Scalable Statistical Inference for Genome-Wide Association Studies

James J. Yang
University of Michigan
GitHub: jyangstat

April 26, 2019

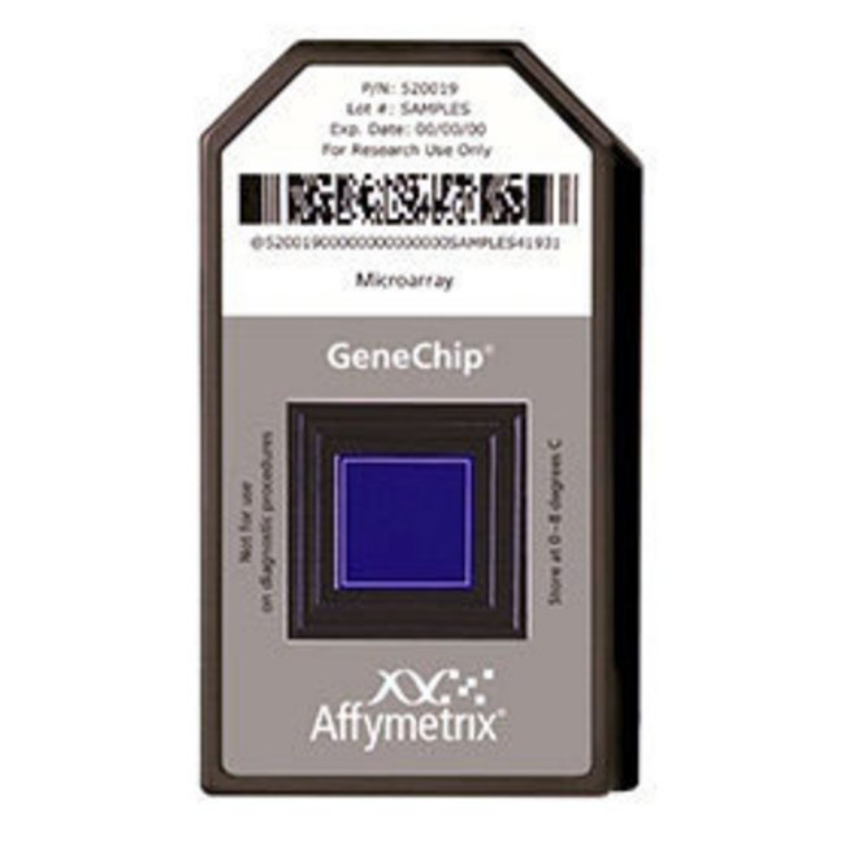
Genome-Wide Association Studies

What are Genome-Wide Association Studies (GWAS)

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

–National Human Genome Research Institute, NIH

Affymetrix GeneChip



Genome-Wide Human SNP Array 6.0: 1.8 million markers (946K probes for copy number variants and 906.6K SNPs)

Genotype Data

- For a SNP with two alleles coded as A and B, there are three possible genotypes: $\{AA, AB, BB\}$.
- Additive (B as reference): $\{AA = 0, AB = 1, BB = 2\}$

Genotype Data: Z_{ij} = number of the reference allele

	ID							
SNP	1	2	3	4	$\cdots$	$n-2$	$n-1$	n
1	0	1	2	2	$\cdots$	2	1	1
2	2	2	0	-1	$\cdots$	2	0	1
3	0	1	2	0	$\cdots$	2	2	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$G-1$	2	0	0	1	$\cdots$	2	-1	2
G	0	0	2	1	$\cdots$	1	2	1

Phenotype Data with Genotype

Phenotype Data:

Y_1	7.32	3.09	5.24	4.57	$\cdots$	7.12	3.96	4.72
Y_2	1	0	1	1	$\cdots$	0	0	0
Y_3	3	5	2	1	$\cdots$	4	1	2

Genotype Data: Z_{ij} = number of the reference allele

	ID							
SNP	1	2	3	4	$\cdots$	$n-2$	$n-1$	n
1	0	1	2	2	$\cdots$	2	1	1
2	2	2	0	-1	$\cdots$	2	0	1
3	0	1	2	0	$\cdots$	2	2	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$G-1$	2	0	0	1	$\cdots$	2	-1	2
G	0	0	2	1	$\cdots$	1	2	1

Issues with GWAS

Confounders: Significant association between SNP and phenotype in GWAS may arise from

- Population Stratification: the presence of a systemic difference in allele frequencies between subpopulations in a population.
- Genetic Admixture: individuals are composed of a mixed ancestry.
- Cryptic Relatedness: the presence of close relatives in a sample of seeming unrelated individuals.

Pleiotropic Genes

Pleiotropy

Pleiotropy refers to the condition where a single mutation causes more than one observable phenotypic effect or change in characteristic.

Multivariate Phenotypes in GWAS

- Hypertension (SBP, DBP).
- Diabetes mellitus (glucose level, HbA_{1C}).
- Substance abuse (nicotine dependence, alcohol dependence).
- Arabidopsis thaliana (107 phenotypes).
  Atwell *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627-31.

Challenges in Identifying Pleiotropic Genes

1. Confounders (Population stratification, Genetic admixture. Cryptic relatedness).
2. The number of phenotypic variables varies.
3. Goodness of fit and power of the models.
4. Computation efficiency.
5. Interpretations of statistical findings.

Both Variables (Y_1 and Y_2) are Continuous

- Let (Y_1, Y_2) be the bivariate continuous response variables
- For a given SNP,

$$Y_1 \text{ and SNP } \Rightarrow p_1$$

$$Y_2 \text{ and SNP } \Rightarrow p_2$$

Both Variables (Y_1 and Y_2) are Continuous

- Let (Y_1, Y_2) be the bivariate continuous response variables
- For a given SNP,

$$Y_1 \text{ and SNP } \Rightarrow p_1$$

$$Y_2 \text{ and SNP } \Rightarrow p_2$$

- Consider a continuous, non-increasing combination function: $(0,1) \times (0,1) \to R$.
- Using Fisher combination function:

$$T = -2\log(p_1) - 2\log(p_2)$$

- The p-values of T can be calculated using (time-consuming) permutation method.

The Distribution of T under H_0

- Under the null hypothesis of no association:

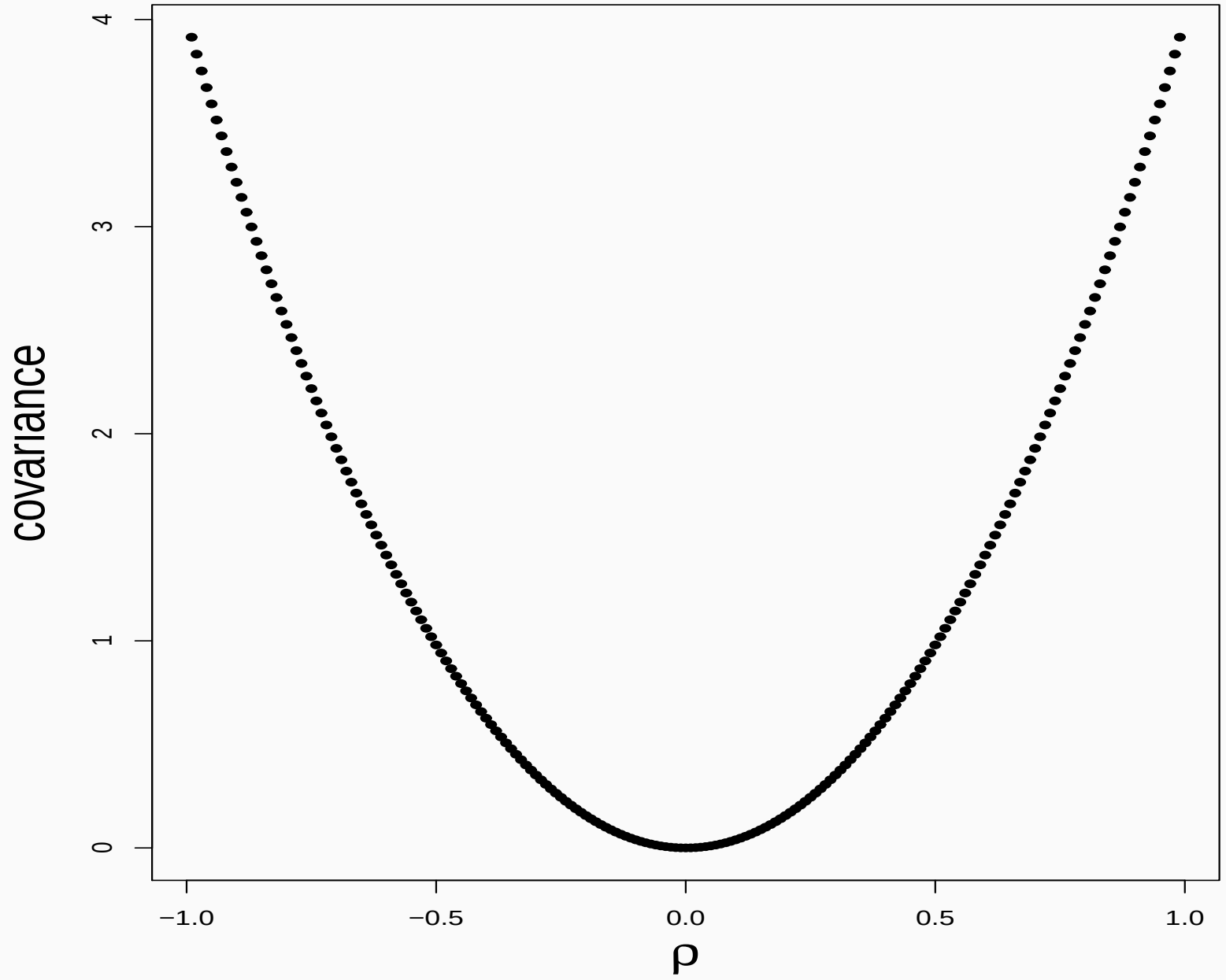
$$T = -2\log(p_1) - 2\log(p_2) \sim \text{Gamma}(k, s)$$

- RHS: $E[T] = ks$ and $Var[T] = ks^2$
- LHS: First two moments of T:

$$E[T] = 4$$
$$Var[T] = 8 + 2\,\text{cov}\left\{-2\log(p_1), -2\log(p_2)\right\}$$

1. $cov\left\{-2\log(p_1), -2\log(p_2)\right\}$ is a function of $\rho = \rho(Y_1, Y_2)$
2. Numerical integration for a given ρ:

$$cov\left\{-2\log(p_1), -2\log(p_2)\right\} =$$
$$4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\log\left\{2\Phi(-|z_1|)\right\}\log\left\{2\Phi(-|z_2|)\right\}dF(z_1, z_2) - 4$$


covariance
0
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
ρ

- $\text{cov}\{-2\log(p_1), -2\log(p_2)\} \doteq$
$$c_1\rho_{1,2}^2 + c_2\rho_{1,2}^4 + c_3\rho_{1,2}^6 + c_4\rho_{1,2}^8 + c_5\rho_{1,2}^{10}$$
where
$$\begin{aligned} c_1 &= 3.9081 \\ c_2 &= 0.0313 \\ c_3 &= 0.1022 \\ c_4 &= -0.1378 \\ c_5 &= 0.0941 \end{aligned}$$
- Maximum residual ≤ 0.0001
- Therefore,
$$\begin{aligned} E[T] &= 4 \\ Var[T] &\doteq 8 + 2\left\{c_1\rho_{1,2}^2 + c_2\rho_{1,2}^4 + c_3\rho_{1,2}^6 + c_4\rho_{1,2}^8 + c_5\rho_{1,2}^{10}\right\} \end{aligned}$$

Summary of the Proposed Procedure

Fisher Combination Method

1. Calculate marginal p-values for all SNPs.
2. Estimate $\rho_{1,2}$ from phenotypes.
3. Using the values of $\hat{\rho}_{1,2}$ to calculate the values of k and s.
4. The p-values of $T\ (= -2\log(p_1) - 2\log(p_2))$ is derived from Gamma$(\hat{k}, \hat{s})$.

Summary of the Proposed Procedure

Fisher Combination Method

1. Calculate marginal p-values for all SNPs.
2. Estimate $\rho_{1,2}$ from phenotypes.
3. Using the values of $\hat{\rho}_{1,2}$ to calculate the values of k and s.
4. The p-values of $T\ (= -2\log(p_1) - 2\log(p_2))$ is derived from Gamma$(\hat{k}, \hat{s})$.

- For multivariate phenotypes $Y_1, \ldots, Y_m$ $(m \geq 2)$:

$$T = \sum_{j=1}^{m} -2\log(p_j)$$

Mixed Continuous (Y_1) and Binary (Y_2)

- For a given SNP,

$$Y_1 \text{ and SNP } \Rightarrow p_1$$

$$Y_2 \text{ and SNP } \Rightarrow p_2$$

$$T_{12} = -2\log(p_1) - 2\log(p_2)$$

Mixed Continuous (Y_1) and Binary (Y_2)

- For a given SNP,

$$Y_1 \text{ and SNP } \Rightarrow p_1$$

$$Y_2 \text{ and SNP } \Rightarrow p_2$$

$$T_{12} = -2\log(p_1) - 2\log(p_2)$$

- Let (Y_1, W) be the bivariate continuous response variables, where W is a latent variable,

$$Y_2 = \begin{cases} 1 & \text{if } W \geq C, \\ 0 & \text{if } W < C. \end{cases}$$

- For a given SNP,

$$W \text{ and SNP } \Rightarrow p_w$$

$$T_{1w} = -2\log(p_1) - 2\log(p_w)$$

Both continuous Y_1 and W

1. p-values: p_{y_1} and p_w
2. Statistic: $T_{1w} = -2\log(p_{y_1}) - 2\log(p_w)$
3. $E[T_{1w}] = 4$
4. $Var[T_{1w}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_w)\}$
5. $Var[T_{1w}] \doteq 8 + 2\sum_{j=1}^{5} c_j \rho_{y_1 y_w}^{2j}$

Continuous Y_1 and Binary Y_2

1. p-values: p_{y_1} and p_{y_2}
2. Statistic: $T_{12} = -2\log(p_{y_1}) - 2\log(p_{y_2})$
3. $E[T_{12}] = 4$
4. $Var[T_{12}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_{y_2})\}$

Both continuous Y_1 and W

1. p-values: p_{y_1} and p_w
2. Statistic: $T_{1w} = -2\log(p_{y_1}) - 2\log(p_w)$
3. $E[T_{1w}] = 4$
4. $Var[T_{1w}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_w)\}$
5. $Var[T_{1w}] \doteq 8 + 2\sum_{j=1}^{5} c_j \rho_{y_1 y_w}^{2j}$

Continuous Y_1 and Binary Y_2

1. p-values: p_{y_1} and p_{y_2}
2. Statistic: $T_{12} = -2\log(p_{y_1}) - 2\log(p_{y_2})$
3. $E[T_{12}] = 4$
4. $Var[T_{12}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_{y_2})\}$

$$\text{cov}\{-2\log(p_{y_1}), -2\log(p_w)\} \geq \text{cov}\{-2\log(p_{y_1}), -2\log(p_{y_2})\}$$

Both continuous Y_1 and W

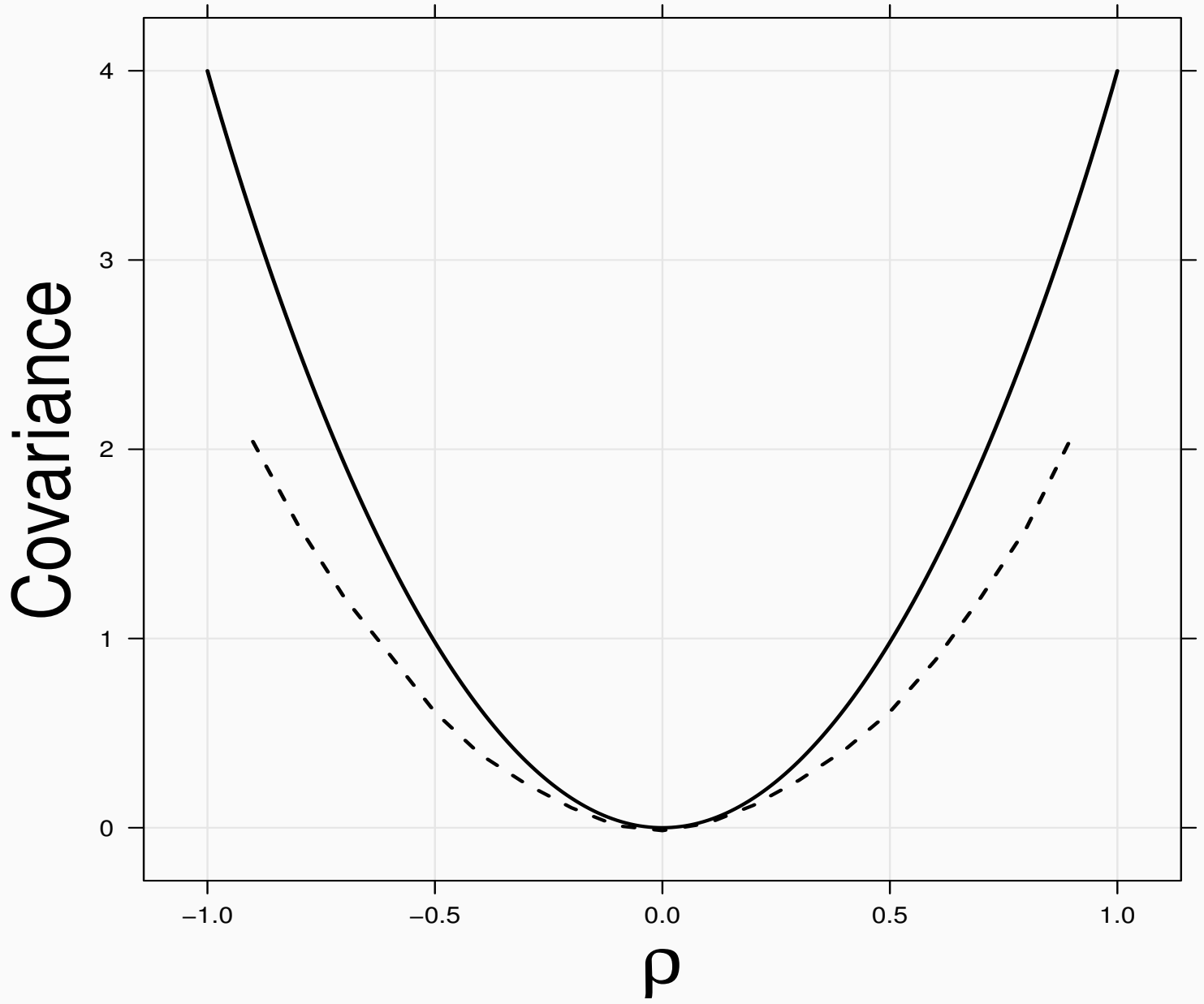
1. p-values: p_{y_1} and p_w
2. Statistic: $T_{1w} = -2\log(p_{y_1}) - 2\log(p_w)$
3. $E[T_{1w}] = 4$
4. $Var[T_{1w}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_w)\}$
5. $Var[T_{1w}] \doteq 8 + 2\sum_{j=1}^{5} c_j \rho_{y_1 y_w}^{2j}$

Continuous Y_1 and Binary Y_2

1. p-values: p_{y_1} and p_{y_2}
2. Statistic: $T_{12} = -2\log(p_{y_1}) - 2\log(p_{y_2})$
3. $E[T_{12}] = 4$
4. $Var[T_{12}] = 8 + 2\,\text{cov}\{-2\log(p_{y_1}), -2\log(p_{y_2})\}$

$$\text{cov}\{-2\log(p_{y_1}), -2\log(p_w)\} \geq \text{cov}\{-2\log(p_{y_1}), -2\log(p_{y_2})\}$$

Pearson's Biserial correlation: $\hat{\rho} = \dfrac{\hat{r}_{y_1 y_2} S_{y_2}}{\phi(\Phi^{-1}(1-\bar{y}_2))} (\to \rho_{y_1 y_w})$


Covariance
4
3
2
1
0
−1.0
−0.5
0.0
0.5
1.0
ρ

Mixed Measurement Scales (Continuous, Binary, and Ordinal)

- Let Y_k be ordinal response variables ($k = 1, 2$).
- W_k is a latent variable,

$$Y_k = \begin{cases} 1 & \text{if } -\infty \leq W_k < C_1, \\ 2 & \text{if } C_1 \leq W_k < C_2, \\ & \vdots \\ R & \text{if } C_{R-1} \leq W_k < \infty, \end{cases}$$

- For a given SNP,

$$Y_1 \text{ and SNP } \Rightarrow p_1$$

$$Y_2 \text{ and SNP } \Rightarrow p_2$$

- Using Fisher combination function:

$$T = -2\log(p_1) - 2\log(p_2)$$

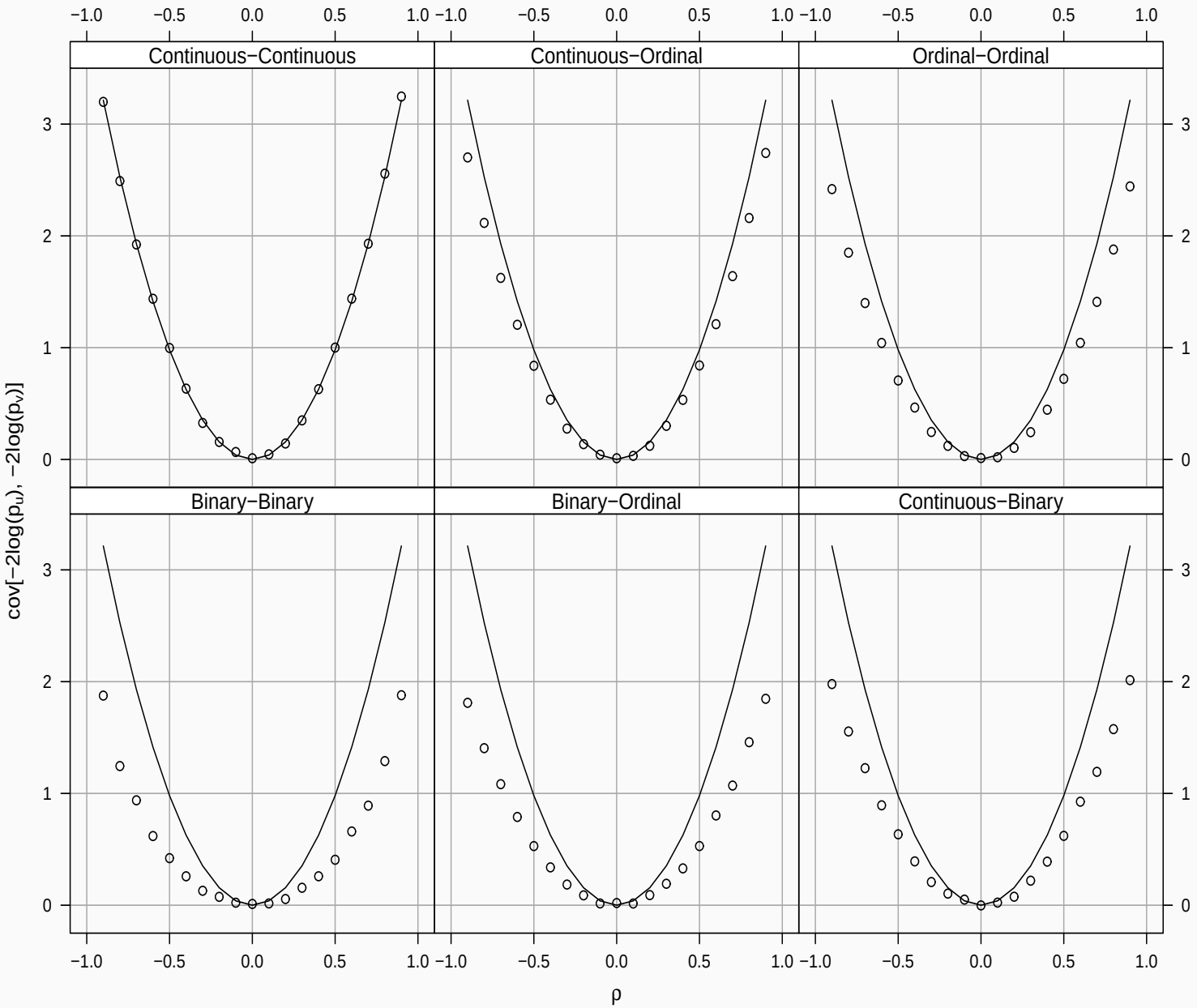
Mixed Measurement Scales (Continuous, Binary, and Ordinal)

- Based on variable types:

		Y_1		
		Continuous	Binary	Ordinal
Y_2	Continuous	Kendall	Biserial	Polyserial
	Binary		Tetrachoric	Polychoric
	Ordinal			Polychoric

- Kendall's τ correlation:

$$\rho = \sin\left(\frac{\pi\tau}{2}\right)$$


Continuous−Continuous
Continuous−Ordinal
Ordinal−Ordinal
Binary−Binary
Binary−Ordinal
Continuous−Binary
cov[−2log(p_u), −2log(p_v)]
ρ
−1.0
−0.5
0.0
0.5
1.0
0
1
2
3

Simulation Study: Multivariate Outcomes

- Latent phenotypes:

$$\begin{pmatrix} W_1 \\ \vdots \\ W_6 \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_6 \end{pmatrix}, \begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}\right)$$

- Effect Size:

$$\mu_i = \begin{cases} -e_i & \text{if } Z = 0 \\ 0 & \text{if } Z = 1 \\ e_i & \text{if } Z = 2 \end{cases}$$

for $i = 1, \ldots, 6$.

- Observed Phenotypes:

$$
\begin{aligned}
Y_1 &= W_1; \\
Y_2 &= W_2; \\
Y_3 &= I(W_3 \leq C_3); \qquad Y_3 = 0 \text{ or } 1 \\
Y_4 &= I(W_4 \leq C_4); \qquad Y_4 = 0 \text{ or } 1 \\
Y_5 &= I(\xi_{m-1} \leq W_5 < \xi_m); \qquad Y_5 = 1, \ldots, 5 \\
Y_6 &= I(\zeta_{m-1} \leq W_6 < \zeta_m); \qquad Y_6 = 1, \ldots, 5
\end{aligned}
$$

ρ	e_1/e_2	e_3/e_4	e_5/e_6	Latent	Mixed	Binary	Continuous
0	0	0	0	0.00009	0.00008	0.00004	0.00009
0.35	0	0	0	0.00057	0.00024	0.00005	0.00025
0.75	0	0	0	0.00027	0.00007	0.00002	0.00012
0	0.5	0.5	0.5	0.96	0.91	0.77	0.91
0.35	0.5	0.5	0.5	0.84	0.73	0.49	0.73
0.75	0.5	0.5	0.5	0.51	0.36	0.19	0.39
0	0/0.7	0/0.7	0/0.7	0.94	0.88	0.68	0.88
0.35	0/0.7	0/0.7	0/0.7	0.82	0.68	0.42	0.70
0.75	0/0.7	0/0.7	0/0.7	0.36	0.20	0.08	0.25
0	0.9	0	0	0.94	0.94	0.67	0.94
0.35	0.9	0	0	0.84	0.84	0.42	0.83
0.75	0.9	0	0	0.37	0.37	0.05	0.38

SAGE Data

- SAGE: The Study of Addiction: Genetics and Environment
  - COGA: The Collaborative Study on the Genetics of Alcoholism
  - FSCD: The Family Study of Cocaine Dependence
  - COGEND: The Collaborative Genetic Study of Nicotine Dependence
- Unrelated individuals: 2,775 (1,288 male, 1,487 female)
- Total number of SNPs (pass QC): 753,238

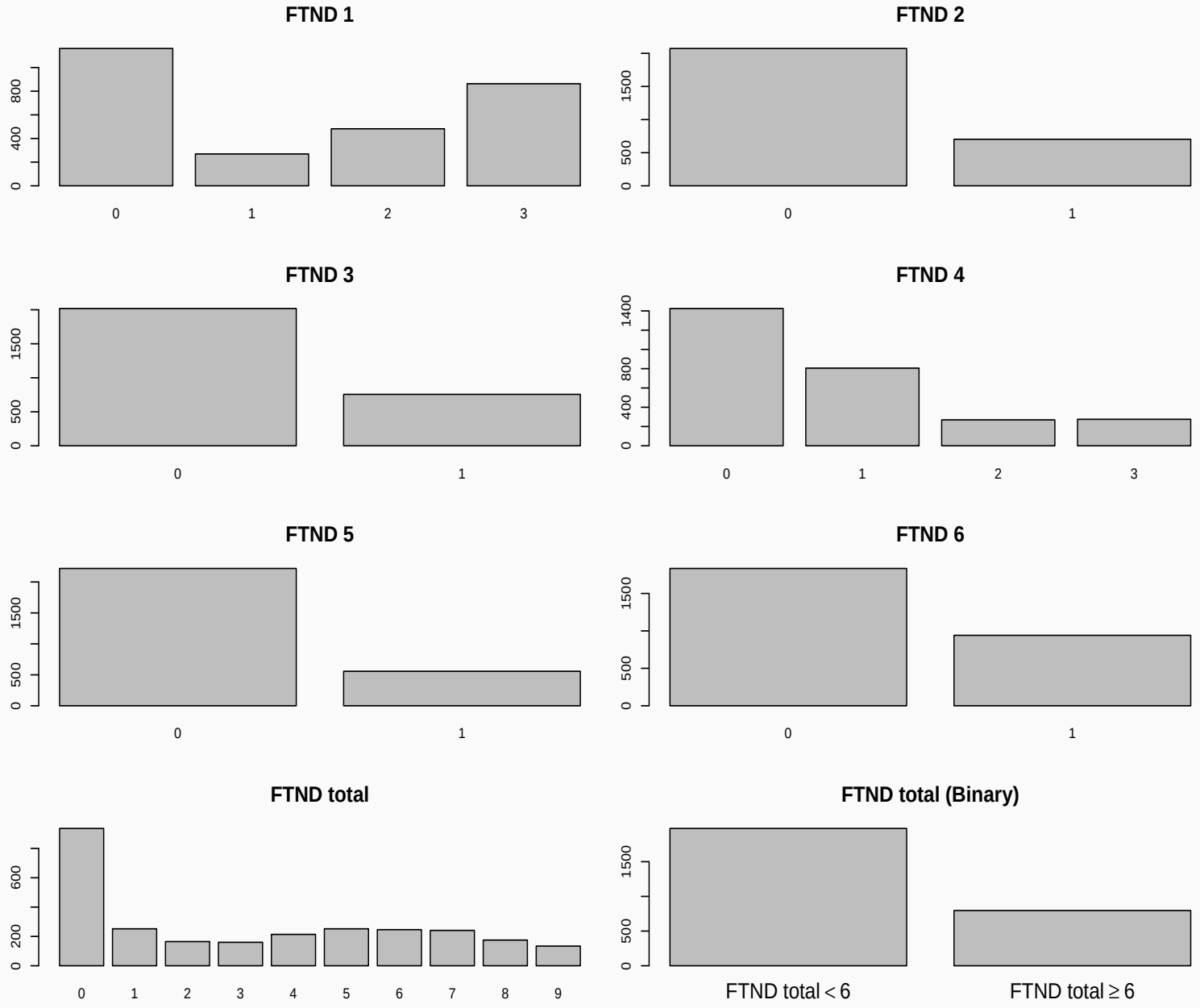
SAGE Data

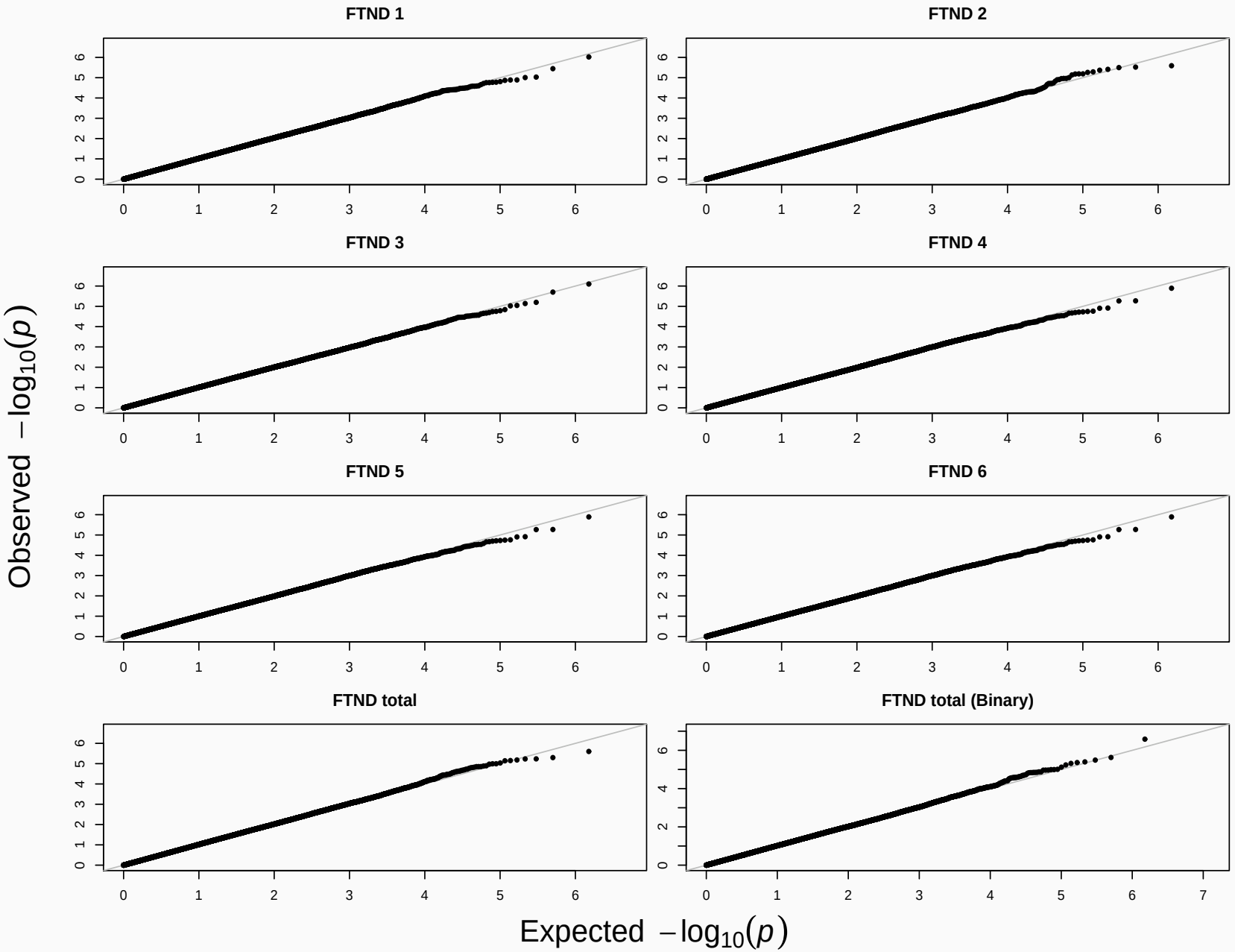
Phenotypes: Participant's lifetime score for Fagerström Test for Nicotine Dependence (FTND) questions:

- `ftnd_1`: How soon after you wake up do you smoke your first cigarette?
  ($0 =$ after 60 min; $1 = 30 - 60$ min; $2 = 6 - 30$ min; $3 =$ within 5 min)
- `ftnd_2`: Do you find it difficult to refrain from smoking in places where it was forbidden?
  ($1 =$ Yes; $0 =$ No)
- `ftnd_3`: Which cigarette would you hate most to give up?
  ($1 =$ First one in morning; $0 =$ All others)

- `ftnd_4`: How many cigarettes per day do you smoke?
  ($3 = 31$ or more; $2 = 21 - 30$; $1 = 11 - 20$; $0 = 10$ or fewer)
- `ftnd_5`: Do you smoke more frequently during the first hours after waking than during the rest of the day?
  ($1 =$ Yes; $0 =$ No)
- `ftnd_6`: Do you smoke if you are so ill that you are in bed most of the day?
  ($1 =$ Yes; $0 =$ No)
- `ftnd_total`: Sum of points scored for `ftnd_1` through `ftnd_6`
  (range $= 0 - 10$)
- `ftnd_total_binary`: `ftnd_total` < 6 or not
  ($1 =$ `ftnd_total`: $6 - 10$; $0 =$ `ftnd_total`: $0 - 5$)

Correlation	ftnd_2	ftnd_3	ftnd_4	ftnd_5	ftnd_6
ftnd_1	0.6815	0.6758	0.7528	0.6446	0.7770
ftnd_2		0.4579	0.5895	0.4403	0.6838
ftnd_3			0.4822	0.6394	0.5294
ftnd_4				0.4452	0.6702
ftnd_5					0.5110


FTND 1
0
400
800
0
1
2
3
FTND 2
0
500
1500
0
1
FTND 3
0
500
1500
0
1
FTND 4
0
400
800
1400
0
1
2
3
FTND 5
0
500
1500
0
1
FTND 6
0
500
1500
0
1
FTND total
0
200
600
0
1
2
3
4
5
6
7
8
9
FTND total (Binary)
0
500
1500
FTND total < 6
FTND total ≥ 6


FTND 1
FTND 2
FTND 3
FTND 4
FTND 5
FTND 6
FTND total
FTND total (Binary)
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

Summary of Analysis Results

- Type I error rate $= 10^{-6}$

  `ftnd_1` rs821722 ($p = 9.54 \times 10^{-7}$)

  `ftnd_3` rs3138134 ($p = 7.94 \times 10^{-7}$)

- No significant finding for `ftnd_2`, `ftnd_4`, `ftnd_5`, `ftnd_6`, and `ftnd_total`.
- Our proposed combination method with 6 phenotypes: 9 SNPs
  (rs17538699, rs17798885, rs2245261, rs4077464, rs4658846, rs4658847, rs6553017, rs7672047, rs944582)

Conclusions

- We can add baseline demographic variables and principal components in the regression model to control for potential confounders.
- It is much easy to evaluate goodness-of-fit in the marginal models.
- The computation efficiency of the proposed method is proportional to that of marginal test.
- Once we have marginal p-values, the multivariate tests are extremely efficient.

Permutation Tests

Permutation Methods

- Advantage: The procedure is very general that it is applicable to most test statistics.
- Disadvantage: The procedure is time consuming to conduct exact tests or to estimate very small p-values.

Permutation Methods

- Null hypothesis: $H_0 : F_1 = F_2$
- Data: $\boldsymbol{z} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$
  $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1n_1})$ is a sample from F_1
  $\boldsymbol{x}_2 = (x_{21}, \ldots, x_{2n_2})$ is a sample from F_2.
- Under null H_0, we permute the index of $\boldsymbol{z}$ and results in $\boldsymbol{z}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$.
  Let the set $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_B^*\}$ contain all possible permutations of the observed data.
- The observed statistic: $T_0 = T(\boldsymbol{z})$.
- The statistics based on permuted data:
  $\Omega = \{T_1^* = T(\boldsymbol{z}_1^*), \ldots, T_B^* = T(\boldsymbol{z}_B^*)\}$
  where the size of the set Ω is $B = \binom{n_1+n_2}{n_1}$.

The p-value of the Permutation Test

- Exact test:

$$p_{\text{exact}} = \frac{\sum_{b=1}^{B} I(T_b^* \geq T_0) + 1}{B + 1}$$

$n_1 = n_2 = 30 \Rightarrow B = 1.18 \times 10^{17}$

- Monte Carlo approximation:

$$p_{\text{MC}} = \frac{\sum_{j=1}^{M} I(T_j^* \geq T_0) + 1}{M + 1}$$

where $\Omega_{\text{MC}} = \left\{ T_j^* : j \in (1, \ldots, M) \subset (1, \ldots, B) \right\}$.

- Normal approximation:

$$p_{\text{Normal}} = 2\Phi(-|T - \mu_t|/s_t)$$

where $\mu_t = \sum_{j=1}^{M} T_j^*/M$ and $s_t^2 = \sum_{j=1}^{M} (T_j^* - \mu_t)^2/(M-1)$

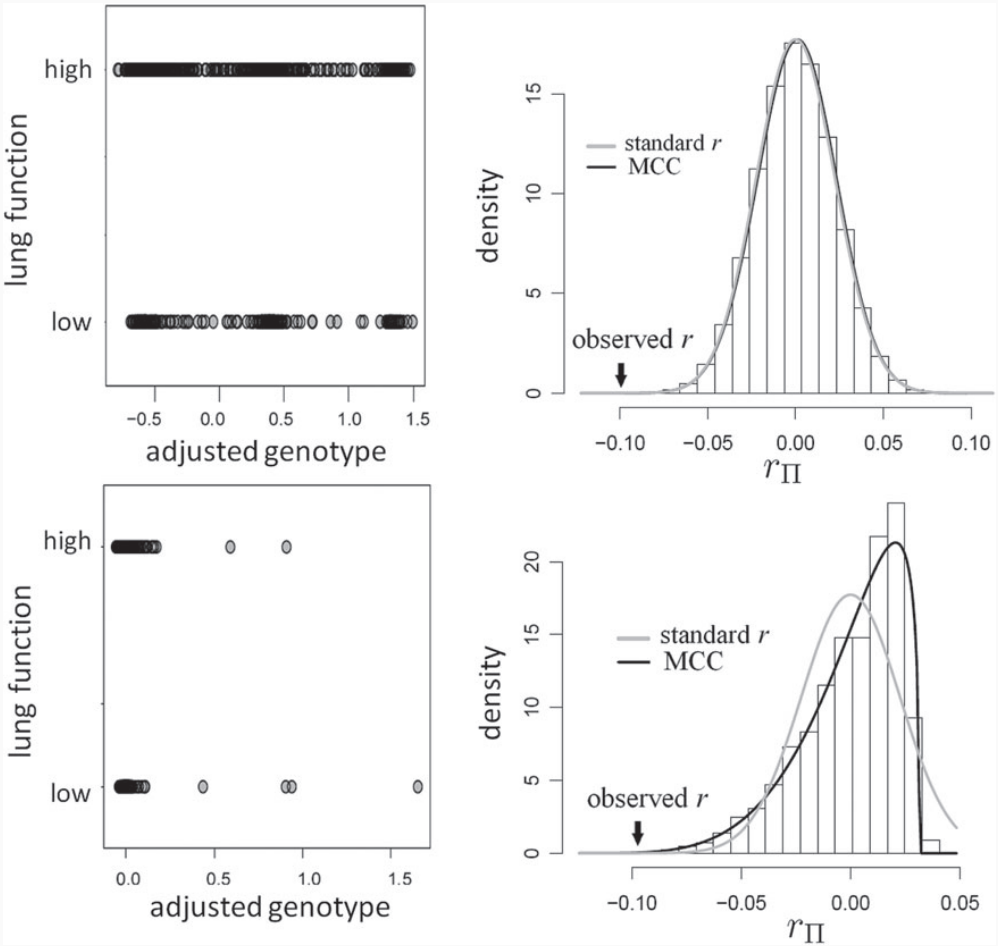


Fig. 1. MCC for genotype association testing. Upper left: data for SNP rs2956073. Although SNP genotypes were initially coded as 0, 1, 2, after covariate adjustment they appear as shown. Upper right: histogram of r_Π, with standard r and MCC fitted densities. Lower left: SNP rs180784621, with a low minor allele frequency producing considerable skew in the adjusted genotypes. Lower right: histogram of r_Π shows that MCC fits much better than standard r.

Zhou and Wright (2015). Hypothesis testing at the extremes: fast and robust association for high-throughput data. *Biostatistics*, 16(3), 611-625.

Sequential Monte Carlo Method

- In addition to $M =$ total number of permutations, define $K =$ maximum number of $(T_j^* > T_0)$
- Let r_h be the cumulative number of permuted statistics larger than T_0 at the h-th permutation ($h < M$). Besag and Clifford (1991):

$$p_{\text{seqMC}} = \begin{cases} \frac{K+1}{h+1} & \text{if } r_h = K \\ \frac{r_h+1}{M+1} & \text{if } r_h < K. \end{cases}$$

- $M = 1900$ and $K = 115$ for $\alpha = 0.05$ (Che et al. 2014).
- $M = 10^9$ and $K = 121$ for $\alpha = 5 \times 10^{-8}$.

Che et al (2014). An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. *BioData Mining*, 7, 9.

Besag and Clifford (1991) Sequential Monte Carlo p-Values. *Biometrika*, 78(2): 301-304.

Edgeworth Expansion Approximation

- Assume the *studentized/pivot* T_n is nonsingular and $E|T_n|^4 < \infty$.
- The Edgeworth expansion of the distribution of T_n

$$Pr[T_n \leq t] = \Phi(t) + \frac{1}{\sqrt{n}} q_1(t)\phi(t) + \frac{1}{n} q_2(t)\phi(t) + O(n^{-\frac{2}{3}})$$

  where $q_1(t) = \frac{m_3}{6} a_2(t)$,
  $q_2(t) = \frac{1}{36}\left\{3m_4 a_3(t) - 2m_3^2 a_5(t) - 9b_3(t)\right\}$, and
  $a_2(t), a_3(t), b_5(t)$ are polynomials of t.
- The first term is the Normal approximation.
- The skewness m_3 and kurtosis m_4 are estimated using permuted statistics in the sequential Monte Carlo step.

SAGE Data

- GWAS data from the Study of Addiction: Genetics and Environment (SAGE)
- Phenotypes indicative of risk for alcohol dependence:
  1. `age_first_drink`: the age when the participant had a drink containing alcohol for the first time.
  2. `ons_reg_drink`: the age of regular drinking onset (defined as drinking once a month for 6 months or more).
  3. `age_first_got_drunk`: the age when the participant got drunk for the first time.
  4. `alc_sx_tot`: the number of alcohol dependence symptoms endorsed.

Real Data Analysis

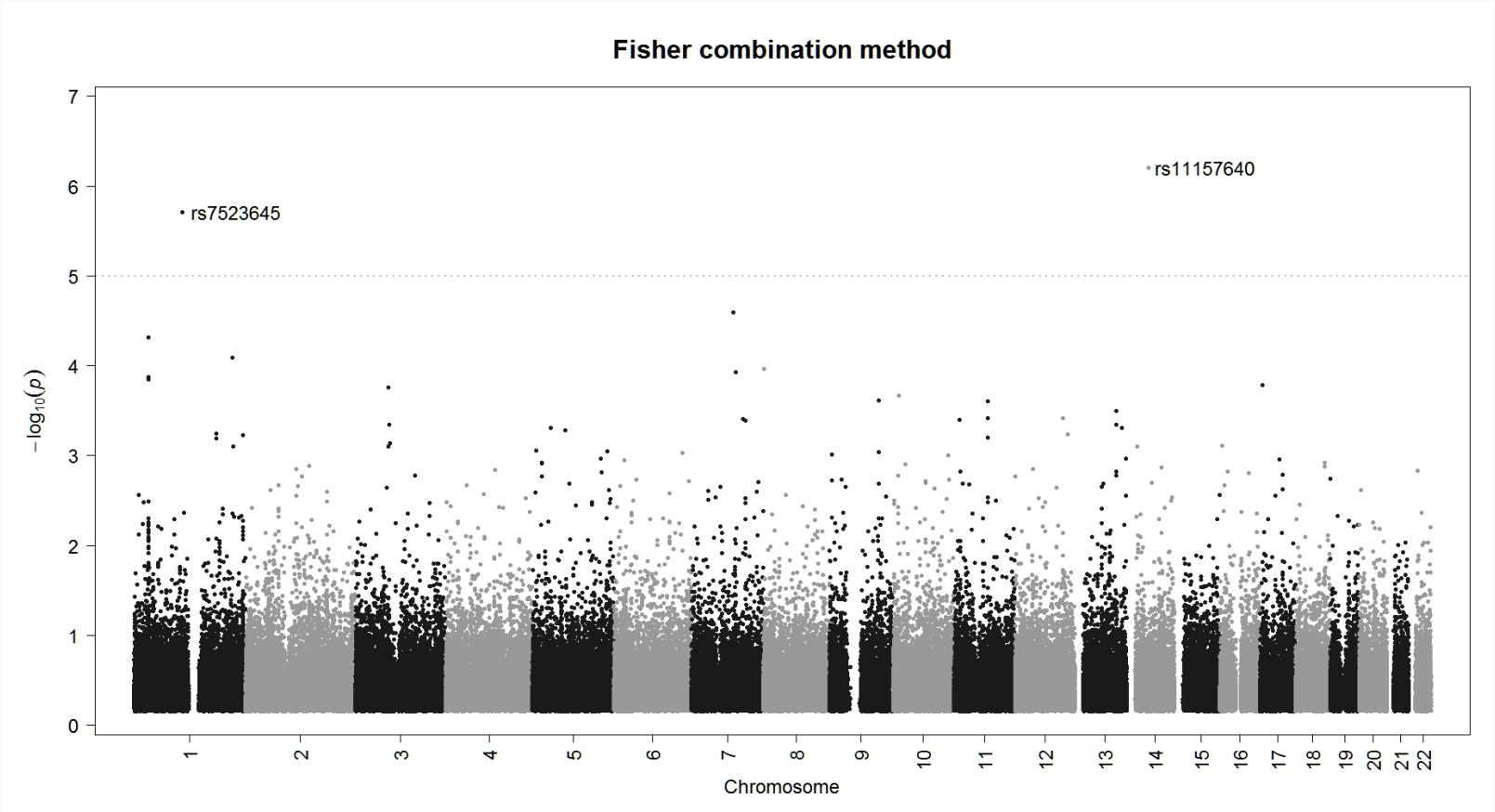
- Test statistics:

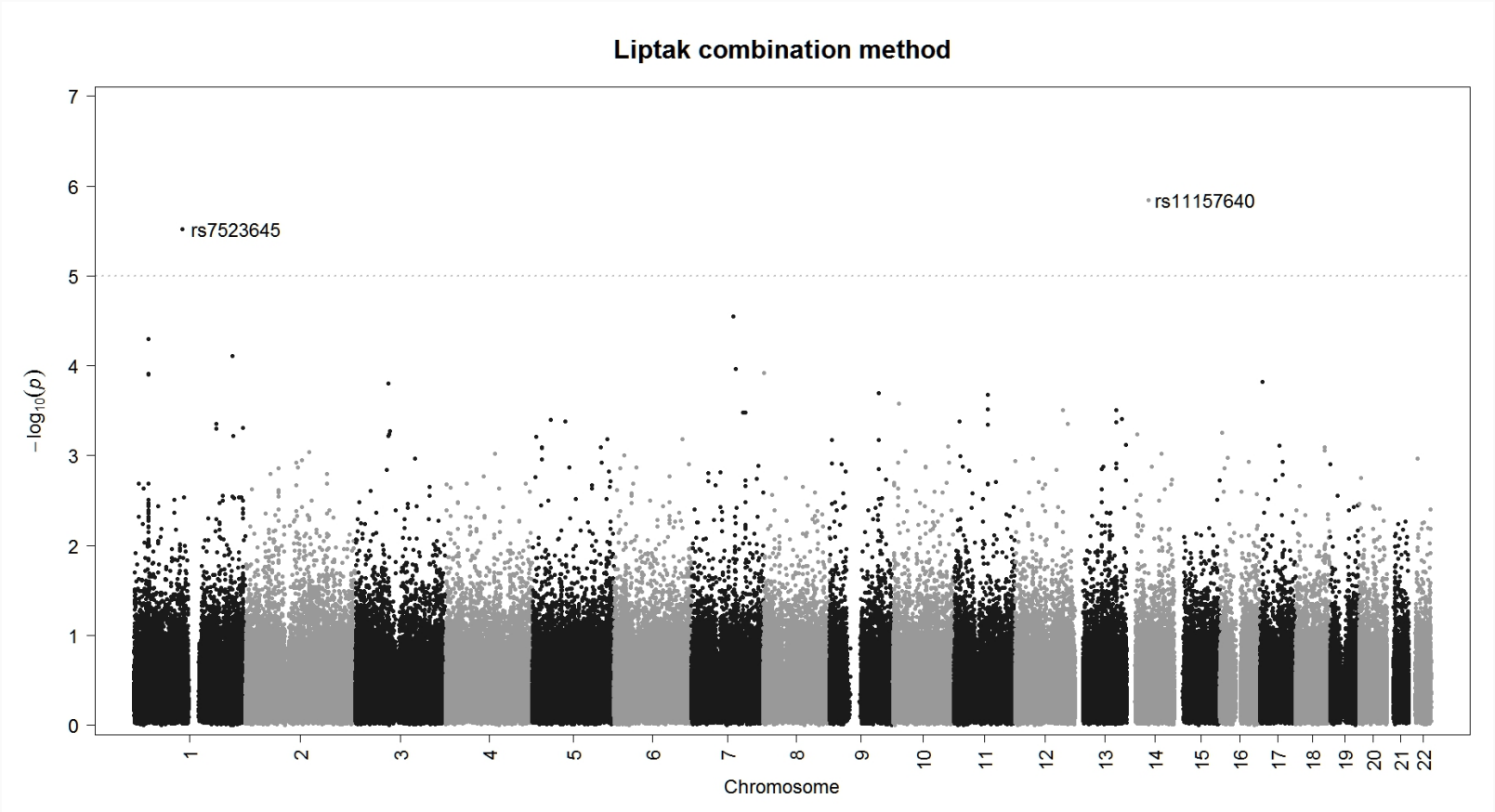
$$T_{\text{Fisher}} = \sum_{j=1}^{K} -2\log(p_j) \sim Gamma$$

$$T_{\text{Liptak}} = \sum_{j=1}^{K} \log\left(\frac{1-p_j}{p_j}\right) \sim \text{Unknown}$$

- Sequential Monte Carlo:

$$M = 2,000 \text{ and } K = 120$$


Fisher combination method
rs7523645
rs11157640
$-\log_{10}(p)$
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Liptak combination method
rs7523645
rs11157640
$-\log_{10}(p)$
Chromosome
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Conclusions

- The proposed method is applicable to a variety of test statistics.
- The sequential Monte Carlo method is unbiased and efficient for estimating moderate or large p-values.
- The Edgeworth expansion approximation can be used to estimate small or very small p-values.
- The studentized test statistic ensures that the first term $\Phi(\cdot)$ doest no depend on any moments. Therefore higher accuracy is reached.
- The Edgeworth expansion does not require the test statistic to be studentized. For non-studentized test statistic, the first term of the Edgeworth expansion depends on the first two moments.


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
спасибо
faafetai lava
vinaka
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
спасибі
thank you
gracias
хвала
nanni
nandri
dhanyavad
mauruuru
asante
manana
bedankt
bayarlalaa
gràcie
hvala
köszönöm
aciū
obrigada
murakoze
tenki
enkosi
akun dankon
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dĕkuji
sukriya
kop khun krap
taiku
obrigado
sagolun
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
shukriya
mercé
мерси
kam sah hamnida
rahmat
terima kasih
rahmet
diolch
dhanyavadagalu
ευχαριστώ
merci
তোমাকে ধন্যবাদ
감사합니다
xièxie