Genome-Wide Association Studies for Bivariate Sparse Longitudinal Data

Authors: Kiranmoy Das,Jiahan Li,Guifang Fu,Zhong Wang,Rongling Wu

Presented by: Adrija Bhar

Indian Statistical Institute

Introduction

- The successful completion of the Human Genome Project (in 2005) made a revolution in detecting genes controlling various traits and diseases by genome-wide association studies (GWAS).
- The genetic interactions and anomalous nature of complex genes have been better studied by geneticists and biomedical experts.
- paramount importance in developing advanced treatment methods and powerful drugs for diseases with high complexity.
- Despite their potential effect on biomedical sciences, traditional GWAS suffer from several severe limitations.

Powerful Statistical Approach *f*GWAS

- GWAS consider a single time point measurement per subject.
- Most GWAS have found only a small proportion of genetic variation.
- For correlated phenotypic traits are measured longitudinally for many subjects, GWAS focus on single phenotype-genotype analyses, without taking into account the possible dependence among the biomarkers.
- In all senses, a curve is more informative than a single point.

Past Significant work on Correlated Biomarkers

- For the analysis of univariate repeated measures, the linear mixed model-based approach available in standard statistical packages and are very popular.
- A bivariate linear mixed model approach using the standard statistical package (SAS) was proposed by Thiebaut et al.; Sithole and Jones used such a model for detecting prescribing change in two drugs simultaneously with correlated errors.
- A multivariate linear model is more powerful than using a separate model for each biomarker.

Objective of this Article

- Longitudinal measurements with bivariate response have been analyzed by several authors using two separate models for each response.
- For most of the biological or medical experiments, the two responses are highly correlated.
- A single model considering a bivariate response provides a more powerful inference modelling the correlation between the responses appropriately.
- The authors have proposed a dynamic statistical model, based on a direct relationship between phenotype and genotype, to detect significant SNPs associated with blood pressure by considering the gender-gene interaction.

Data Overview

- Used a similar bivariate longitudinal model to analyze data from the Framingham Heart Study (FHS)
- Collectively the data provide records of phenotypes at many time points(total number of measurements vary from subject to subject)
- 977 subjects (all Caucasians: people from Europe, the Middle East, and parts of Central Asia): 500 males and 477 females.
- Their systolic and diastolic blood pressure, BMI and many other variables measured at multiple time points.
- The number of serial measurements for a subject can be as low as 3, but the times and intervals of measurements are highly variable among the different subjects.(reduces the limitations due to sparse data).

Raw Data

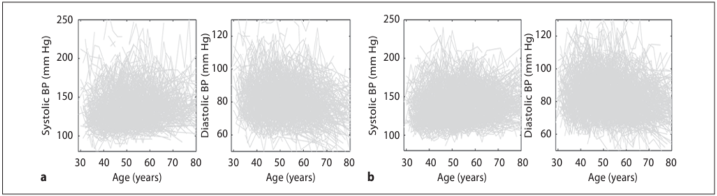

Fig. 1. Raw data for systolic and diastolic blood pressure for males (**a**) and females (**b**), respectively.

SNP- Description

- Total 550,000 SNP data from the entire human genome.
- Excluded SNPs with minor allele frequencies (MAF) of less than 0.10(similar to GWAS).
- The numbers and percentages of non-rare allele SNPs vary among different chromosomes and range from $4,417$ to $28,771$ and from 0.64 to 0.72, respectively.

Statistical Model

- Let
$$\boldsymbol{y}_{ik} = [y_{ik}(t_{i1}), y_{ik}(t_{i2}), \ldots, y_{ik}(t_{iT_i})]$$
denote the vector of the trait of type k ($k = 1$ for systolic blood pressure and $k = 2$ for diastolic blood pressure) for the i-th subject measured at time points $t_i = [t_{i1}, \ldots, t_{iT_i}]$.
- Consider n subjects measured at different time points.
- Consider a SNP with two alleles, A and a, three genotypes: AA (coded as 1) with n_1 observations, Aa (coded as 2) with n_2 observations, and aa (coded as 3) with n_3 observations.

Statistical Model

- The k-th type phenotypic value for subject i at time $t_{i\tau}$ $(\tau = 1, \ldots, T_i)$ is expressed as

$$y_{ik}(t_i) = \sum_{j=1}^{3} \xi_{ij}\mu_{jk}(t_{i\tau}) + \beta x_i(t_{i\tau}) + e_{ik}(t_{i\tau}),$$

- ξ_{ij}: indicator variable taking value 1 if the ith subject is of genotype j and 0 if otherwise
- $\mu_{jk}(t_{i\tau})$: mean value for genotype j of the k-th response at time t_i,
- $x_i(t_{i\tau})$ is the time-variant covariate (in this case, BMI),
- β: regression coefficient of the covariate, assumed to have the same effect on both systolic and diastolic blood pressure.
- $e_{ik}(t_{i\tau})$: residual error for subject i, assumed to be distributed as multivariate normal with mean 0 and covariance matrix Σ_i.

Brief Overview of Orthogonal Legendre Polynomial(LP)

- For modeling the mean curves, a nonparametric approach based on orthogonal Legendre polynomials (LP) is used.
- These polynomials have already been proven a powerful tool to model longitudinal or functional data by several authors
- The robustness of nonparametric regression using LP has been investigated effectively.

Brief Overview of Orthogonal Legendre Polynomial(LP)

- The LP are solutions to the Legendre differential equation

$$(1-x^2)\frac{d^2z}{dx^2} - 2x\frac{dz}{dx} + r(r+1)z = 0$$

- The general form of an LP of order r is given by the following sum

$$P_r(x) = \sum_{l=0}^{L} (-1)^l \frac{(2r-2l)!}{2^r l!(r-l)!(r-2l)!} x^{r-2l}$$

where $L = \frac{r}{2}$ or $\frac{r-1}{2}$ whichever is an integer. These polynomials are defined over $[-1, 1]$ and are orthogonal to each other in this interval in the sense that

$$\int_{-1}^{1} P_r(x)P_s(x)\,dx = 0 \quad \text{when} \quad r \neq s.$$

Let $P_r(t)$ denote the rth order LP at time t.

Estimation of the Mean Curves

- The following transformation of the original time points $t_1, t_2, t_3, \ldots$ to $t'_1, t'_2, t'_3, \ldots$, are done where:

$$t'_i = -1 + 2\frac{(t_i - t_{\min})}{t_{\max} - t_{\min}},$$

- $t_{\min}$ and $t_{\max}$ are the shortest and the longest time points, respectively.
- A family of orthogonal LP is denoted by

$$P(t) = [P_0(t), P_1(t), \ldots, P_r(t)]^T,$$

- The genotype-specific mean value is expressed as a linear combination of the polynomials, such as

$$\mu_{jk}(t) = u_{jk}^T P(t),$$

  where $u_{jk} = (u_{jk0}, u_{jk1}, \ldots, u_{jkr})^T$ is called the base vector.
- The order of the LP is chosen by an information criterion like AIC/BIC, etc.

Other approaches

- Fourier series, B-splines, wavelets, etc.
- If the trait under consideration is well studied, a known parametric structure might be more effective(logistic equation for growth for plants).
- Semiparametric models can also be implemented in some situations for a better understanding of the process under consideration.

Estimation of the Covariance Structures

- In longitudinal data analysis, it is also fundamentally important to model the within-subject covariance structure in a robust and powerful way.
- If the trait variation within subjects is constant, one might use a stationary autoregressive model(to estimate only two parameters).
- In order to handle more complex covariance structures, the autoregressive moving average model (ARMA(p, q)) might be implemented too.
- More recently, Fan and Wu have developed a semiparametric kernel-based covariance function which is robust and effective for modeling subject-specific covariance structure for longitudinal data analysis.

Estimation of the Covariance Structures

- The covariance matrix for the bivariate response longitudinal data as a Kronecker product of two covariance structures following Sithole and Jones.
- For a fixed k ($k = 1$ and $k = 2$), the longitudinal measurements from the same subject are assumed to have an AR(1) structure.
- The correlation between the systolic and diastolic blood pressure remains the same over time for a fixed subject.
- The covariance structure can be expressed as UN $\otimes$ AR(1), where

$$\text{UN} = \begin{pmatrix} \sigma_s^2 & \sigma_{sd} \\ \sigma_{sd} & \sigma_d^2 \end{pmatrix}, \quad \text{AR}(1) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Features of such Structure

- This assumes that an intra-blood pressure correlation is the same for systolic and diastolic.
- inter-blood pressure correlation is proportional to an intra-blood pressure correlation.
- can be modeled easily using the standard statistical packages.
- exploratory data analysis is performed to investigate the validity of such a structure for irregular longitudinal data using SAS, and it turned out that SAS is equally powerful in this case as well.

Additive and Dominant Effects

- The additive (a) and dominant (d) effects over time of the SNP under consideration can be expressed as follows:

$$a_k(t) = \frac{1}{2}\left[\mu_{1k}(t) - \mu_{3k}(t)\right]$$

$$d_k(t) = \mu_{2k}(t) - \frac{1}{2}\left[\mu_{1k}(t) + \mu_{3k}(t)\right]$$

- Note here

$$a_k(t) = 0 \Rightarrow \mu_{1k}(t) = \mu_{3k}(t)$$

$$d_k(t) = 0 \Rightarrow \mu_{2k}(t) = \mu_{1k}(t) = \mu_{3k}(t)(\text{given } a_k(t) = 0)$$

Likelihood of our Study

- Assuming the subjects are independent, the joint likelihood function can be written as

$$L = \prod_{i=1}^{n_1} f_1(\boldsymbol{y}_i) \prod_{i=1}^{n_2} f_2(\boldsymbol{y}_i) \prod_{i=1}^{n_3} f_3(\boldsymbol{y}_i),$$

- $y_i = [y_{i1}, y_{i2}]^T$: response vector for the i-th subject
- $f_j(\boldsymbol{y}_i)$: multivariate normal density for the i-th subject carrying SNP genotype j ($j = 1, 2, 3$), with the following mean vectors:

$$\begin{array}{ll} \left[\mu_{11}(t_{i\tau}) + \beta x_i(t_{i\tau}), \mu_{12}(t_{i\tau}) + \beta x_i(t_{i\tau})\right], & \text{for genotype AA} \\ \left[\mu_{21}(t_{i\tau}) + \beta x_i(t_{i\tau}), \mu_{22}(t_{i\tau}) + \beta x_i(t_{i\tau})\right], & \text{for genotype Aa} \\ \left[\mu_{31}(t_{i\tau}) + \beta x_i(t_{i\tau}), \mu_{32}(t_{i\tau}) + \beta x_i(t_{i\tau})\right], & \text{for genotype aa} \end{array}$$

  and the subject-specific covariance matrix Σ_i.
- Maximized the log likelihood to obtain the mle for the model parameters.

Hypothesis Testing

- The hypotheses can be formulated in the following way:

  $$H_0 : \mu_{1k} = \mu_{2k} = \mu_{3k}$$

  vs

  $$H_1 : \text{at least one equality in } H_0 \text{ does not hold}$$

  , for $k = 1, 2$.
- The LR test statistic $-2\log(LR)$ asymptotically follows a χ^2 distribution with the degree of freedom equal to the difference in the numbers of unknown parameters under H_0 and H_1 .
- Because of the large number of SNPs under consideration, it is necessary to adjust for multiple comparison and a standard false discovery rate (FDR) approach is used as proposed by Benjamini and Hochberg.

Optimal order of LP

- **Optimal polynomial order:** $r = 2$ (based on BIC).

Order (r)	BIC (Male)	BIC (Female)
1	4.62	3.85
2	**1.62**	**1.95**
3	3.84	2.77
4	1.99	2.01

Table: BIC values for Legendre polynomial order selection

Significant SNPs

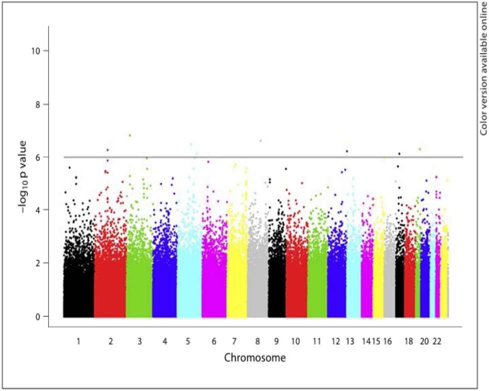

Fig. 2. Manhattan plot for males: selection of the most significant SNPs controlling blood pressure.

Significant SNPs

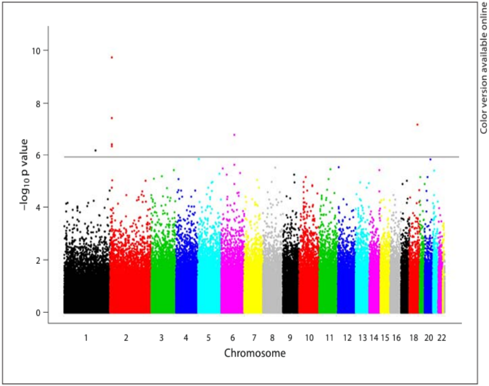

Fig. 3. Manhattan plot for females: selection of the most significant SNPs controlling blood pressure.

Significant SNPs

- With a significance level = 10–6 (with an estimated FDR = 0.002), 8 significant SNPs for the male and 7 significant SNPs for the female population have been selected.
- For males: the detected SNPs are
  - rs66475406 on chromosome 2
  - rs66149495 on chromosome 3
  - rs66299297 and rs66501706 on chromosome 5,
  - rs66484226 on chromosome 8
  - rs66379521 on chromosome 12
  - rs66154967 on chromosome 17
  - rs66092412 on chromosome 19.
- For females:
  - rs66076226 on chromosome 1
  - rs66053327, rs66123984, rs66114880 and rs66347789 on chromosome 2
  - rs66447584 on chromosome 6
  - rs66225752 on chromosome.

Genotype-Specific Mean Curves

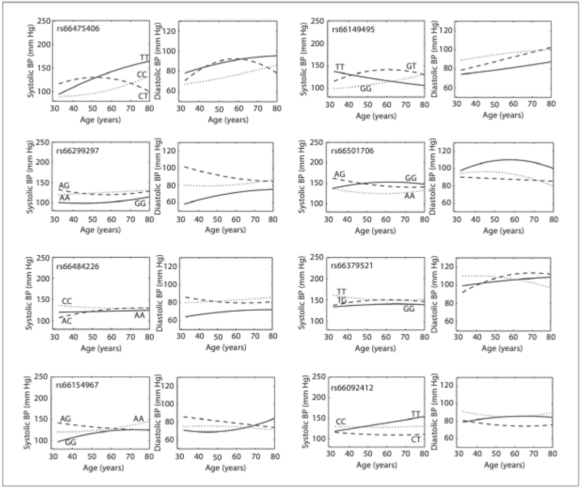

Fig. 4. Genotype-specific mean curves for the most significant SNPs for males.

Genotype-Specific Mean Curves

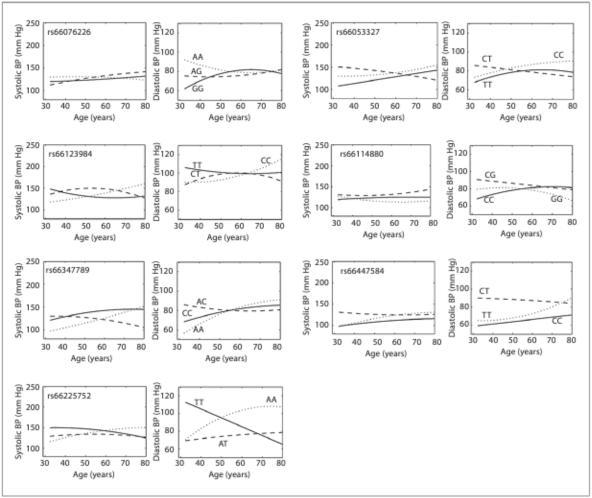

Fig. 5. Genotype-specific mean curves for the most significant SNPs for females.

Observed vs Expected p-value Plot

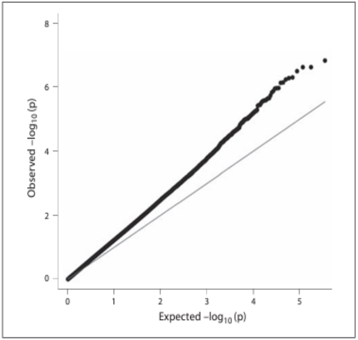

Fig. 6. Observed versus expected P-plot for males.

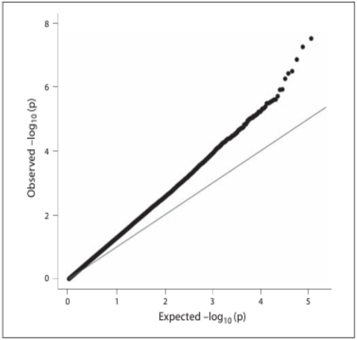

Fig. 7. Observed versus expected P-plot for females.

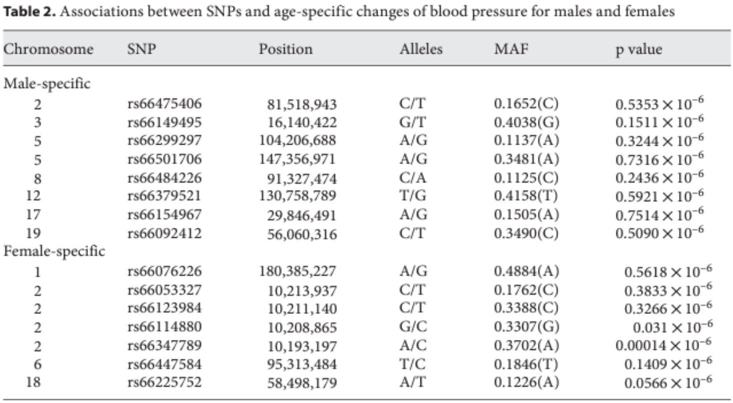

Table 2. Associations between SNPs and age-specific changes of blood pressure for males and females

Chromosome	SNP	Position	Alleles	MAF	p value
Male-specific					
2	rs66475406	81,518,943	C/T	0.1652(C)	0.5353×10^{-6}
3	rs66149495	16,140,422	G/T	0.4038(G)	0.1511×10^{-6}
5	rs66299297	104,206,688	A/G	0.1137(A)	0.3244×10^{-6}
5	rs66501706	147,356,971	A/G	0.3481(A)	0.7316×10^{-6}
8	rs66484226	91,327,474	C/A	0.1125(C)	0.2436×10^{-6}
12	rs66379521	130,758,789	T/G	0.4158(T)	0.5921×10^{-6}
17	rs66154967	29,846,491	A/G	0.1505(A)	0.7514×10^{-6}
19	rs66092412	56,060,316	C/T	0.3490(C)	0.5090×10^{-6}
Female-specific					
1	rs66076226	180,385,227	A/G	0.4884(A)	0.5618×10^{-6}
2	rs66053327	10,213,937	C/T	0.1762(C)	0.3833×10^{-6}
2	rs66123984	10,211,140	C/T	0.3388(C)	0.3266×10^{-6}
2	rs66114880	10,208,865	G/C	0.3307(G)	0.031×10^{-6}
2	rs66347789	10,193,197	A/C	0.3702(A)	0.00014×10^{-6}
6	rs66447584	95,313,484	T/C	0.1846(T)	0.1409×10^{-6}
18	rs66225752	58,498,179	A/T	0.1226(A)	0.0566×10^{-6}

Additive and Dominant Effect Graphs

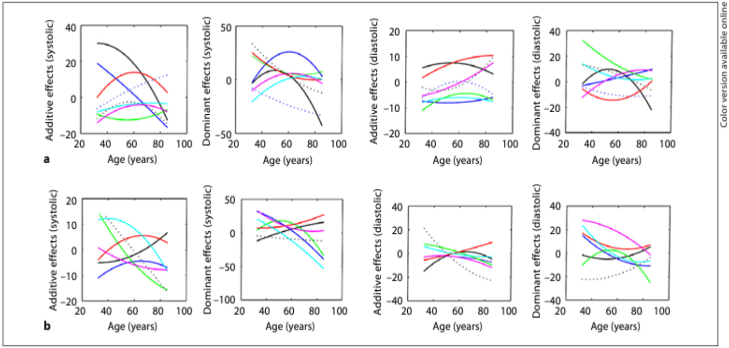

Fig. 8. Additive and dominant effects for the most significant SNPs for males (**a**) and females (**b**), respectively.

Simulation Results and Power Analysis

- To examine the power, statistical properties practical and applicability and usefulness of the proposed joint analysis was assessed by further simulation studies.
- Sparse trait values were simulated from the model and the covariance structure we used for original data analysis.
- Under two different setups, when the bivariate data are affected by pleiotropic genes and when they arenot affected by such genes, power and false positive rate (FPR) analyses are performed with different values for MAF.
- We considered two different sample sizes ($n = 1,000$ and $2,000$) for each situation.

Genotype-Specific Mean Curves from Simulation Studies

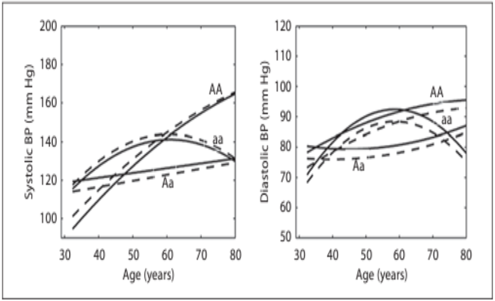

Fig. 9. Simulation results: actual (solid) versus fitted (dashed) curves for different genotypes.

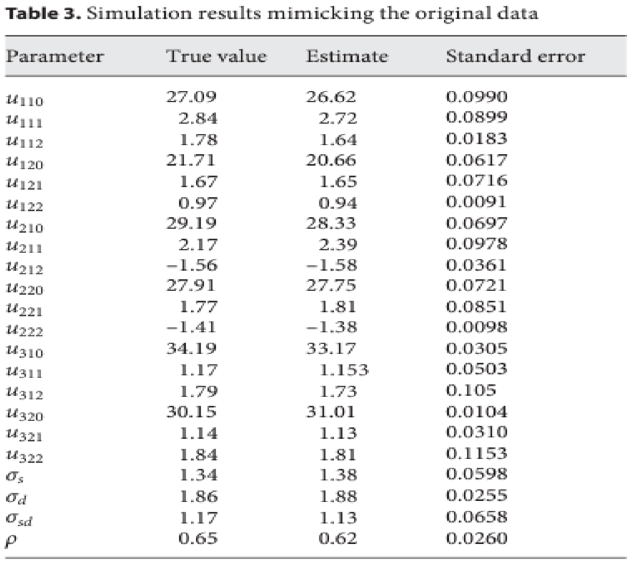

Table 3. Simulation results mimicking the original data

Parameter	True value	Estimate	Standard error
u_{110}	27.09	26.62	0.0990
u_{111}	2.84	2.72	0.0899
u_{112}	1.78	1.64	0.0183
u_{120}	21.71	20.66	0.0617
u_{121}	1.67	1.65	0.0716
u_{122}	0.97	0.94	0.0091
u_{210}	29.19	28.33	0.0697
u_{211}	2.17	2.39	0.0978
u_{212}	−1.56	−1.58	0.0361
u_{220}	27.91	27.75	0.0721
u_{221}	1.77	1.81	0.0851
u_{222}	−1.41	−1.38	0.0098
u_{310}	34.19	33.17	0.0305
u_{311}	1.17	1.153	0.0503
u_{312}	1.79	1.73	0.105
u_{320}	30.15	31.01	0.0104
u_{321}	1.14	1.13	0.0310
u_{322}	1.84	1.81	0.1153
σ_s	1.34	1.38	0.0598
σ_d	1.86	1.88	0.0255
σ_{sd}	1.17	1.13	0.0658
ρ	0.65	0.62	0.0260

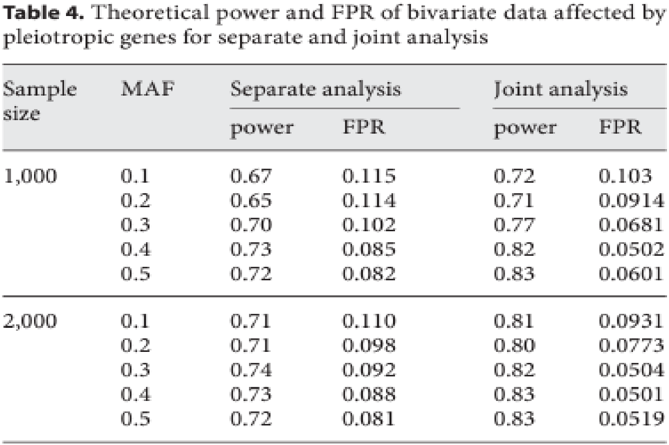

Table 4. Theoretical power and FPR of bivariate data affected by pleiotropic genes for separate and joint analysis

Sample size	MAF	Separate analysis		Joint analysis	
		power	FPR	power	FPR
1,000	0.1	0.67	0.115	0.72	0.103
	0.2	0.65	0.114	0.71	0.0914
	0.3	0.70	0.102	0.77	0.0681
	0.4	0.73	0.085	0.82	0.0502
	0.5	0.72	0.082	0.83	0.0601
2,000	0.1	0.71	0.110	0.81	0.0931
	0.2	0.71	0.098	0.80	0.0773
	0.3	0.74	0.092	0.82	0.0504
	0.4	0.73	0.088	0.83	0.0501
	0.5	0.72	0.081	0.83	0.0519

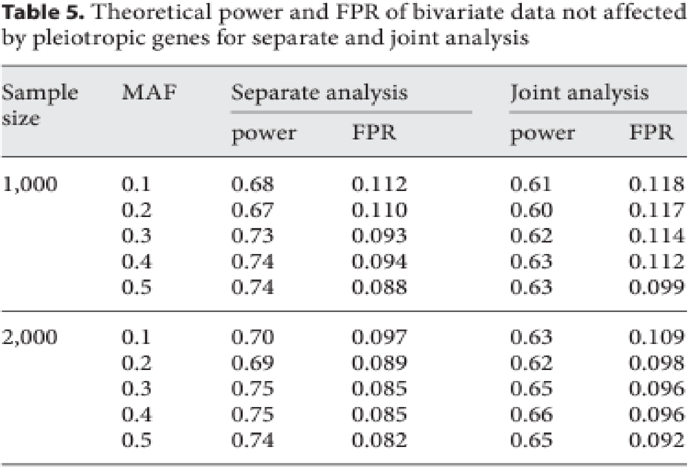

Table 5. Theoretical power and FPR of bivariate data not affected by pleiotropic genes for separate and joint analysis

Sample size	MAF	Separate analysis		Joint analysis	
		power	FPR	power	FPR
1,000	0.1	0.68	0.112	0.61	0.118
	0.2	0.67	0.110	0.60	0.117
	0.3	0.73	0.093	0.62	0.114
	0.4	0.74	0.094	0.63	0.112
	0.5	0.74	0.088	0.63	0.099
2,000	0.1	0.70	0.097	0.63	0.109
	0.2	0.69	0.089	0.62	0.098
	0.3	0.75	0.085	0.65	0.096
	0.4	0.75	0.085	0.66	0.096
	0.5	0.74	0.082	0.65	0.092

Challenges

- The challenge here is to model the covariance structure appropriately. we have seen nonparametric modeling of the mean function and a parametric covariance structure assuming that the intra-response correlation is the same for each response and the inter-response correlation is directly proportional to the intra-response correlation.
- Authors are currently developing a new approach relaxing those assumptions.

Main Usefulness of this Article

- Once genes having significant effects on blood pressure are identified, heart diseases which are the consequences of abnormal blood pressures can be treated and controlled more effectively.
- Also the same procedure could be applied to detect genes controlling other traits and diseases too.
- Both traits need to be measured concurrently. Even a simple extension of this method works for multivariate traits under the same underlying assumption.

Some Limitations

- The underlying assumption for this work is that when one trait is observed for a specific subject at a particular time point, we must also have measurement for the second trait at that point.
- But it might not be the case in a general setting. A more sophisticated method needs to be developed to handle those situations.
- Also all subjects under study are treated independent of each other.
- However, in reality, the subjects might be genetically related or even come from the same family.
- Once the information is there for genetic dependence of the subjects, that can be incorporated in the method and this will reflect biostatistically more informative results.

Thank You