Genome-wide Association Studies

BIOS 770

02/15/2022

Jingjing Yang (jingjing.yang@emory.edu)

Outline

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

Rare Variants

Genetic variants identified in the TOPMed project (2015 - Present)

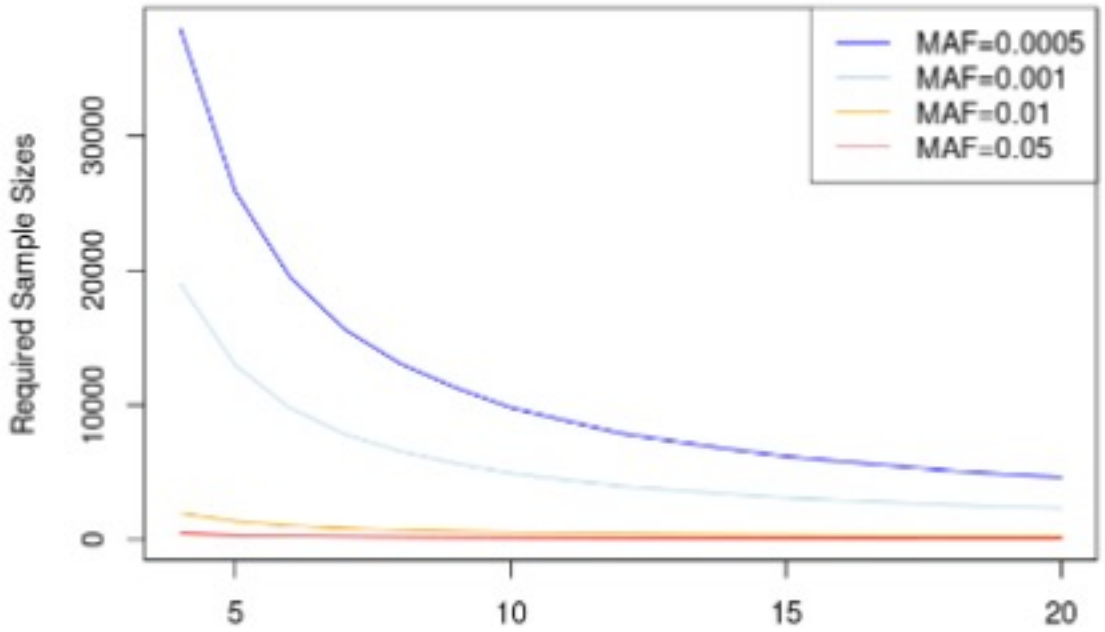
Variant Type	Category	# PASS	# FAIL	% dbSNP (PASS)	Known/Novel Ts/Tv (PASS)
SNP	**All**	**438M**	**85M**	**22.9%**	**1.93 / 1.69**
	Singleton	202M	24M	8.5%	1.23 / 1.54
	Doubleton	69M	8.8M	12.6%	1.61 / 1.74
	Tripleton ~ 0.1%	142M	24M	34.9%	2.23 / 1.99
	0.1% ~ 1%	13M	4.5M	98.2%	2.17 / 1.79
	1 ~ 10%	6.5M	2.9M	99.6%	1.82 / 1.75
	>10%	5.3M	2.0M	99.8%	2.11 / 1.88
Indels	**All**	**33.4M**	**26.2M**	**20.1%**	
	Singleton	15.7M	4.7M	10.1%	
	Doubleton	5.3M	1.8M	12.6%	
	Tripleton ~ 0.1%	10.7M	8.0M	26.7%	
	0.1% ~ 1%	2.8M	968K	88.9%	
	1 ~ 10%	432K	2.3M	98.5%	
	>10%	298K	1.4M	99.6%	

Why Study Rare Variants?

- Most genetic variants are rare
- Functional variants tend to be rare
- Number of samples N needed to observe a rare variant with at least 99.9% probability :

MAF	0.1	0.01	0.001	0.001
N	33	344	3453	34537

- Number of samples needed to achieve 80% power by single variant test (underpowered for rare variants):



Region-based (or Gene-based) Test

Test the joint effect of rare/common variants within a defined genome region (e.g., gene, regulatory region)

- Consider a total of p variants within a test genome region
- Genotype data: $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})', \ x_{ij} = 0, 1, 2$, for individual i
- Genetic effect-size vector $\beta_{1\times p}$
- Consider covariate vector Z_i for individual i, with extended intercept term
- General linear regression model

$$E[f(\mu_i)] = Z_i'\alpha + X_i'\beta \ ,$$

  where $f(\cdot)$ is a link function, e.g., logistic function for dichotomous traits, identity function for quantitative traits.

- Test region-based association

$$H_0 : \beta = (\beta_1, \cdots, \beta_p) = 0$$

Burden Test

- Assume there is a shared genetic effect across all variants, i.e., $\beta_j = w_j \beta_{burden}$
- One commonly used variant weight is based on MAF,
  $w_j = 1/\sqrt{MAF_j(1 - MAF_j)}$
- Equivalent to consider a Burden genotype score

$$C_i = \sum_j w_j x_{ij}$$

- Model is equivalent to

$$E[f(\mu_i)] = Z_i'\alpha + C_i\beta_{burden}$$

- Region-based test is equivalent to test

$$H_0 : \beta_{burden} = 0$$

Burden Test

- Score test statistic can be used

$$T_{score} = \sum_{i=1}^{n} C_i(Y_i - \widehat{\mu_{0i}}) = \sum_{j=1}^{p} w_j X'_{\cdot,j}(Y - \widehat{\mu_0}),$$

$$Var(T_{score}) = C'\left(\widehat{P} - \widehat{P}Z(Z'\widehat{P}Z)^{-1}Z'\widehat{P}\right)C;$$

$$\frac{T_{score}}{\sqrt{Var(T_{score})}} \sim N(0,1), \text{ Under } H_0$$

  - Let $\widehat{\alpha}$ denote the covariate coefficients estimated under the NULL model $E[f(\mu_i)] = Z_i\alpha$
  - $\widehat{\mu_{0i}} = Z_i'\widehat{\alpha}$ and $\widehat{P} = \widehat{\sigma_\epsilon^2}I$, with error variance estimate $\widehat{\sigma_\epsilon^2}$ under NULL model, for standard linear regression model
  - $\widehat{\mu_{0i}} = logit^{-1}(Z_i\widehat{\alpha})$ and $\widehat{P} = diag\left(\widehat{\mu_{01}}(1-\widehat{\mu_{01}}), \cdots, \widehat{\mu_{0n}}(1-\widehat{\mu_{0n}})\right)$ for logistic regression model

- Sum of weighted single variant score test statistics $T_j = X'_{\cdot,j}(Y - \widehat{\mu_0}), j = 1, \cdots, p.$

Underpowered when genetic variants have Effect-sizes of Opposite Signs

Burden Test

- Burden test
  - Cohort Allelic Sum Test (CAST, Morgenthaler & Thilly, 2006) : collapses information on all rare variants within a region into a single dichotomous variable per sample
  - Weighted sum test (WST, Madsen & Browning, 2009) : collapses rare variants into a single weighted average of the number of rare alleles per sample

- Limitations: all rare variants are assumed influencing the phenotype in the same direction and with the same magnitude of effect, after incorporating known weights

Variance Component Test

- Consider the general linear regression model

$$E[f(\mu_i)] = Z_i'\alpha + X_i'\beta \ ,$$

- Assume $\beta_j \sim N(0, w_j^2\tau)$, sharing a common variance component τ
- Then region-based test is equivalent to test

$$H_0 : \tau = 0$$

Variance Component Test : SKAT

- Test statistic (Sum of squared weighted single variant score test statistics) :

$$Q_{SKAT} = (Y - \widehat{\mu_0})XWWX'(Y - \widehat{\mu_0}) = \sum_{j=1}^{p} \left(w_j X_j'(Y - \widehat{\mu_0})\right)^2$$

- Let $K = XWWX'$ denote the kernel matrix, $W = diag(w_1, \cdots, w_p)$
- $\widehat{\mu_0}$ is estimated under the NULL hypothesis
- Q_{SKAT} asymptotically follows a mixture of $\chi^2_{(1)}$ distribution under the NULL hypothesis

$$Q_{SKAT} \approx \sum_{j=1}^{p} \lambda_j \chi^2_{(1)}$$

- λ_j are eigenvalues of $P_0^{1/2} K P_0^{1/2}$, with projection matrix $P_0 = \widehat{P} - \widehat{P}Z(Z'\widehat{P}Z)^{-1}Z'\widehat{P}$.
- The mixture of $\chi^2_{(1)}$ distribution can be approximated with the computationally efficient Davies method, which will be used to calculate p-value.

Wu et al. (2011) AJHG

Power Comparisons of SKAT and Burden Tests

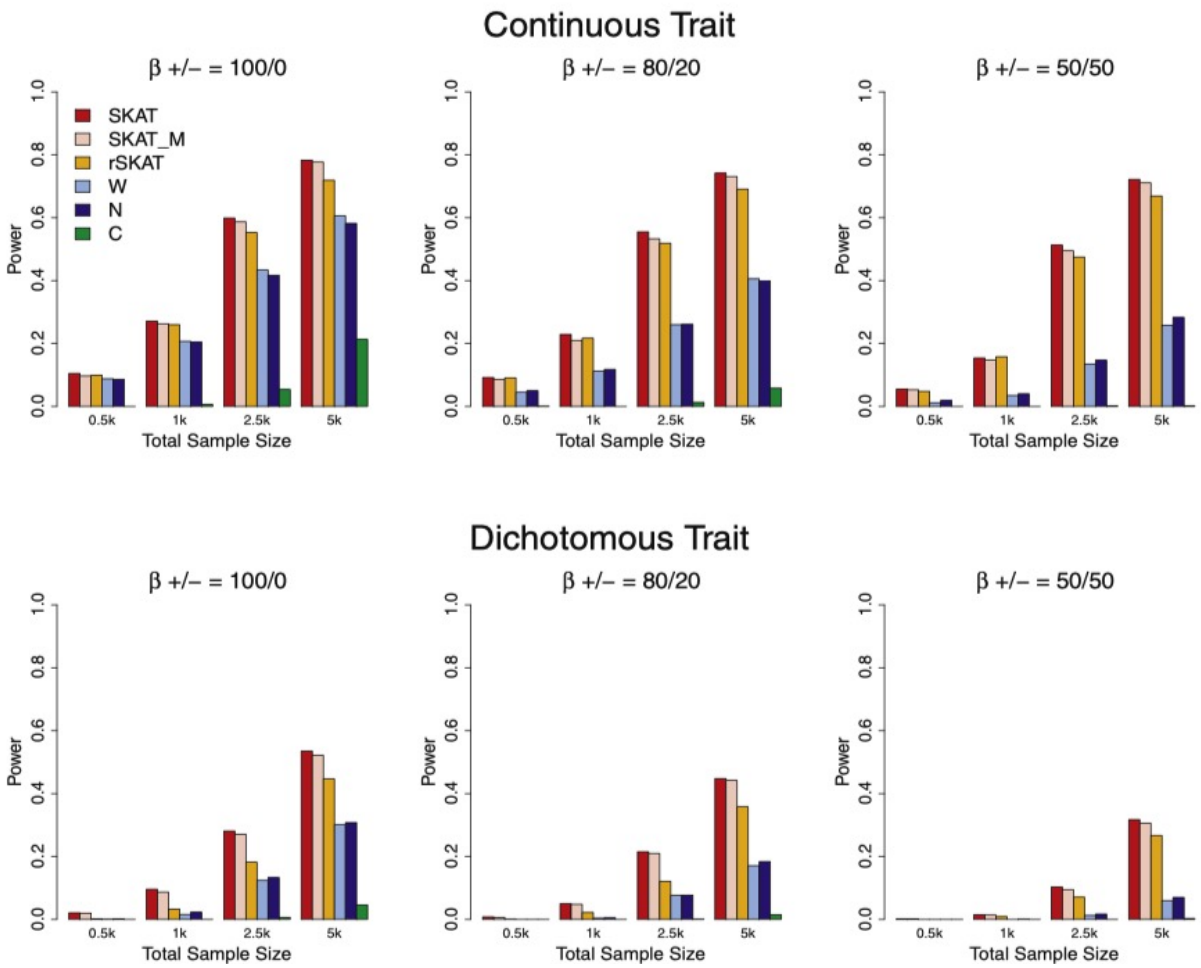


Figure 1. Simulation-Study-Based Power Comparisons of SKAT and Burden Tests
Empirical power at $\alpha = 10^{-6}$ under an assumption that 5% of the rare variants with MAF < 3% within random 30 kb regions were causal. Top panel: continuous phenotypes with maximum effect size (|β|) equal to 1.6 when MAF = 10^{-4}; bottom panel: case-control studies with maximum OR = 5 when MAF = 10^{-4}. Regression coefficients for the *s* causal variants were assumed to be a decreasing function of MAF as $|\beta_j| = c\,|\log_{10} MAF_j|$ $(j = 1,\ldots,p$ [see Figure S2]), where *c* was chosen to result in these maximum effect sizes. From left to right, the plots consider settings in which the coefficients for the causal rare variants are 100% positive (0% negative), 80% positive (20% negative), and 50% positive (50% negative). Total sample sizes considered are 500, 1000, 2500, and 5000, with half being cases in case-control studies. For each setting, six methods are compared: SKAT, SKAT in which 10% of the genotypes were set to missing and then imputed (SKAT_M), restricted SKAT (rSKAT) in which unweighted SKAT is applied to variants with MAF < 3%, the weighted sum burden test (W) with the same weights as used by SKAT, counting-based burden test (N), and the CAST method (C). All the burden tests used MAF < 3% as the threshold. For each method, power was estimated as the proportion of p values < α among 1000 simulated data sets.

Sample Sizes Required for Reaching 80% Power

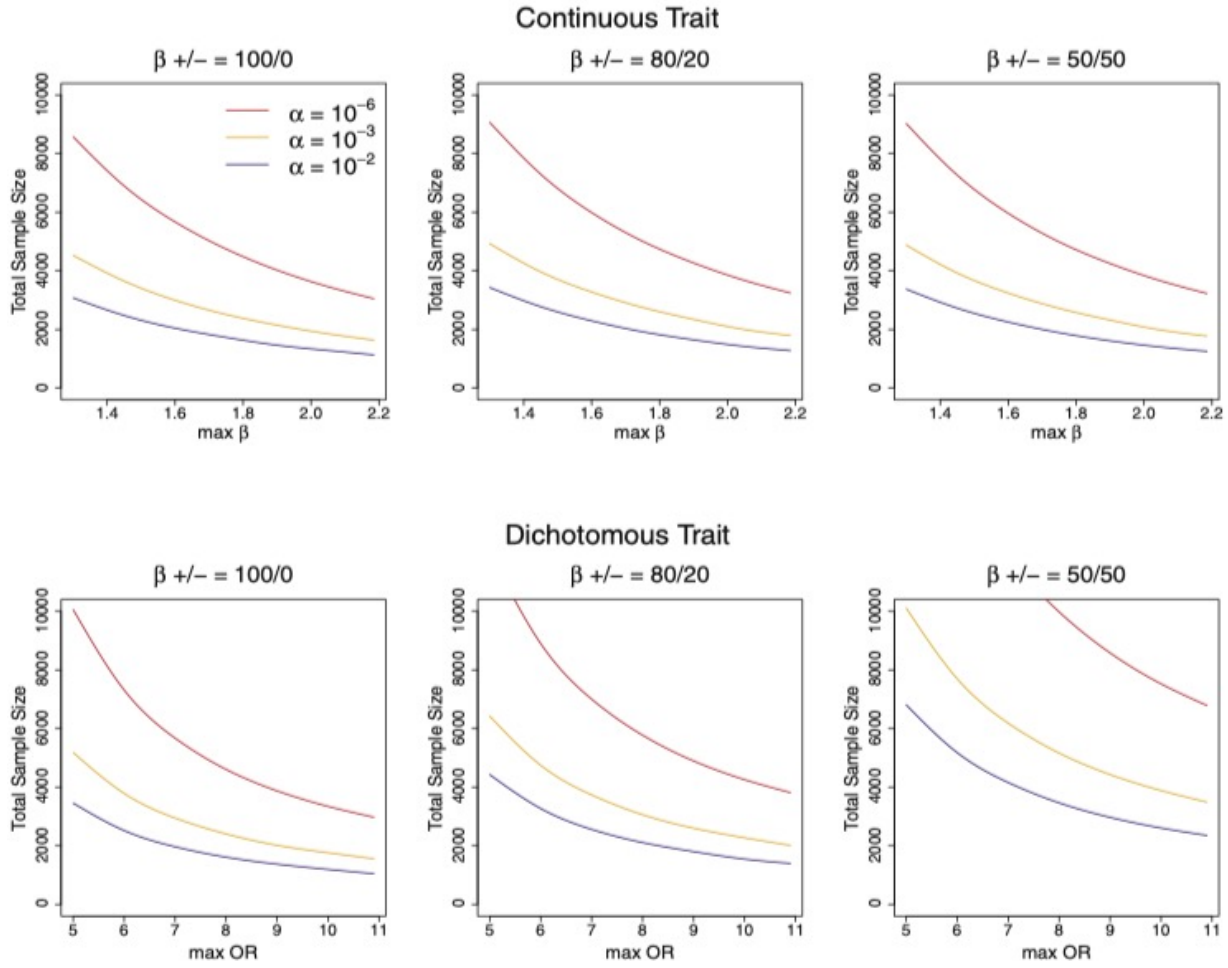


Figure 2. Sample Sizes Required for Reaching 80% Power
Analytically estimated sample sizes required for reaching 80% power to detect rare variants associated with a continuous (top panel) or dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the $\alpha = 10^{-6}$, 10^{-3}, and 10^{-2} levels, under the assumption that 5% of rare variants with MAF < 3% within the 30 kb regions are causal. Plots correspond to 100%, 80%, and 50% of the causal variants associated with increase in the continuous phenotype or risk of the dichotomous phenotype. Regression coefficients for the *s* causal variants were assumed to be the same decreasing function of MAF as that in Figure 1. The absolute values of Required total sample sizes are plotted again the maximum effect sizes (ORs) when MAF = 10^{-4}. Estimated total sample sizes were averaged over 100 random 30 kb regions.

Practices of Rare Variant Tests

- QC
- Filter out common variants (especially those associated with the phenotype of interest)
- Group by functional annotations (i.e., known biological functions of the genetic variants)
- Account for possible confounding covariates: age, sex, genotype PCs, etc.
- Genomic control factor and meta-analysis methods also apply
- Visualize by Manhattan and QQ plots

Pleiotropy

- Pleiotropy: One gene can affect multiple traits
- Example pleiotropy in chickens and the Frizzle Trait:

  In 1936, researchers Walter Landauer and Elizabeth Upham observed that chickens that expressed the dominant frizzle gene produced feathers that curled outward rather than lying flat against their bodies (Figure 2). However, this was not the only phenotypic effect of this gene — along with producing defective feathers, the frizzle gene caused the fowl to have abnormal body temperatures, higher metabolic and blood flow rates, and greater digestive capacity. Furthermore, chickens who had this allele also laid fewer eggs than their wild-type counterparts, further highlighting the pleiotropic nature of the frizzle gene.

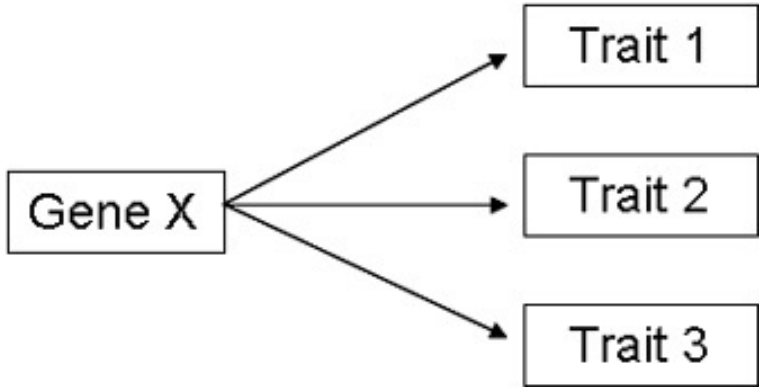


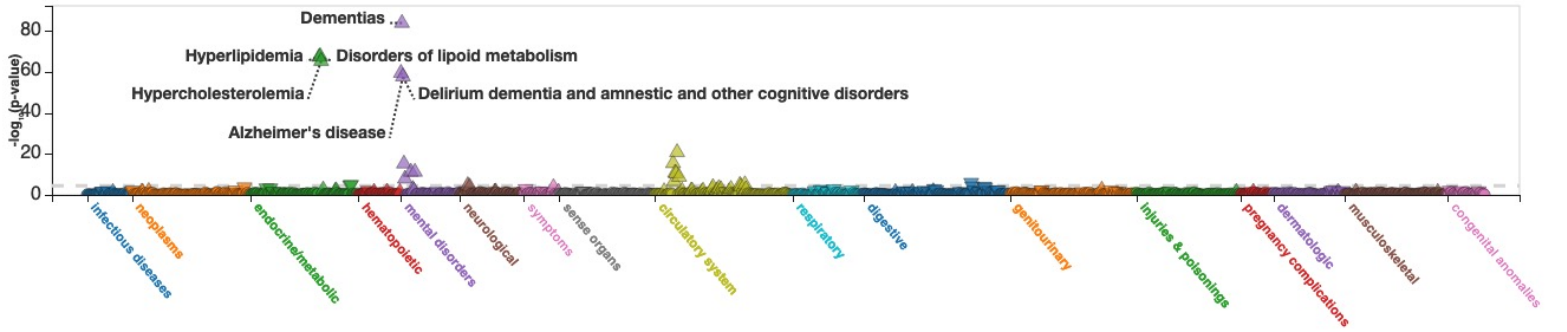
Figure 2: A chicken with the frizzle gene
© 2004 Richard Blatchford, Dept. of Animal Science UC Davis. All rights reserved.

Example Pleiotropy at *APOE E4* allele (rs429358) in Human

19 : 45,411,941 T / C (rs429358)

Nearest gene: *APOE*
AF ranges from 0.15 to 0.16
View on UCSC , GWAS Catalog , dbSNP , **PubMed (224 results)** , **Clinvar**

Save PNG | Save SVG



Search... "427.21", "Diabetes", etc.

1403 total codes

Category	Phenotype	P-value	Effect Size (se)	Number of samples
mental disorders	Dementias	1.9e-84	1.4 (0.074)	956 / 402383
endocrine/metabolic	Hyperlipidemia	3.0e-68	0.20 (0.012)	35844 / 373034
endocrine/metabolic	Disorders of lipoid metabolism	6.0e-68	0.20 (0.012)	35927 / 373034
endocrine/metabolic	Hypercholesterolemia	6.2e-66	0.21 (0.012)	33242 / 373034
mental disorders	Delirium dementia and amnestic and other cognitive disorders	3.9e-60	0.79 (0.048)	1970 / 402383
mental disorders	Alzheimer's disease	2.3e-58	1.9 (0.12)	404 / 402383
circulatory system	Coronary atherosclerosis	1.7e-21	0.14 (0.015)	20023 / 377103
circulatory system	Ischemic Heart Disease	3.9e-16	0.10 (0.012)	31355 / 377103
mental disorders	Vascular dementia	6.1e-16	1.3 (0.16)	189 / 402383
mental disorders	Neurological disorders	5.7e-12	0.20 (0.029)	4655 / 402383

First | ← Previous | 1 | 2 | 3 | 4 | 5 | Next → | Last

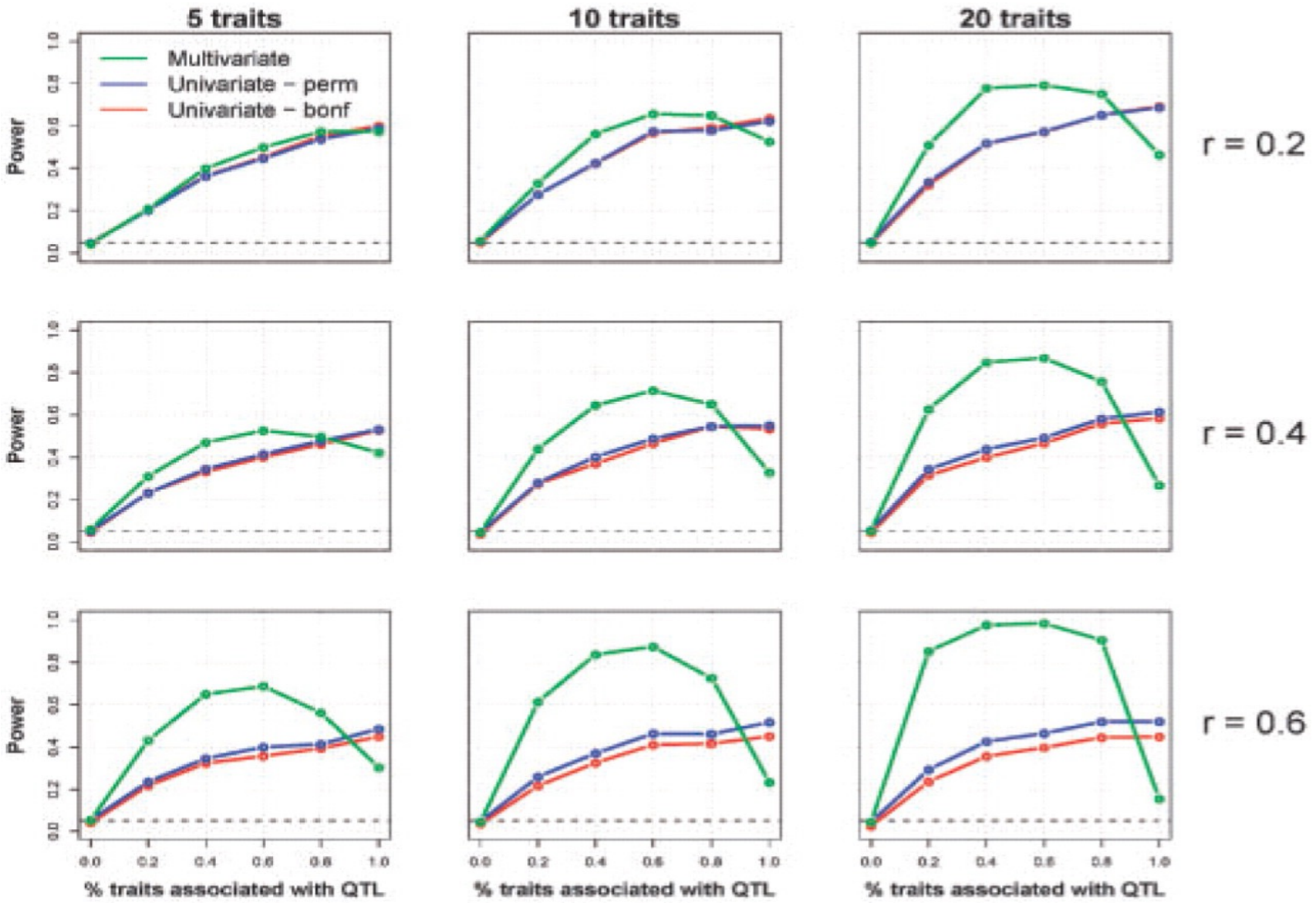
Testing Genetic Pleiotropy

- H_0: No trait ($Y = (Y_1, \ldots, Y_K)$) is associated with the test SNP x
- F-statistic (assume normally distributed quantitative traits)
  - $lm(Y \sim x)$: multiple response variables
  - $lm(x \sim Y)$: reverse regression with multiple explanatory variables
  - Test canonical correlation of $Y \sim x$ (used by plink.multivariate, Ferreira & Purcell, Bioinformatics, 2019)
- Test association between top Principal Component of $Y = (Y_1, \ldots, Y_K)$ and SNP x

Canonical Correlation Analysis (CCA)

- Assume normally distributed traits: $Y_{n\times K} = (Y_1, \dots, Y_K)$
- SNP genotype vector: $x_{1\times n}$
- CCA extracts the linear combination of traits that explain the largest possible amount of the covariation between the marker and all traits
  - $b' = \arg\_\max\_corr(x, b'Y)$
  - Solving for b' that maximizes $\rho = corr(x, b'Y)$
- Test statistic Wilk's lambda : $\lambda = 1 - \hat{\rho}^2$,
  - Canonical correlation estimate $\hat{\rho}$ is given by the square root of the eigenvalue of
    $$\Sigma_{x,x}^{-\frac{1}{2}}\Sigma_{x,Y}\Sigma_{Y,Y}^{-1}\Sigma_{Y,x}\Sigma_{x,x}^{-\frac{1}{2}}$$
  - b is an eigenvector of $\Sigma_{Y,Y}^{-1}\Sigma_{Y,x}\Sigma_{x,x}^{-1}\Sigma_{x,Y}$
  - F-approximation: $\mathrm{F_{K,n-K-1}} = [(1-\lambda)\lambda] * \left[\frac{n-K-1}{K}\right]$

Performance of the multivariate test of association.



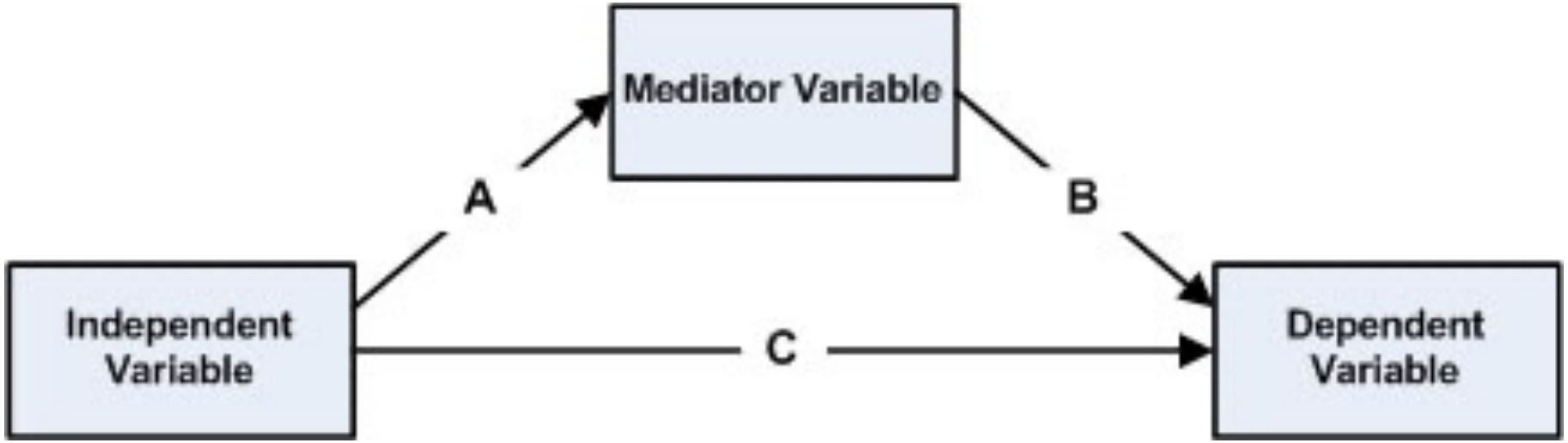
- 60% heritability
- 20% minor allele frequency
- r=0.2, 0.4, 0.6 denoting residual cross-trait correlation

Bioinformatics, Volume 25, Issue 1, 1 January 2009, Pages 132–133, https://doi.org/10.1093/bioinformatics/btn563

The content of this slide may be subject to copyright: please see the slide notes for details.

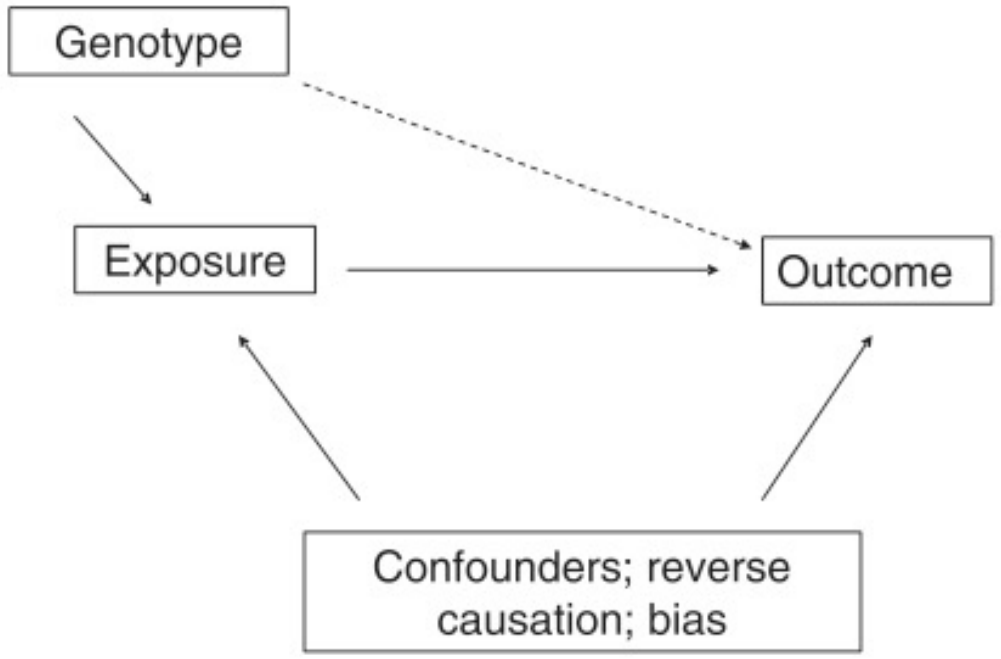
Mediation Analysis

- Mediation Analysis seeks to identify the mechanism that the Independent Variable (Instruments) affects the Dependent Variable (Response) through the Mediator Variable



Mendelian Randomization

- <u>Assumptions for Instruments</u> (Didelez and Sheehan, 2007).
  1. Not associated with any confounder of the Exposure-Outcome association (IV 1)
  2. Associated with the Exposure/Mediator (IV 2)
  3. Conditionally independent of the outcome given the Exposure and confounders (IV 3)

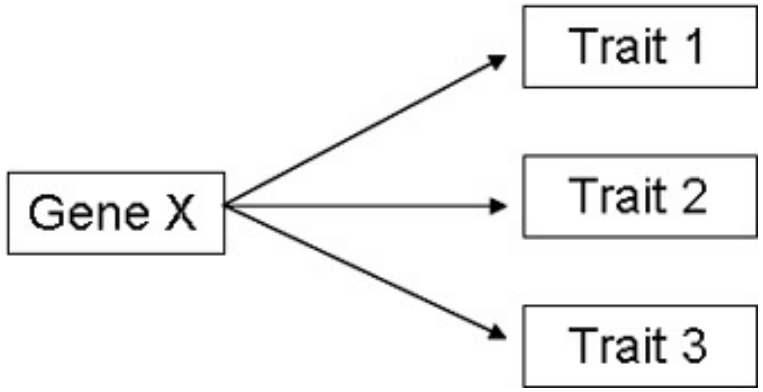


Mendelian Randomization

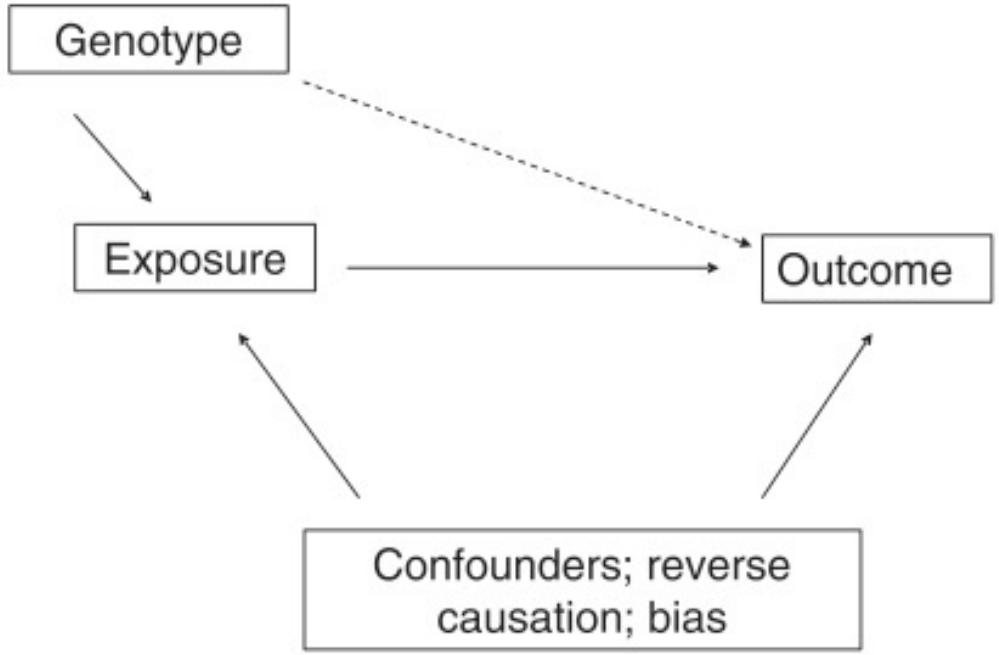
- Mendelian Randomization uses genetic variants as instruments for Mediation Analysis
  - Under the Mendel's Second Law, genotypes are assigned randomly when passed from parents to offspring during meiosis (independent assortment)
  - Alternative of traditional randomized trials to estimate a putative causal effect of the mediator on phenotype
  - Instruments are proxies of the exposure that are free of confounders

Pleiotropy vs. Mediation

- Pleiotropy



- Mediation



Model

Let X denote genotype, M denote mediator, Y denote outcome

- Genetic model for M: $M = \beta_0 + \beta_{XM}X + \varepsilon$
- Genetic model for Y: $Y = \widetilde{\beta}_0 + \beta_{XY}X + \varepsilon$
- Joint model for Y

$$Y = \beta_0 + \beta_{direct}X + \beta_{causal}M + \varepsilon$$

$$Y = \widetilde{\beta}_0 + \beta_{direct}X + \beta_{causal}\beta_{XM}X + \varepsilon$$

$$Y = \widetilde{\beta}_0 + (\beta_{direct} + \beta_{causal}\beta_{XM})X + \varepsilon$$

- If $\beta_{direct} = 0$, $\beta_{XY} = \beta_{causal}\beta_{XM}$, equivalently $\beta_{causal} = \beta_{XY}/\beta_{XM}$
- If $\beta_{direct} \neq 0$, $\beta_{XY} = \beta_{direct} + \beta_{causal}\beta_{XM}$

Goal: test if β_{causal} significantly different from 0.

Test Methods

- Inverse-Variance Weighted (IVW) method : Estimate the ratio between SNP-outcome association and SNP-exposure association using a meta-analysis approach
- MR-Egger Regression (Bowden et al., 2015) : Weighted linear regression of the SNP-outcome coefficients on the SNP-exposure coefficients. Without an intercept term in the regression model, MR-Egger slope estimate will equal to the IVW estimate.
- SMR & GSMR (Zhu et. al., 2016, Zhu et. al., 2018):

Inverse-Variance Weighted (IVW) method

- Single variant test results are available for testing $\mathrm{M} \sim SNP_j\ X_j$
- Single variant test results are available for testing $\mathrm{Y} \sim SNP_j\ X_j$
- Assume no direct causal effect $SNP_j \rightarrow Y$ (Assumption IV3), the causal effect mediated through M is estimated by

$$\hat{\beta}_{causal,j} = \frac{\hat{\beta}_{X_jY}}{\hat{\beta}_{X_jM}}$$

- With multiple genetic variants (in linkage equilibrium), the above causal effect of $\mathrm{SNP_j} \rightarrow M \rightarrow Y$ can be estimated using the weighted average of the ratio estimates per SNP (analogous to the inverse-variance meta-analysis method)

$$\hat{\beta}_{causal} = \frac{\sum_{j=1,\ldots,J} w_j \hat{\beta}_{causal,j}}{\sum_{j=1,\ldots,J} w_j}, w_j = \hat{\beta}^2_{X_jM} / \sigma^2_{\{Y \sim X_j\}}$$

IVW Example

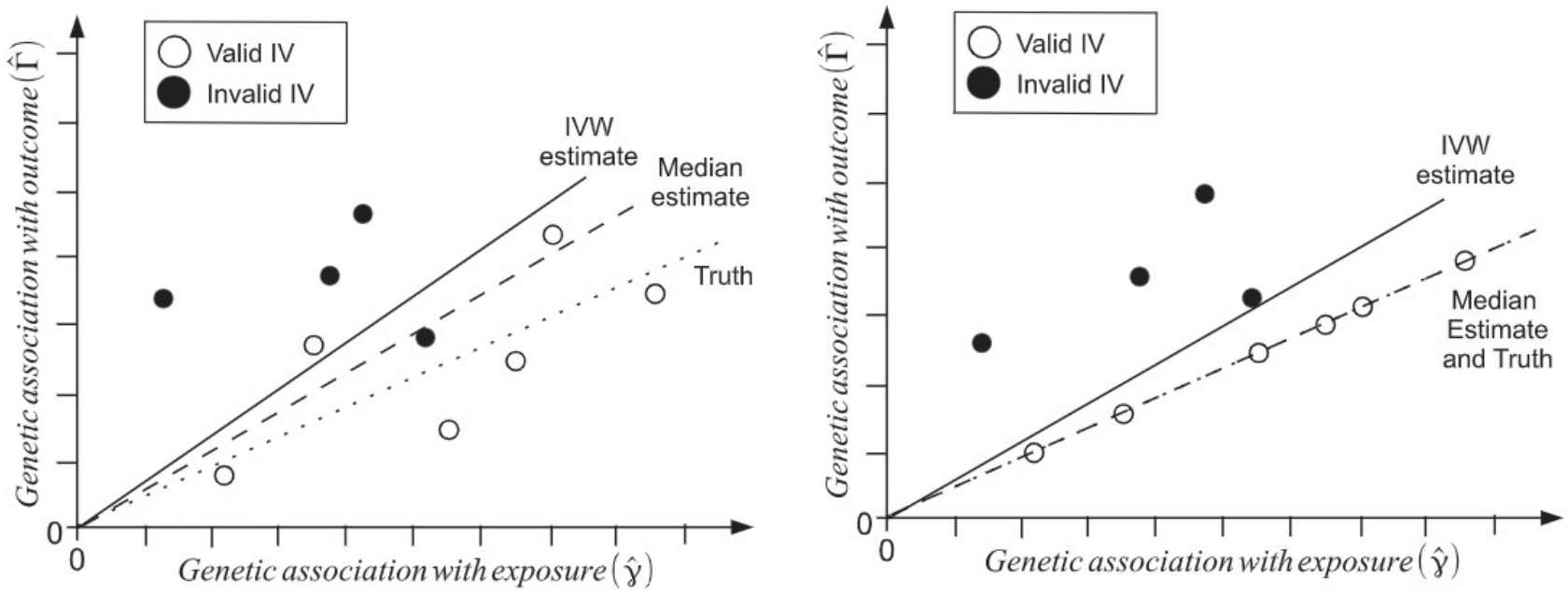


Figure 2. Fictional example of a Mendelian randomization analysis with 10 genetic variants—six valid instrumental variables (hollow circles) and four invalid instrumental variables (solid circles) for finite sample size (left) and infinite sample size (right) showing IVW (solid line) and simple median (dashed line) estimates compared with the true causal effect (dotted line). The ratio estimate for each genetic variant is the gradient of the line connecting the relevant datapoint for that variant to the origin; the simple median estimate is the median of these ratio estimates.

Weaker assumption IV3 (exclusion restriction assumption), Bowden et al., 2015

- Direct causal effect $SNP_j \rightarrow Y$ (Assumption IV3) is not 0 for all Ivs
- Assume Instrument Strength Independent of Direct Effect (InSIDE)
  - The distributions of $\beta_{indirect,j}$ and $\beta_{causal,j}$ are independent
- Egger regression: Over causal effect of $\mathrm{SNP}_{j=1,\ldots,J} \rightarrow M \rightarrow Y$ can be estimated by the linear regression slope of $\hat{\beta}_{X_jY} \sim \hat{\beta}_{X_jM}, j = 1, \ldots, J$

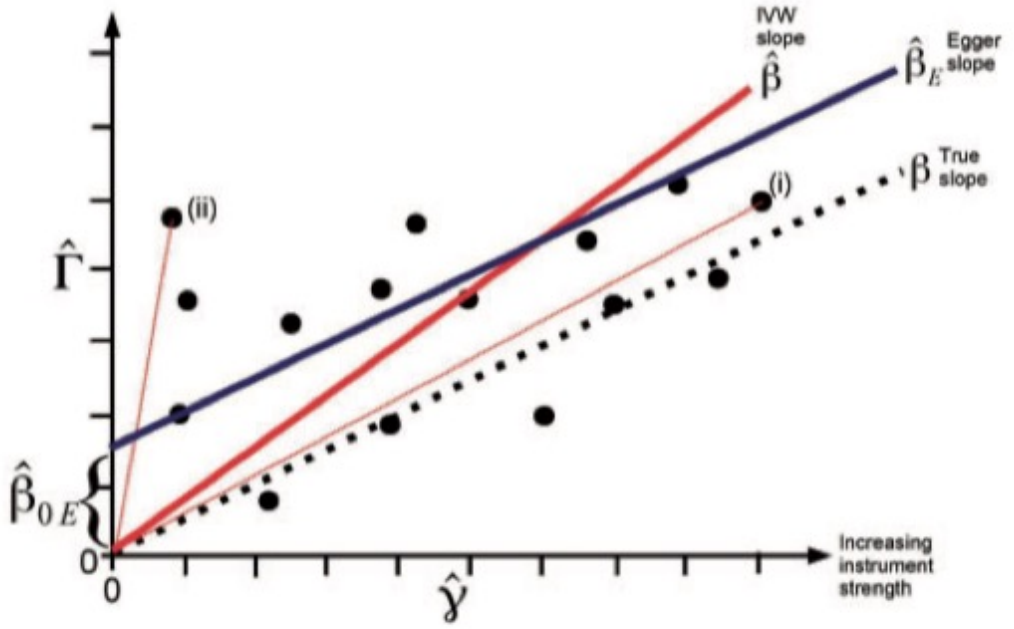


Figure 2. Plot of the gene–outcome ($\hat{\Gamma}$) vs gene–exposure ($\hat{\gamma}$) regression coefficients for a fictional Mendelian randomization analysis with 15 genetic variants. The true slope is shown by a dotted line, the inverse-variance weighted (IVW) estimate by a red line, and the MR-Egger regression estimate by a blue line. Refer to text for explanation of points (i) and (ii).

Example MR Studies of Blood Pressure and Coronary Artery Disease

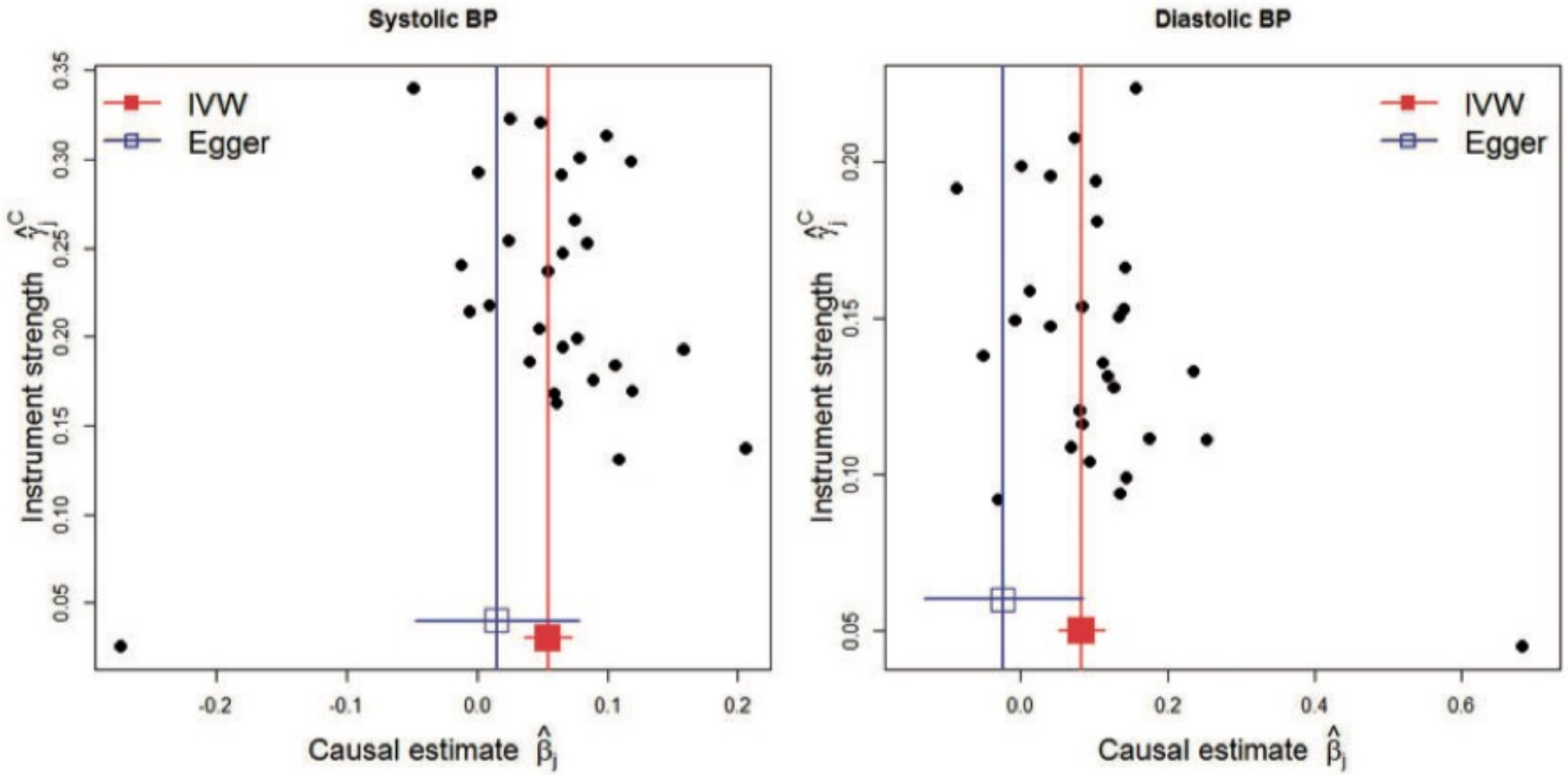
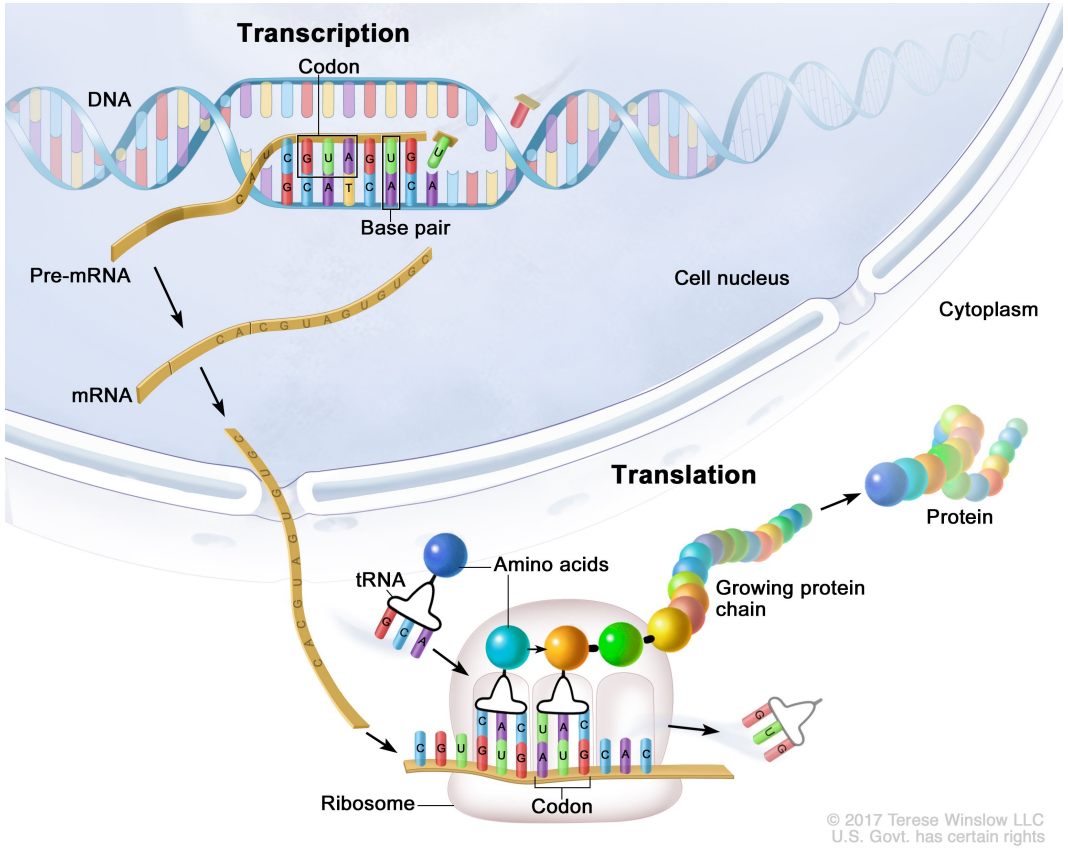


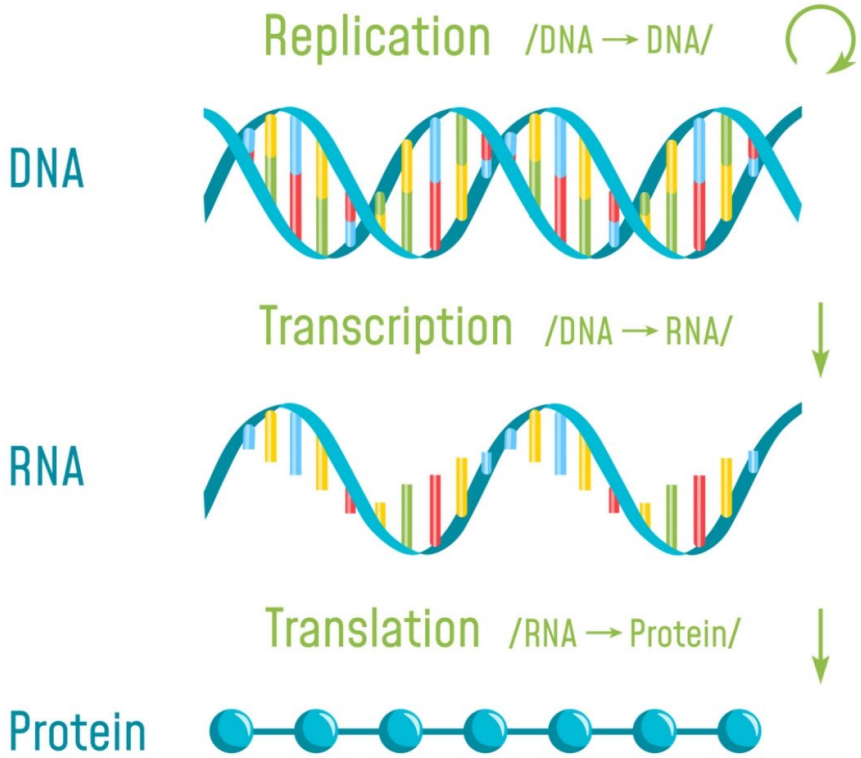
Figure 4. Genetic associations with blood pressure and coronary artery disease risk from 29 variants—funnel plots of minor allele frequency corrected genetic associations with blood pressure ($\hat{\gamma}_j^C$) against causal estimates of blood pressure on CAD based on each genetic variant individually ($\hat{\beta}_j$). Left: funnel plot for systolic blood pressure. Right: funnel plot for diastolic blood pressure. The inverse-variance weighted (IVW) and MR-Egger causal effect estimates are also shown.

Study multi-omics data by MR methods

- Test genetic-phenotype causality mediated through epigenome (DNA methylation, histone markers), transcriptome (gene expressions), proteome (protein abundance)
- Select IV SNPs
  - GWAS
  - Molecular Quantitative Trait Loci (e.g., eQTL)
- Apply MR methods
  - MR-Egger, Bowden et. al., 2015
  - SMR, Zhu et. al., 2016
  - GSMR, Zhu et. al., 2018

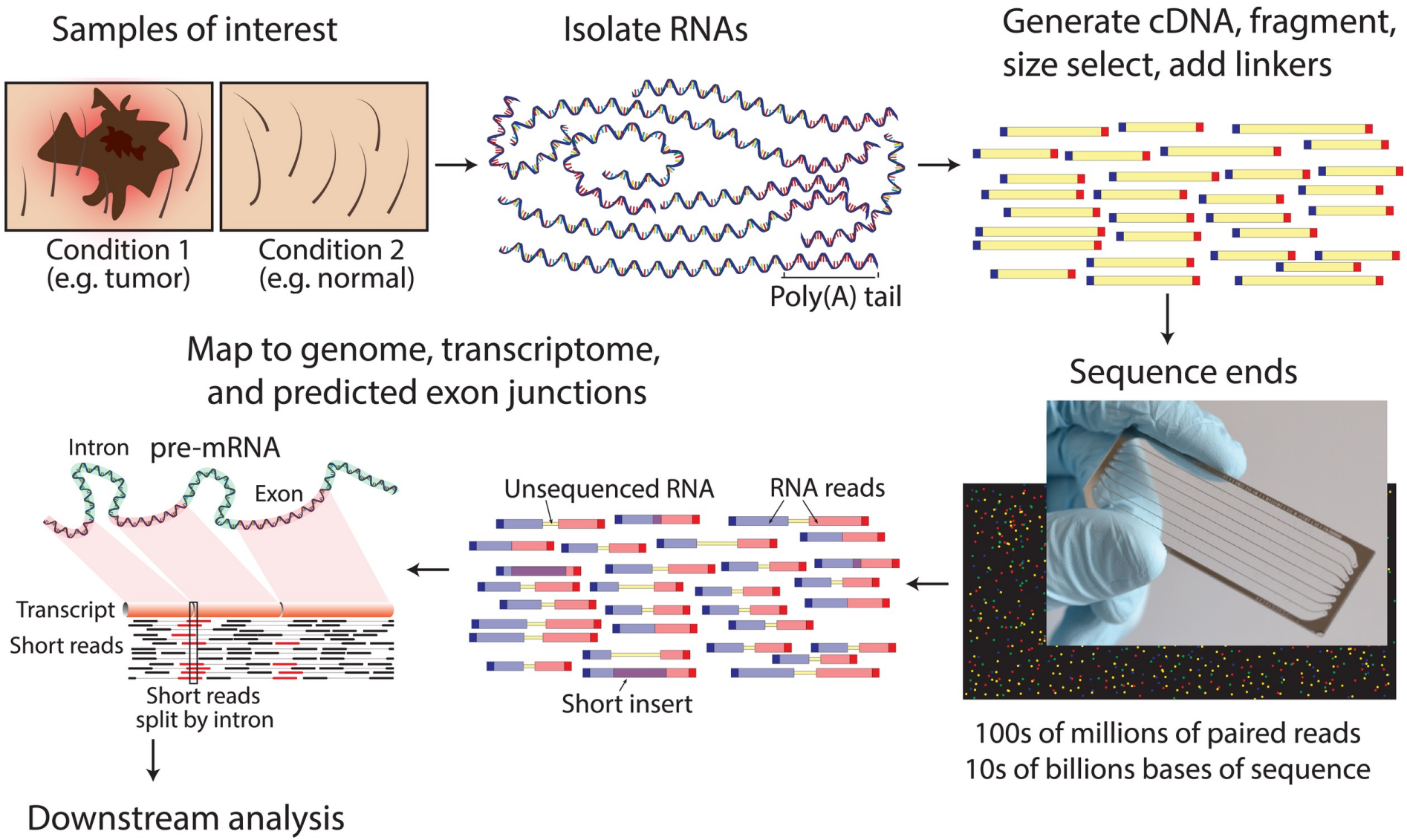
Transcription





https://www.thoughtco.com/dna-transcription-373398

Profile Gene Expression Levels by RNA-sequencing



eQTL

- Consider the profiled gene expression levels as the quantitative trait Y in the following single variant linear regression model:

$$Y = \beta_0 + \alpha C + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

  - X represents the genotype data (0, 1, 2) or dosage [0, 2] of the test SNP
  - C represents the confounding covariates or other environmental variables
  - ϵ represents the error term, other unknown factors
- eQTL :SNPs significantly associated with a gene expression quantitative trait ($H_0: \beta_1 = 0$ is significantly rejected) are referred as **e**xpression **Q**uantitative **T**rait **L**oci
- Cis-eQTL : nearby the test gene (e.g., located within the +-1MB region of the transcription starting site). Need to screen thousands SNPs per gene.
- Trans-eQTL : distant from the test gene (e.g., located out of the +-1MB region of the transcription starting site, or on different chromosome). Need to screen ~10M SNPs per gene.

Summary data-based Mendelian Randomization (SMR) Zhu et. al., 2016.

- Consider SNP genotype/IV z, phenotype/response y, mediator/gene expression x
- Study the causal effects from SNPs / IVs -> Gene Expression / Mediator -> Disease/Outcome

$$b_{xy} = b_{zy}/b_{zx}$$

- Use summary-level GWAS and eQTL data (p-values, effect sizes, effect size variances, sample sizes, minor allele frequencies)

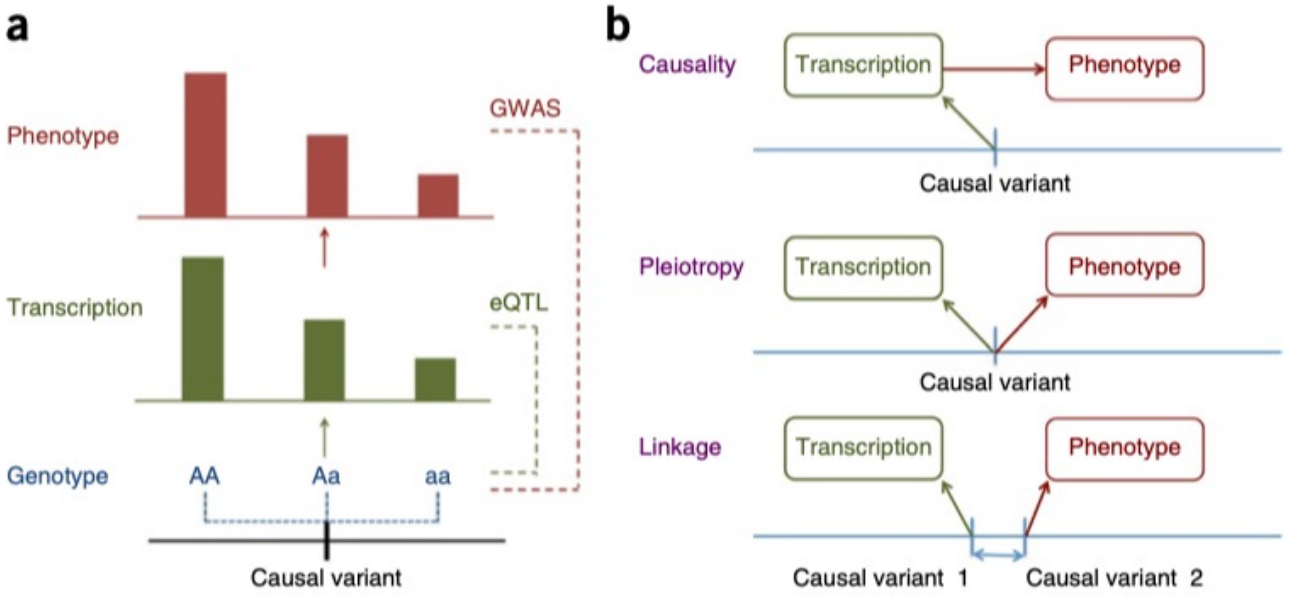


Figure 1 Association between gene expression and phenotype through genotypes. (**a**) A model of causality where a difference in phenotype is caused by a difference in genotype mediated by gene expression (transcription). (**b**) Three possible explanations for an observed association between a trait and gene expression through genotypes.

Available Tools

- R library “SKAT”: Rare variant test
  https://cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf
- PheWeb : Visualize Biobank GWAS results of multiple phenotypes, search phenotypes, genes, SNPs
  https://pheweb.org/
- Test pleiotropy effect : plink.multivariate
  https://genepi.qimr.edu.au/staff/manuelF/multivariate/main.html
- GSMR : Generalized Summary-data-based Mendelian Randomization
  https://yanglab.westlake.edu.cn/software/gsmr/

Lecture on 03/01

- Transcriptome-wide Association Studies (TWAS)
  - PrediXcan based on Elastic-Net model (Gamazon et. al. Nature Genetics, 2015).
  - TIGAR based on Bayesian Dirichlet Process Regression model (Nagpal et. al, AJHG 2019).
  - PMR-Egger based on a MR likelihood framework that unifies existing TWAS and MR methods (Yuan et. al, Nature Communications, 2020).