (Benner et al., 2016)
- Requires only GWAS summary statistics and reference LD
- Provide probabilistic measures of causality for individual variants

Bayesian Method for Fine-mapping

- Likelihood based on Multivariate Linear Regression Model

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

- MLE estimates of β depends on column-standardized X, y only through SNP correlation (LD) matrix R and single-SNP Z-score test statistic $\hat{z}$:

$$\hat{\beta} = (X^T X)^{-1} X^T y = n^{-\frac{1}{2}} \sigma R^{-1} \hat{z}$$

$$R = n^{-1} X^T X, \qquad \hat{z} = \frac{X^T y}{\sqrt{n}\sigma}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = n^{-1} \sigma R^{-1}$$

$$E[\hat{\beta}] = \beta$$

Bayesian Method for Fine-mapping

- Single-SNP Z-score test statistic $\hat{z}$ can be obtained from GWAS summary statistics
- SNP correlation matrix R can be approximated from a reference panel with the same ethnicity
- The likelihood function for β can be approximated by

$$\hat{\beta} \sim MVN(\beta, Var(\hat{\beta}))$$

- Use a Bayesian approach with a prior distribution to account for sparsity among causal effects

Priors for β with a binary indicator vector γ

- Assume an indicator vector γ : $\gamma_l = 1$ if the lth variant has non-zero causal effect $\beta_l \neq 0$; $\gamma_l = 0$ if $\beta_l \neq 0$.
- For non-zero effect sizes, the likelihood is given by

$$\beta|\gamma \sim MVN(0, s_\beta^2 \sigma^2 \Delta_\gamma)$$

Δ_r: Diagonal matrix with γ on the diagonal

- σ^2 can be taken as 1 for quantitative traits or $1/(\varphi(1-\varphi))$ with φ denoting the proportion of cases among n individuals
  - Assuming standardized phenotype vector
  - Assuming no other confounding covariates
- Taking $s_\beta^2 = 0.05^2$ means with 95% probability a causal SNP explains less than 1% of the phenotype variation (FINEMAP)

Prior of binary indicator vector γ with respect to the number of assumed true causal SNPs

- $p_k = \Pr(\#\ of\ k\ causal\ SNPs)\,, k = 1, \ldots, K;\ \ K \ll m$ total number SNPs
- $p_0 = 0$, assuming there is at least one causal SNP for the fine-mapped region
- Assume the same probability for each configuration with k causal SNPs (FINEMAP)

$$p(\gamma) = p_k / \binom{m}{k},\ \sum_{l=1}^{m} \gamma_l = k$$

Likelihood function of indicator vector γ by integrating out β

- Posterior distribution of the indicator vector γ infers the posterior causal probability per SNP : $\mathrm{P}(\gamma|y,X)$
- Likelihood function of indicator vector γ by integrating out β:

$$\mathrm{L}(\gamma) = \mathrm{P}(y|\gamma,X) = \int P(y\,|\beta,X)P(\beta|\gamma)d\beta$$
$$= N\left(\hat{\beta}\middle|0,\sigma^2(nR)^{-1} + s_\beta^2\sigma^2\Delta_\gamma\right)$$
$$= N\left(\hat{z}\middle|0,R + R\Sigma_\gamma R\right), \Sigma_\gamma = ns_\beta^2\Delta_\gamma$$

Δ_r: Diagonal matrix with γ on the diagonal

- The likelihood function $L(\gamma)$ need to be evaluated per γ
- Computational efficiency is needed because of all $\sum_{k-1}^{K}\binom{m}{k}$ causal configurations

Evaluate likelihood function $L(\gamma)$ by FINEMAP

- Partition $\hat{z}$ into components for the Causal SNPs $\hat{z}_C$ and Non-causal SNPs $\hat{z}_N$
- Partition R, Σ_γ, and $R\Sigma_\gamma R$

$$R = \left[\begin{array}{c|c} R_{CC} & R_{CN} \\ \hline R_{NC} & R_{NN} \end{array}\right] \qquad \Sigma_\gamma = \begin{bmatrix} \Sigma_{CC} & 0 \\ 0 & 0 \end{bmatrix}$$

$$R + R\Sigma_\gamma R = \left[\begin{array}{c|c} R_{CC} + R_{CC}\Sigma_{CC}R_{CC} & R_{CN} + R_{CC}\Sigma_{CC}R_{CN} \\ \hline R_{NC} + R_{NC}\Sigma_{CC}R_{CC} & R_{NN} + R_{NC}\Sigma_{CC}R_{CN} \end{array}\right]$$

- Use the properties of conditional multivariate normal distribution

$$\mathbb{E}[\hat{z}_N|\hat{z}_C] = R_{NC}R_{CC}^{-1}\hat{z}_C$$

$$\mathbb{V}[\hat{z}_N|\hat{z}_C] = R_{NN} - R_{NC}R_{CC}^{-1}R_{CN}$$

Evaluate likelihood function $L(\gamma)$ by FINEMAP

- Rewrite the marginal likelihood function $L(\gamma)$:

$$L(\gamma) = P\left(\hat{z}\middle|\gamma, R, \Sigma_\gamma\right) = N\left(\hat{z}\middle|0, R + R\Sigma_\gamma R\right) = P\left(\hat{z}_N\middle|\gamma, \hat{z}_C, R, \Sigma_\gamma\right)P(\hat{z}_C|\gamma, R_{CC}, \Sigma_{CC})$$
$$= N(\hat{z}_C|0, R_{CC} + R_{CC}\Sigma_{CC}R_{CC})N(\hat{z}_N|E[\hat{z}_N|\hat{z}_C], Var(\hat{z}_N|\hat{z}_C))$$

NULL: $L(\gamma = 0) = P(\hat{z}|\gamma = 0, R) = N(\hat{z}_C|0, R_{CC})N(\hat{z}_N|E[\hat{z}_N|\hat{z}_C], Var(\hat{z}_N|\hat{z}_C))$

$$\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R} + \boldsymbol{R}\boldsymbol{\Sigma}_\gamma\boldsymbol{R}) = \mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC}) \times$$
$$\mathcal{N}(\hat{\boldsymbol{z}}_N|\mathbb{E}[\hat{\boldsymbol{z}}_N|\hat{\boldsymbol{z}}_C], \mathbb{V}[\hat{\boldsymbol{z}}_N|\hat{\boldsymbol{z}}_C])$$
$$= \mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC}) \times \frac{\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R})}{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC})}$$

- Bayes factor for assessing the evidence with a given γ against the null model using only causal SNPs (calculation only involves causal SNPs).

Bayes factor (BF) is a likelihood ratio of the marginal likelihood of two competing hypotheses

$$\mathrm{BF}(\boldsymbol{\gamma}:\mathrm{NULL}) = \frac{\mathcal{N}\left(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R} + \boldsymbol{R}\boldsymbol{\Sigma}_\gamma\boldsymbol{R}\right)}{\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R})}$$
$$= \frac{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC})}{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC})}$$

Posterior for γ

- Unnormalized posterior probability

$$\mathrm{P}(\gamma|y,X) = BF(\gamma: NULL) * \left(p_k / \binom{m}{k}\right)$$

- Can be normalized over all $\sum_{k-1}^{K} \binom{m}{k}$ causal configurations
- A Shotgun Stochastic Search (SSS) algorithm (Hans et al. 2007) was used by FINAMAP to rapidly evaluate many configurations and is designed to discover especially those with highest posterior probability

Shotgun Stochastic Search (SSS) algorithm

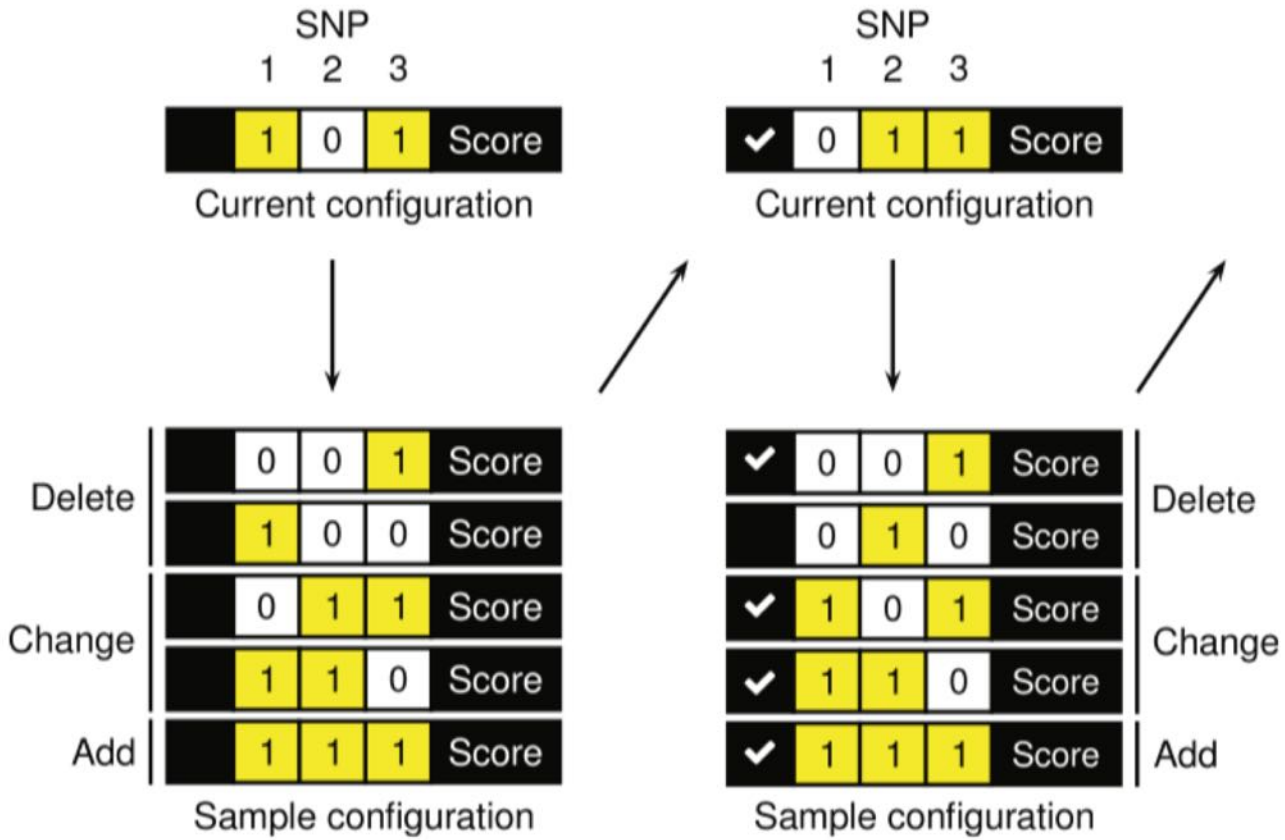


Fig. 2. Shotgun stochastic search rapidly identifies configurations of causal SNPs with high posterior probability. In each iteration, the neighborhood of the current causal configuration is defined by configurations that result from deleting, changing or adding a causal SNP (1) from the current configuration. The next iteration starts by sampling a new causal configuration from the neighborhood based on the scores normalized within the neighborhood. The unnormalized posterior probabilities remain fixed throughout the algorithm and can thus be memorized (✔) to avoid recomputation when already-evaluated configurations appear in another neighborhood

Single-SNP Bayes factor

- Marginal posterior probability that the lth SNP is causal, i.e., single-SNP inclusion probability:

$$p(\gamma_\ell = 1|\boldsymbol{y}, \boldsymbol{X}) = \sum_{\boldsymbol{\gamma} \in \Gamma^*} 1(\gamma_\ell = 1) p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}).$$

- Single-SNP Bayes factor

$$\mathrm{BF}(\gamma_\ell = 1 : \gamma_\ell = 0) = \frac{p(\gamma_\ell = 1|\boldsymbol{y}, \boldsymbol{X})}{p(\gamma_\ell = 0|\boldsymbol{y}, \boldsymbol{X})} \Big/ \frac{p(\gamma_\ell = 1)}{p(\gamma_\ell = 0)},$$

where the prior probability of the ℓth SNP being causal is

$$p(\gamma_\ell = 1) = \sum_{k=1}^{K} \left(\frac{k}{m}\right) p_k.$$

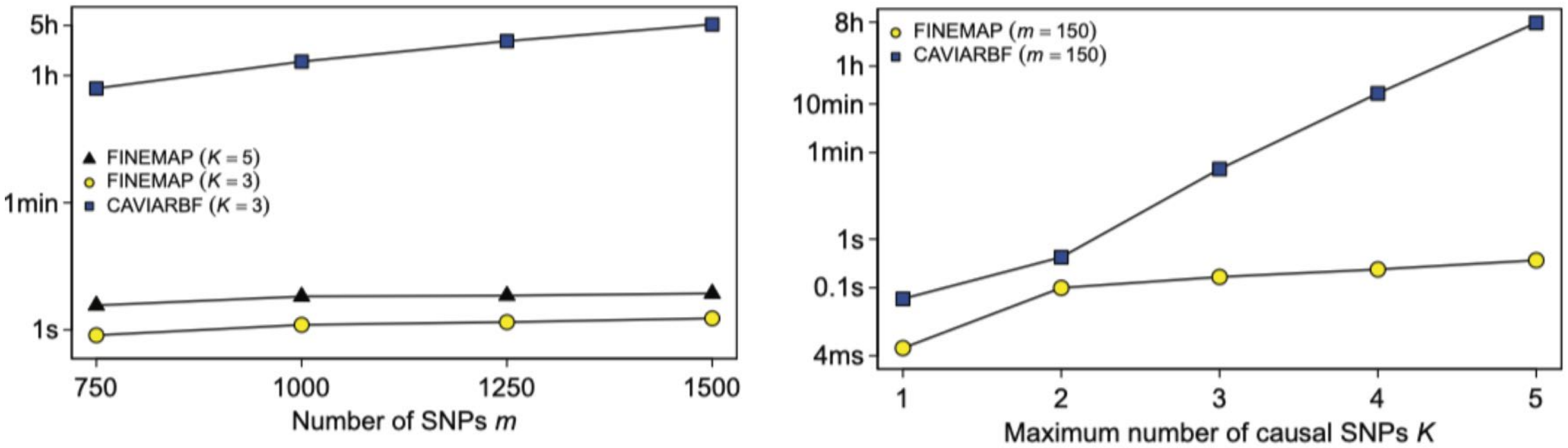


Fig. 3. Processing time of one locus with FINEMAP and CAVIARBF on $\log_{10}$ scale. Top panel: Scenario A with increasing number of SNPs allowing $K = 3$ or $K = 5$ causal SNPs. Bottom panel: Scenario B with 150 SNPs considering causal configurations with different maximum numbers of SNPs. All processing times are averaged over 500 datasets using one core of a Intel Haswell E5-2690v3 processor running at 2.6 GHz

Fine-mapping accuracy

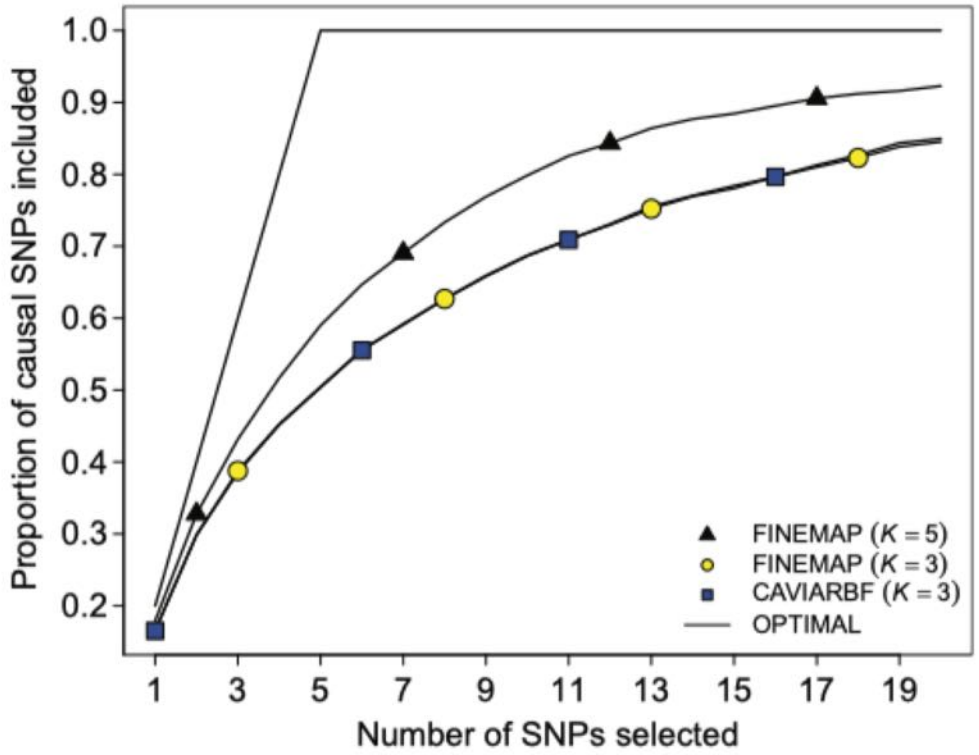


Fig. 5. Fine-mapping accuracy of FINEMAP and CAVIARBF on data with five causal SNPs, allowing either $K = 3$ or $K = 5$ causal SNPs. The proportion of causal SNPs included is plotted against the number of top SNPs selected on the basis of ranked single-SNP inclusion probabilities. Proportions are averaged over 500 datasets with 1500 SNPs. Case $K = 5$ is computationally intractable for CAVIARBF

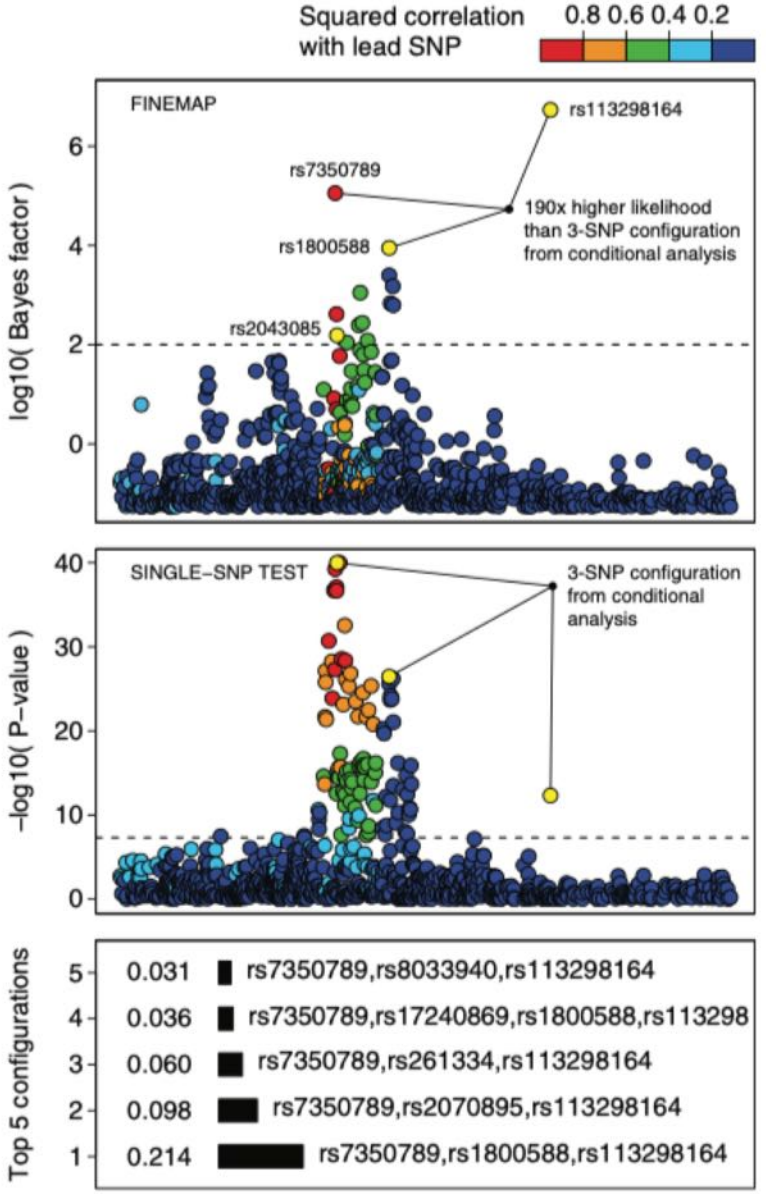


Fig. 7. Fine-mapping of 15q21/*LIPC* region associated with high-density lipoprotein cholesterol. Independent association signals in conditional analysis are highlighted by ○. Dashed lines correspond respectively to a single-SNP Bayes factor of 100 and *P*-value of 5×10^{-8}. Squared correlations are shown with respect to rs2043085

Account for LD in GWAS by Multivariate Regression Model

- Consider the multivariate regression model with all genome-wide variants in the genotype matrix X

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

- Variable selection is needed
- Memory and computation issues for ~10 Million SNPs

LASSO : Least Absolute Shrinkage and Selection Operator

- Lasso-penalized least squares objective function

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|.$$

- Tunning penalty parameter λ (S. Yang et. al., Bioinformatics, 2020).
- Solve for estimates of genetic effect sizes: β
- R function “glmnet()”

Bayesian Variable Selection Regression

- Consider the multivariate regression model with a point-and-spike prior on β_l

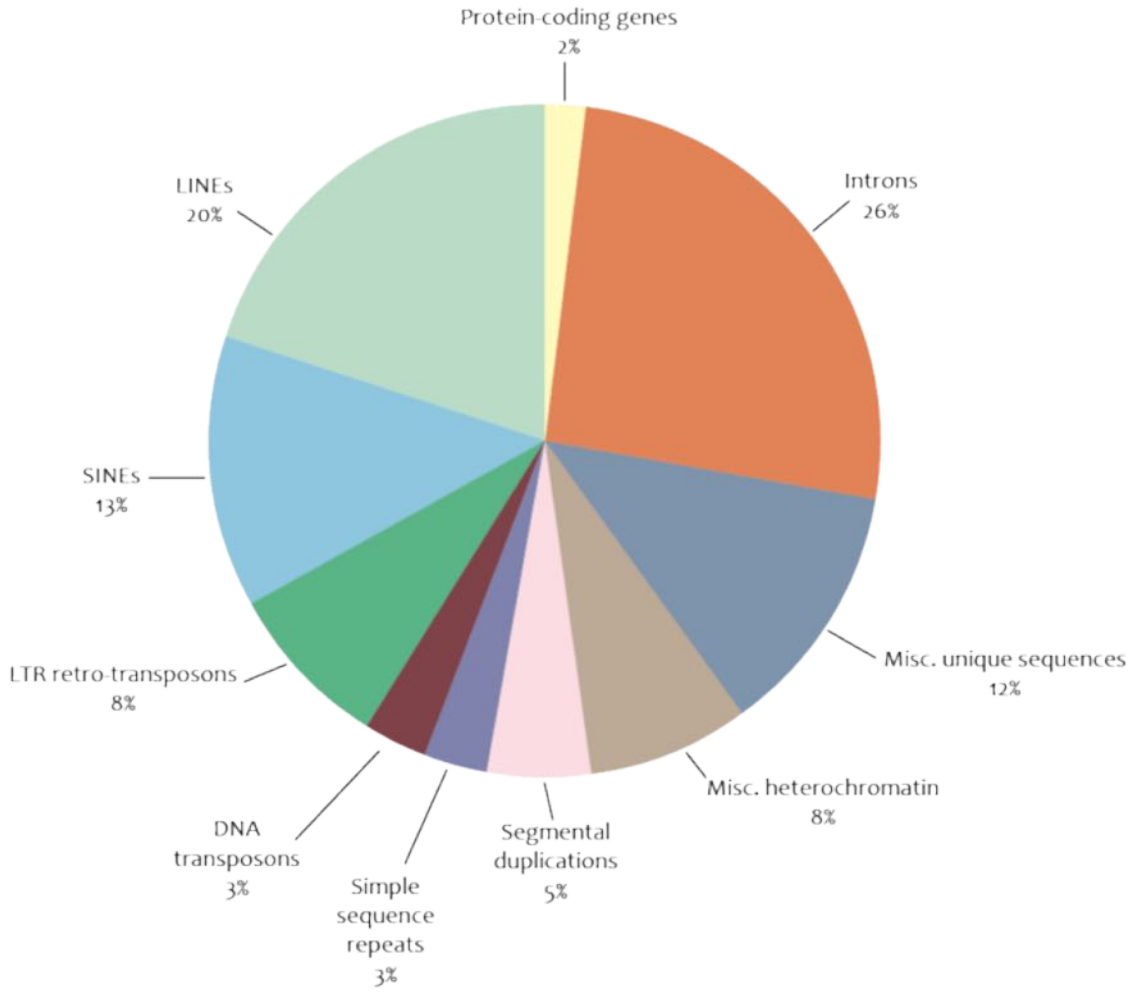
$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \tau^{-1}I)$$
$$\beta_l \sim \pi N\left(0, \tau^{-1}\sigma_\beta^2\right) + (1-\pi)\delta_0(\beta_l), l = 1, \dots, m$$
$$\sigma_\beta^2 \sim InverseGamma(k_1, k_2), \pi \sim Beta(a, b)$$
$$\tau \sim Gamma(k_3, k_4),$$

- Assume an indicator vector γ, that is,

$$\gamma_l \sim Bernoulli(\pi)$$

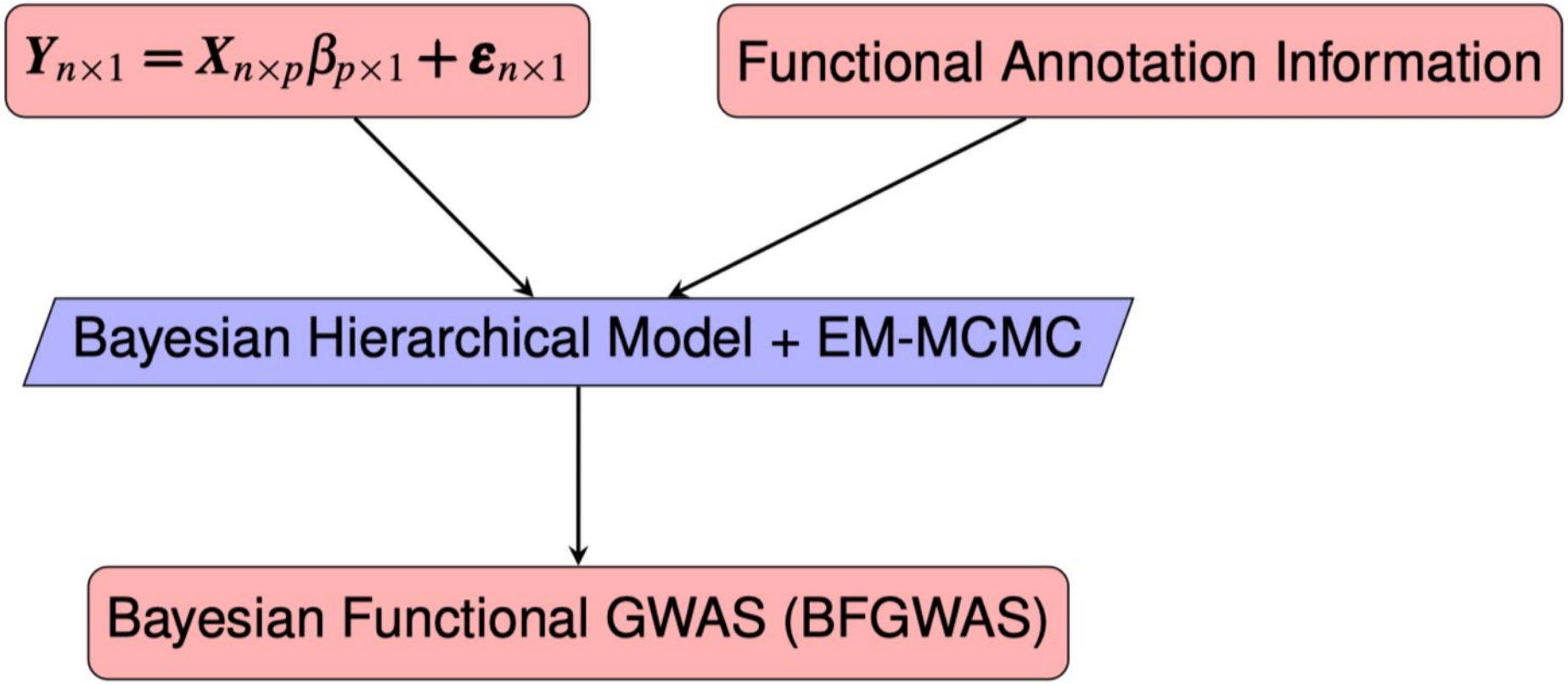
- Inference goal: estimate $\beta_l, E[\gamma_l], \sigma_\beta^2, \pi$
- Approach: Monte Carlo Markov Chain (MCMC) (Guan and Stephens, 2011)
- Convergence would be an issue for studying ~10 Millions SNPs

Account for Functional Annotation



- Prioritize functional SNPs
- Quantify enrichment of each type of functional SNPs with respect to GWAS associations

Bayesian Functional GWAS



Bayesian Hierarchical Model

Joint linear regression model

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

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

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

J. Yang et. al. AJHG. 2017

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
  - $\pi_q \sim Beta(a_q, b_q)$;
  - $\sigma_q^2 \sim InverseGamma(k_1, k_2)$;
  - $\tau \sim Gamma(k_3, k_4)$

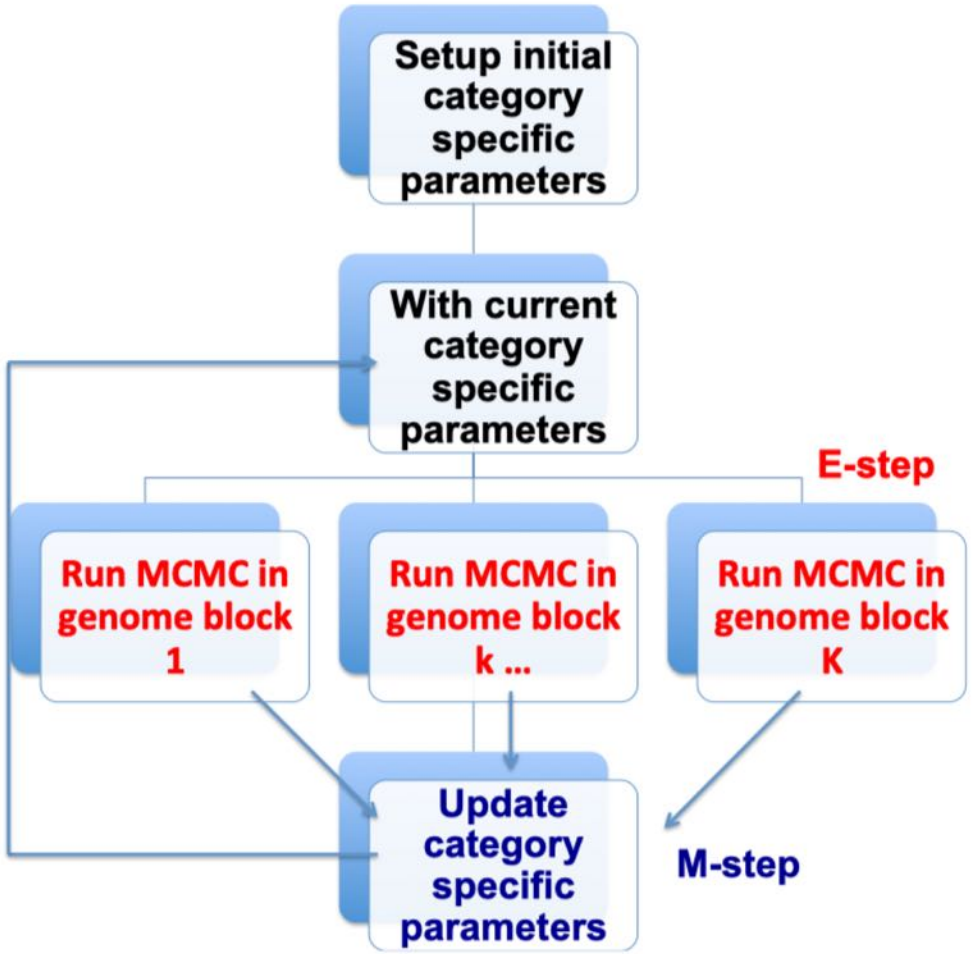
- The joint posterior distribution

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

- 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

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

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

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

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

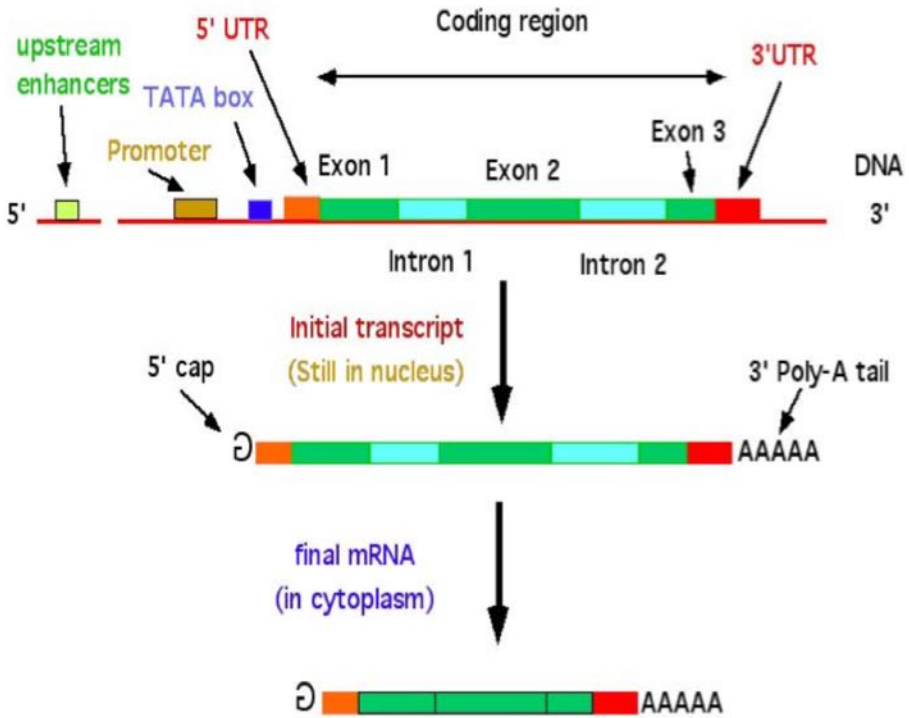
Apply BFGWAS to Study Age-related Macular Degeneration (AMD)

- ∼10M low-frequency and common variants (MAF>0.5%)
- ∼ 16K cases and ∼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:

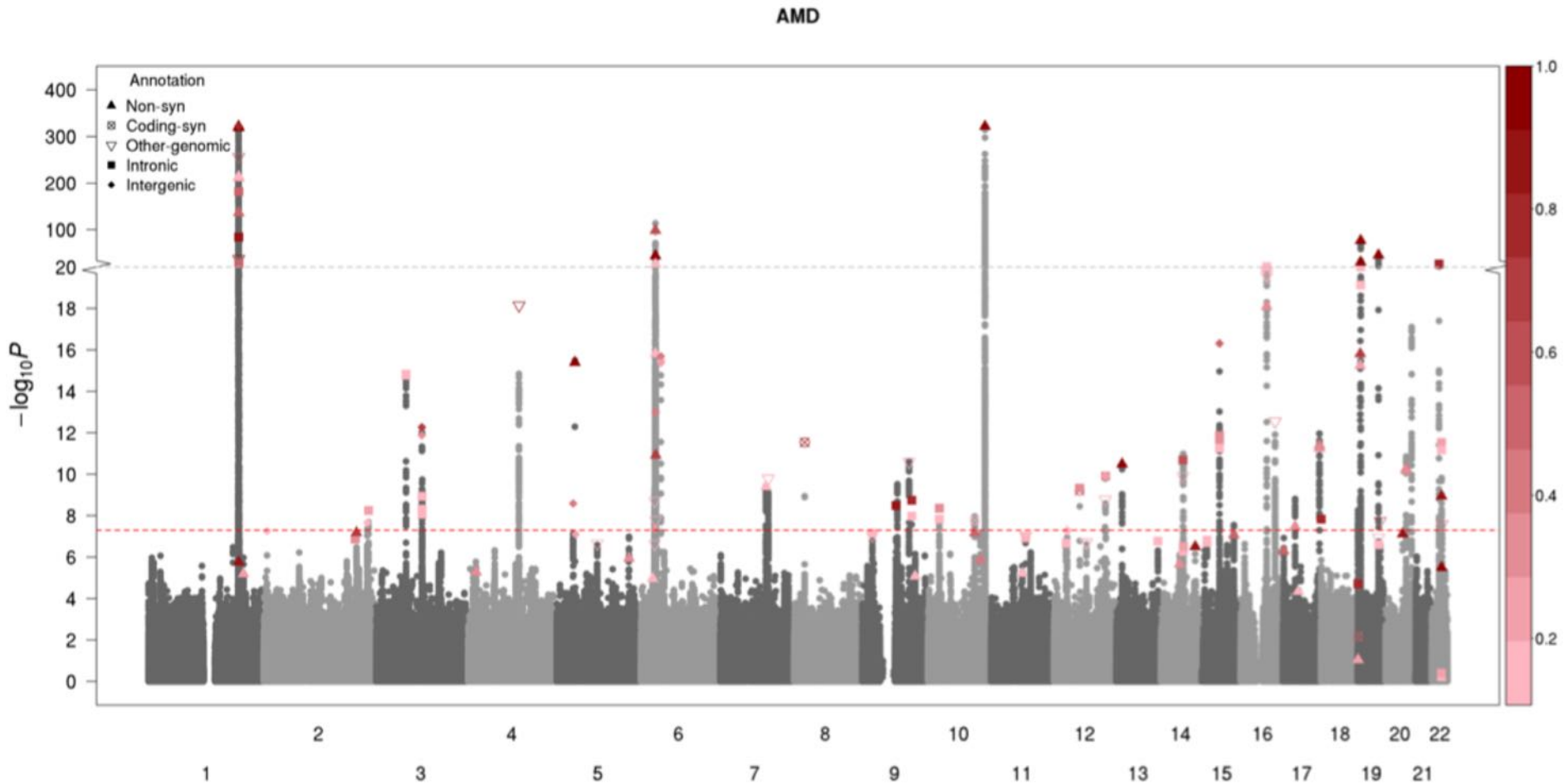
- Non-synonymous (42,005)
- Synonymous (67,165)
- Intronic (3,679,235)
- Intergenic (5,512,423)
- Other genomic (565,916, UTR, non-coding exons, upstream and downstream)



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).

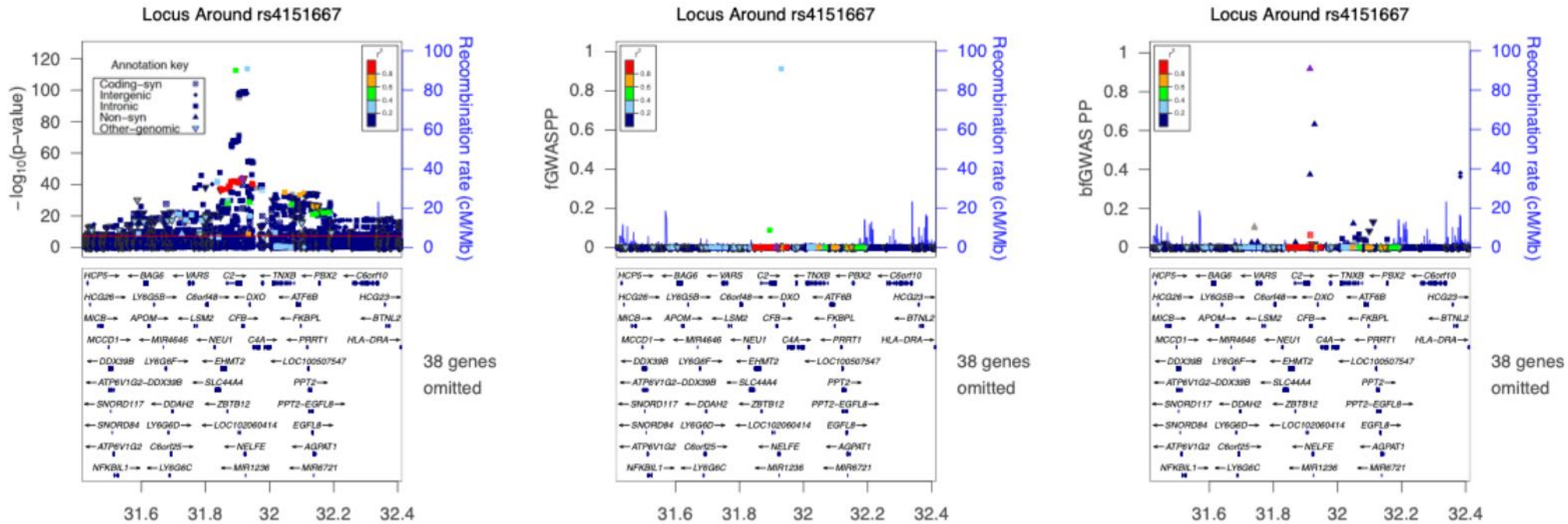


Figure 3: GWAS (left) vs. FGWAS (middle; Pickrell JK, AJHG 2014) vs. BFGWAS (right) for example locus #8.

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*.

Enrichment Results

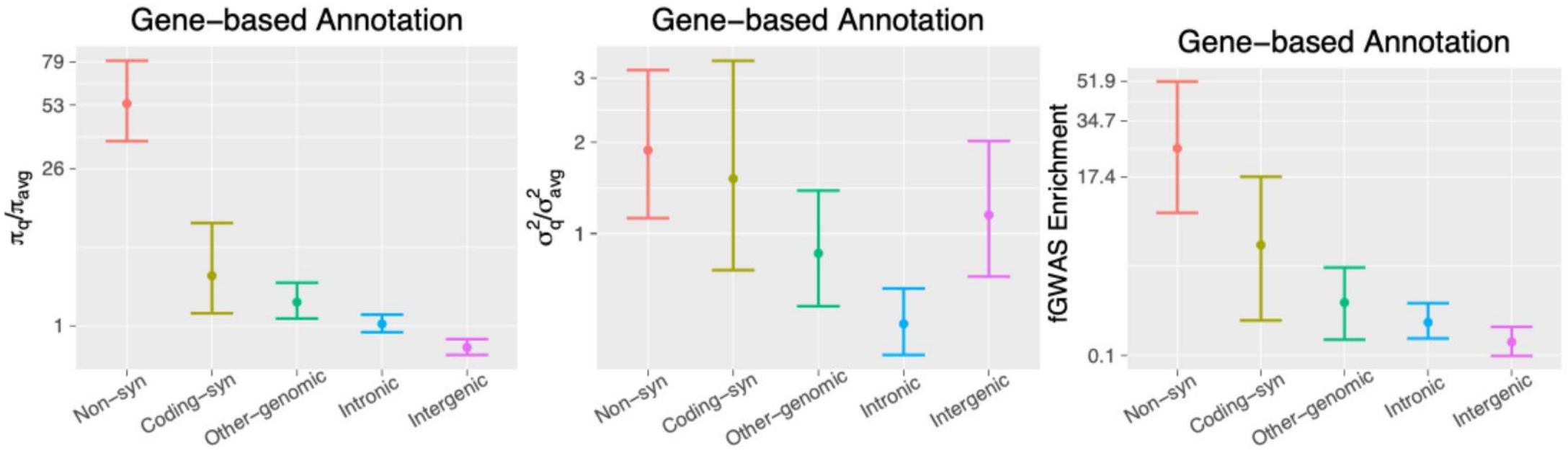


Figure 5: BFGWAS enrichment Results (left, middle) vs. FGWAS (right).

Available Tools

- GEMMA: GWAS and SNP heritability estimation by LMM, BVSR
  https://github.com/genetics-statistics/GEMMA

- FINEMAP: Fine-mapping GWAS results
  http://www.christianbenner.com

- BFGWAS: Bayesian Functional GWAS
  https://github.com/yjingj/bfGWAS

Topics for Next Lecture

- Rare Variant Test
  - Burden Test
  - Variance Component Test
- Pleiotropy
  - Model Multiple Phenotypes
- Mendelian Randomization
  - Mediation Analysis