HETEROSCEDASTIC APPROACHES FOR DECIPHERING

MULTIETHNIC GENOMIC SEQUENCES AND

MICROARRAYS

A DISSERTATION

SUBMITTED ON THE THIRD DAY OF APRIL 2017

TO THE DEPARTMENT OF GLOBAL BIOSTATISTICS AND DATA SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF PUBLIC HEALTH AND TROPICAL MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

Weiwei Ouyang

Weiwei Ouyang, BS, MS

APPROVED:

Huaizhen Qin, PhD; Date

4/3/2017

John Lefante, PhD; Date

4/3/2017

Sudesh Srivastav, PhD; Date

4/3/2017

Tanika Kelly, PhD; Date

4/3/2017

Fang Zhide, PhD; Date

4/3/17

ABSTRACT

Advanced omics technologies have been generating abundant multi-ethnic multi-omics data, including DNA sequences, methylations, gene expressions, and copious clinical traits. Such big data pose unprecedented challenges due to the high complexity of heterogeneous networks between biomarkers. Heteroscedasticity (aka, dispersion heterogeneity of trait residuals) is a common phenomenon in multi-omics data mining. It can be caused by interactions such as gene×gene, gene×enviroment, linkage disequilibrium (LD) between marker loci, and pleiotropic traits as well. Especially, it occurs in the data mining of the multi-omics data of admixed individuals subjects due to broad admixture LD and gene×ancestry interactions. Meanwhile, it can be induced by background confounders, e.g., population structure, cryptic relatedness, polygenetic effects, and correlations between residuals of multiple traits. However, existent univariate and multivariate methods neglect all the high-order effects of both test biomarkers and background confounders. This dissertation contributes systematic harmonious signal augmentation methods with applications for distilling high-order information from multiethnic DNA sequences to microarrays. In Chapter I, we proposed a novel harmonious signal augmentation schemes in single-based association tests. The harmonious single-based association test (HSAT) is more powerful then existent single-based methods in both simulations and real data application. In Chapter II we put forth harmonious gene-based association tests (HGAT) to incorporate high-order effects.

Within a gene, the importance of a test variant is measured by the signal of marker-wise high-order effects. Leveraging high-order effects of genetic variants has proven to improve power for identifying susceptive genes. By extensive simulations under published designs, the proposed method properly controlled type I error rates and appeared strikingly more powerful than existent prominent gene-based sequence association methods. We apply HGAT methods in homogeneous population and admixed population. There are two parts in Chapter III, the first part introduced integrating informative mean and variance effects to identify differentially expressed (DE) genes. The second part illustrated the application of harmonious integration of mean and high order effects to identify differentially expressed (DE) genes. In summary, this dissertation demonstrated tremendous potential of explicitly distilling informative higher-order effects in big multiethnic multi-level data mining and offered paradigm applications for integrating high-order information resources while effectively calibrating major heteroscedastic confounders.

Keywords: Single-based and gene-based tests, Harmonious signal augmentation, High-order heterogeneities, admixed population, differentially expressed (DE) genes

ACKNOWLEDGEMENT

First and foremost, I must acknowledge and thank Dr Qin, my advisor who gave me constant guidance and encouragement in the past several years. With his guidance, I was trained to be a professional researcher in Biostatistics and learned lots of things in how to conduct research. It is my pleasure to acknowledge and thank the members of my dissertation committee members- Dr Qin, Dr John Lefante, Dr Sudesh Srivastav, Dr Tanika Kelly and Dr Zhide Fang - for their cooperation, supervision and availability. My sincere thanks, gratitude and appreciation to my mother and father for their love, support and understanding during the long years of my education. I am grateful to the Department of Global Biostatistics and Data Science at Tulane University's School of Public Health and Tropical Medicine where I pursued knowledge and spent a good time during these years. Lastly, I would like to thank my friends and colleagues for their support, feedback and friendship.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

HARMONIOUS SIGNAL AUGMENTATION SCHEMES IN SINGLE-BASED ASSOCIATION TESTS

1.1 Abstract

Current prominent single-based association methods of complex disease phenotypes are based on homoscedasticity working models, i.e., linear models (LMs), generalized linear models (GLMs) and generalized linear mixed models (GLMMs), which only aim to exploit the mean effects of variants on disease traits. All these models assume homoscedasticity that model residuals are independent of all predictors (covariates and variants). As shown by real-world genetic data, the assumption of homoscedastic residuals is incompetent to account for phenotypic variation induced by the complex structure of biological networks. In this paper, we proposed a novel harmonious signal augmentation schemes to solve the so-called “Searching Needles in the Haystack” problem in single-based association tests. Two advantages are highlighted in our novel schemes: (1) Integrating mean effect and high-order effect, which indicates association signal of genotypes on the high-order moments of quantitative traits beyond the first order moment (e.g. the mean), can effectively select single variants that involve in potential interactions and causal networks, latent covariates. (2) The few association methods integrating mean and variance effects of genotypes sacrifice statistical power and had poor association power for susceptible low and rare frequency variants(i.e. minor

allele frequency (MAF)<5%). By extensive simulations and real data application to COGA data, our harmonious augmentation method brought about dramatic association power gain for detecting low and rare frequency variants and demonstrate the superiority of the novel method to existing mean-only and mean-variance association tests for continuous trait in homogeneous populations under situations of variance heterogeneity and $G \times E$ interactions.

Key words:

Harmoniously integration, Variance heterogeneity, $G \times E$ and $G \times G$ interactions, High-order effects, Single-based method

1.2 Introduction

In recent years, the development of sequencing technologies is accelerating the process of localize genetic determinants (i.e., susceptible variants, genes) which govern the underlying disease risk and trait value. Current prominent single-based sequence association methods of complex disease phenotypes are based on homoscedastic regression models. i.e., generalized linear models (GLMs)[1], and generalized linear mixed models (GLMMs)[2, 3],which only aim to exploit the effects of variants on the alteration of first-order moment of disease traits (i.e. the mean). All these models assume homoscedasticity that model residuals are independent of all predictors (covariates and variants) and consistent across all individual values. As shown by real-world genetic data, the homoscedastic models are too simple to effectively capture high complex disease genetic mechanisms.

The violation of homoscedasticity is called the heteroscedasticity problem and was noticed as early as in 2000 in genetic field[4]. The emerging interests of heteroscedasticity relied on the point that rather than a concern of impediment to statistical modeling of genetic data, heteroscedasticity could be of biological interest. Exploring heteroscedasticity can be regarded as an alternative to identify single-variant that involves in potential interactions and causal networks with latent covariates. Variants that display variance heterogeneity can be caused by biological disruption, linkage disequilibrium (LD), gene-by-gene (G × G), or gene-by-environment (G × E) interaction. For example, variability-controlling quantitative trait loci (vQTL)[5-11] are genetic variants whose allelic states associate with phenotypic variability, namely the variance of phenotype values around the mean. vQTL shows some evidence of potential interactions with other genes (e.g., Gene×Gene, namely, epistasis) or environmental factors (e.g., Gene×Enviroment), and the locus with an inflation of variance within its genotypes due to being a mixture of genotypes from the genuine causal loci[11]. Another example is Expression variability QTLs (evQTLs)[12] that were reported as marker loci whose allelic states are associated with variances of gene expression.

If a genetic locus is genuinely functional to the disease, it would lead to the alteration of trait distribution instead of solely the trait mean. From a statistical perspective, the distribution of a random variable can be completely determined by all its moments. High-order moments can capture extra information beyond mean heterogeneity of the outcome. For single-SNP analysis, the main advantage of few methods integrating mean and variance effects is to detect SNPs that are related to the alteration of trait distribution in presence of $G \times G$, $G \times E$ and dependence of variance of traits on

genotype. For example, Cao et al.[11] considered mean and variance differences simultaneously by proposing a versatile likelihood joint test (LRT). And Soave et al.[13] proposed another joint location-scale (JLS) testing framework that simultaneously tests the mean and variance by aggregating association evidence from the location-only (i.e. the partial t-test on mean effect) and scale-only tests (i.e. Levene's test [14] for dispersion heterogeneity of phenotypic residuals among three genotypic groups) using Fisher's combination method[15]. They demonstrated the superior power of the JLS method when $G \times E$ interactions exist and are not explicitly modeled. The two integrative methods mentioned above combine the association signals from mean test and variance test orthogonally. In such orthogonal integration, the mean and variance test statistics are independent to each other, for both causal SNP and neutral SNP. Their perceived disadvantage are the essential power loss than conventional mean-only association test (MT) when association signal from variance tests is weak relative to that from mean test.

Heteroscedasticity can be driven by the effects of genotypes on high-order moments of trait values beyond the first order moment (namely, the mean). But it is a very narrow conception to only indicate variance effects of genotypes on the variance of quantitative traits. The independence of mean and variance tests would bring about power loss as illustrated. Therefore in contrast to the orthogonal joint tests above, our framework incorporates high-order association signals instead of only the variance effect of genotypes on trait values harmoniously. Herein we propounded the novel ideology of harmonious tests: A pair of (mean and high-order) tests are harmonious if (1) *Null independence*: they are independently distributed at neutral markers; (2) *Alternative dependence*: they are statistically dependent at causal markers and their flanking markers.

The null independence warrants us to properly control of type I error rate. This is crucial to prevent false positives. The merit of alternative dependence is the core novelty of our method. In presence of latent factors, this merit can warrant dramatic power gains than do famed orthogonal combination methods, i.e., the recently published JLS and LRT and can effectively augment the association signal even if it only has mean heterogeneity; while the famed methods lead to essential power loss.

The major challenge of genome-wide sequencing studies is like "searching needles in the haystack". For a complex trait, the functional variants are usually sparsely scattered along the genomes and the minor allele frequency of the SNPs are usually small or moderate. In this paper, we proposed our novel single-based test framework of harmoniously integrating mean and high-order effects of test markers while easily calibrating both the mean and dispersion effects of global covariates. Two major advantages are highlighted in our novel test framework: (1) Integrating mean effect and high-order effect, which indicates association signals of genotypes on the high-order moments of quantitative traits beyond the first order moment (e.g. mean), can effectively select single-variant that involves in potential interactions, causal networks, latent covariates. (2) The existent association methods had poor association power of susceptible low and rare frequency variants in sequencing studies. Our harmonious augmentation method brought about dramatic association power gain for detecting low and rare frequency variants (e.g. $MAF < 5\%$) and demonstrate the superiority to existing mean-only and mean-variance association tests for continuous trait in homogeneous populations under situations of variance heterogeneity and $G \times E$ interactions. In addition, our novel gene-based method are capable of obtaining analytical p-values and is

convenient to implement on genome-wide scale. By extensive simulations for single variants under different scenarios, our novel method presents strikingly power gains than existing methods. Moreover, we have applied our method to the COGA on the alcohol addiction of 991 whites. The results demonstrate the noteworthy superiority of our method to existing tests in replicating and identifying novel susceptive variants.

1.3 Materials and methods

1.3.1 Model Notation and Construction of Harmonious Single-Based Association Test (HSAT)

Firstly, we demonstrate the notations and assumptions of our HSAT method. To be specific, let Y_i be a quantitative trait residual for individual i after adjusting for heteroscedastic effects of environmental covariates, G_i be the copy number of minor alleles at the test SNP ($G_i = 0,1,2$). n is the population size in the study. In context, the symbols "*mo*", "*ho*" and "*mh*" stands for "modeling the genetic mean effects only", "modeling genetic high-order effects only" and "modeling genetic mean and high-order effects jointly". The novel model framework can be written as:

Primary Test (*mo*)

$$Y_i = \mu_1 + G_i\beta_1 + e_i, \qquad \textbf{Eq. 1-1}$$

where μ_1 is the intercept and e_i represents the random error term. The distribution assumption of e_i can be loosed to symmetric distribution with $E(e_i) = E(e_i^3) = 0, E(e_i^4) < \infty$, in which $E(.)$ is expectation function. β_1 is the effect size of genotype G_i

on trait value Y_i. The null hypothesis of testing the mean effect is $H_{0.mo}: \beta_1 = 0$. The test statistic is

$$T_1 = \frac{\sqrt{n}\hat{\sigma}_{Y,G}}{\sqrt{\hat{\sigma}_Y^2\hat{\sigma}_G^2 - \hat{\sigma}_{Y,G}^2}}, \quad \textbf{Eq. 1-2}$$

where $\hat{\sigma}_{Y,G} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(G_i - \bar{G})$, $\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $\hat{\sigma}_G^2 = \frac{1}{n}\sum_{i=1}^{n}(G_i - \bar{G})^2$, in which $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y$, $\bar{G} = \frac{1}{n}\sum_{i=1}^{n} G_i$. T_1 is equivalent to $\frac{\sqrt{n-2}r(Y,G)}{\sqrt{1-r^2(Y,G)}}$, where $r(Y,G)$ is the sample Pearson coefficient of correlation between G and Y.

Auxiliary Test (*ho*)

$$Y_i^2 = \mu_2 + G_i^2\beta_2 + e_i', \quad \textbf{Eq. 1-3}$$

where e_i' is the random error term. The highlight of auxiliary test is to capture the second-order moment information beyond the mean by regressing the squared trait residual Y_i against the square of genotype G_i. β_2 is the effect coefficient of the squared genotype G_i^2 on Y_i^2. The null hypothesis of testing the high-order effect is $H_{0.ho}: \beta_2 = 0$. High-order effect is a broader concept than dispersion effect. The association statistic of testing β_2 is

$$T_2 = \frac{\sqrt{n}\hat{\sigma}_{Y^2,G^2}}{\sqrt{\hat{\sigma}_{Y^2}^2\hat{\sigma}_{G^2}^2 - \hat{\sigma}_{Y^2,G^2}^2}}, \quad \textbf{Eq. 1-4}$$

where $\hat{\sigma}_{Y^2,G^2} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - \overline{Y^2}\right)\left(G_i^2 - \overline{G^2}\right)$, $\hat{\sigma}_{Y^2}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - \overline{Y^2}\right)^2$, in which $\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$ and $\overline{G^2} = \frac{1}{n}\sum_{i=1}^{n} G_i^2$. T_2 is equivalent to $\frac{\sqrt{n-2}r(Y^2,G^2)}{\sqrt{1-r^2(Y^2,G^2)}}$, where $r(Y^2, G^2)$ is the sample pearson coefficient of correlation between G^2 and Y^2.

If G is associated with Y ($\beta_1 \neq 0$), then G^2 is associated with Y^2 (namely, $r(Y^2, G^2) \neq 0$). Conversely, a test SNP G has nothing to do with the Y of interest (precisely, namely, it does not harbor causal allele and are not in any LD block with any causal loci of the trait), then the mean heterogeneity model (*mo*) is true. Under the *mo* model, it can be mathematically proved that $\beta_2 = \beta_1^2$. Therefore both the primary and auxiliary models would hold with $\beta_1 = \beta_2 = 0$. In detail, the primary and auxiliary tests are called to be harmonious if (1) Null independence: T_1 and T_2 are asymptotically independent if and only if $\beta_1 = 0$; (2) Alternative dependence: T_1 and T_2 are asymptotically dependent if and only if $\beta_1 \neq 0$. The harmonious properties of T_1 and T_2 can be guaranteed by the following proposition.

Proposition: *Under primary model, if* $E(e_i^4) < \infty$ *and* $E(e_i) = E(e_i^3) = 0$, *then* $T_1 - \delta_1$ *and* $T_2 - \delta_2$ *converge in distribution to a bivariate with unit variance and correlation coefficient* $\rho = \beta_1\delta_1^{-1}\delta_2^{-1}[\sigma_e^2(3\mu_4 - 3\mu_2^2 + \mu_1^2\mu_2 - \mu_1\mu_3) + \beta_1^2(\mu_6 - 2\mu_2\mu_4 + \mu_1\mu_5 - \mu_1^2\mu_4 + \mu_2^3 - \mu_1\mu_2\mu_3 + \mu_1^2\mu_2^2)]$, *in which* $\mu_k \overset{\text{def}}{=} E(G^k)$ *for integer* k *and* $var(e_i) \overset{\text{def}}{=} \sigma_e^2$.

$$\begin{pmatrix} T_1 - C_1 \\ T_2 - C_2 \end{pmatrix} \xrightarrow{a.d.} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where C_1 *and* C_2 *are function of* β_1 *and* β_2, *respectively.*

The detailed mathematical proof of the asymptotical bivariate normal distribution of T_1 and T_2 and expressions of C_1 and C_2 are displayed in **APPENDIX A.1**. The joint null hypothesis of our HSAT method is written as:

$$H_{0.mh}: \beta_1 = \beta_2 = 0$$

Based on the proposition, we adopted Fisher's method[15] to define the HSAT statistic as

$$T_{HSAT} = -2(\log(p_m) + \log(p_h)), \qquad \textbf{Eq. 1-5}$$

where p_m be the p value for testing $H_{0.mo}: \beta_1 = 0$ and p_h be the p value for testing $H_{0.ho}: \beta_2 = 0$.

1.3.2 Simulation Designs for single-based analysis

Herein we consider three different scenarios for four methods: HSAT, JLS, LRT and MT, in which MT is the traditional mean association test. Both the LRT and the JLS are orthogonal integrative methods and are not harmonious.

1.3.2.1 *Scenario I*

One main disadvantage for single-based association test is its poor performance on detecting low and rare variants. For *Scenario I*, we focus on the performance of methods on detecting low and rare variants (e.g. $MAF \leq 5\%$). The additive genetic model was applied to generate the data: $Y_i = G_i\beta + e_i$. Where Y_i be a quantitative trait value for individual i, G_i be the copy number of minor alleles at the test SNP ($G_i = 0,1,2$), β is the effect size of genotype on the trait value.

We simulated the SNP with minor allele frequency (MAF) *p* equal to 0.01, 0.025, 0.05, mimicking the low and rare causal SNP. G_i follows binomial distribution with MAF

p. e_i is standard normal distributed error term. Under the additive genetic model, the heritability of genotype (h^2) is defined as $h^2 = \frac{Var(G)}{Var(Y)} = \frac{\beta^2 2p(1-p)}{1+\beta^2 2p(1-p)}$. For one single SNP, we let h^2ranges from 0% to 2% by a grid of 0.1%, which means one single locus can only explain at most 2% of the total trait variance. And the effect size of genotype can be calculated as $\beta = \sqrt{\frac{h^2}{2(1-h^2)p(1-p)}}$. 100,000 replicates were simulated. The sample size is set to 1000 that is close to our real datasets.

1.3.2.2 *Scenario II*

We adopted the simulation design frame by Soave et al[13]. The MAF of genotype G_i was fixed to be 0.3. The trait value Y was simulated from the following model $Y_i = 0.5 * X_{1i} + 0.5 * X_{2i} + 0.3 * X_{3i} + G_i\beta + G_i X_{3i}\delta + e_i$, Where the error term e_i follows a standard normal distribution. X_{1i} is continuous normal distributed covariate (e.g. $X_{1i} \sim N(0,1)$), X_{2i} follows binomial distribution with frequency 0.5 that mimics the binary covariate such as gender. The unobserved exposure variable X_{3i} was binary variable with frequency 0.3 (e.g. $X_{3i} \sim B(2,0.3)$).

The main mean genetic effect β was set to be 0.01, 0.05, and 0.1, and the interaction effect δ of $G_i \times X_{3i}$ was varied between 0 and 1 by a grid of 0.1. This simulation consider the potential latent $G \times E$ interactions in genetic dataset. 100,000 replicates were simulated. The sample size is set to 1000 that is close to our real datasets.

1.3.2.3 *Scenario III*

The simulation design is similar as that by Cao et al. [11]Genotypes G_i affect both mean and variance of quantitative trait Y_i. In other words, a functional locus G_i in this situation may have variance heterogeneity across different genotypes. The quantitative traits Y_i were generated using the following model: $Y_i = 0.5 * X_{1i} + 0.5 * X_{2i} + G_i\beta + e_i$, Where $e_i \sim N(0, \exp(G_i\gamma))$, in which γ is the effect size of genotype on variance and $exp(.)$ is the exponential function which guarantees that variance of the normal distribution is always positive. β is the effect size of genotype. G_i follows binomial distribution with MAF p. For different genotypic score, the trait value Y_i has different variances. X_1 is continuous normal distributed covariate (e.g. $X_{1i} \sim N(0,1)$) and X_{2i} follows binomial distribution with frequency 0.5 that mimics the binary covariate such as gender. The variance effect size γ ranged from 0 to 0.5 by a grid of 0.05. To obtain reasonable power for methods comparisons, we specific β to be 0.5 to mimic the low frequency causal variants ($0.5\% <$MAF$< 5\%$) with relatively large effect size and be 0.25 to mimic common causal variants (MAF$> 5\%$). The MAF p of a common variant is randomly generated from interval (0.05, 0.5) and that of low frequency variant is randomly generated from interval (0.005, 0.05) in each replicate.

1.3.3 Real Data Analysis on Genetics of Alcoholism (COGA) Study

We reanalysis an existing well-characterized sample of 1050 unrelated Africans selected from COGA Study at 936,263 SNPs that span the genome for alcohol dependence (AD). Positions of all SNPs are genome build 36.3. The primary phenotype is DSM-IV AD[16]. SNPs were excluded if minor allele frequency (MAF)<0.5% or call

rates<95%, leaving 856,149 SNPs after genotype quality control. Among 1050 unrelated individuals, 59 were excluded due to missing or extreme trait values. 991 individuals underwent final analysis. Following genotype quality control, we applied the double generalized linear model (DGLM)[17] to adjust for both mean and variance effects of environmental covariates. The DGLM is implemented in R package dglm. The covariates to adjust for in analysis are gender (1=Male, 2=Female), smoking (0~7), normalized age, squared-normalized age and estimated population stratification (e.g. PCs). Since age ranges from 18 to 77, normalizing age can reduce the difference of age profiles. Adding the square of normalized age allows you to model the effect of age that may have a non-linear relationship with the phenotype AD. The inclusion of smoking was to remove possible spurious results caused by effects of smoking considering the moderate relationship ($r^2 = 0.58$) between drinking and smoking. For background population structure adjustment, we didn't follow the routine way to solely account for the top ten PCs. Instead we adjust for PCs with mean effects or variance effects or both on phenotype. After adjusting for both mean and variance effects, global covariates have no significant mean and variance effects in AD. Following all the procedure above, we centralized the trait residuals in the final step. Covariates adjustment and centralization does not remove the genotype information on both the mean and second-order moment of trait value. The flow chart of data-processing was displayed in **Figure 1-1**.

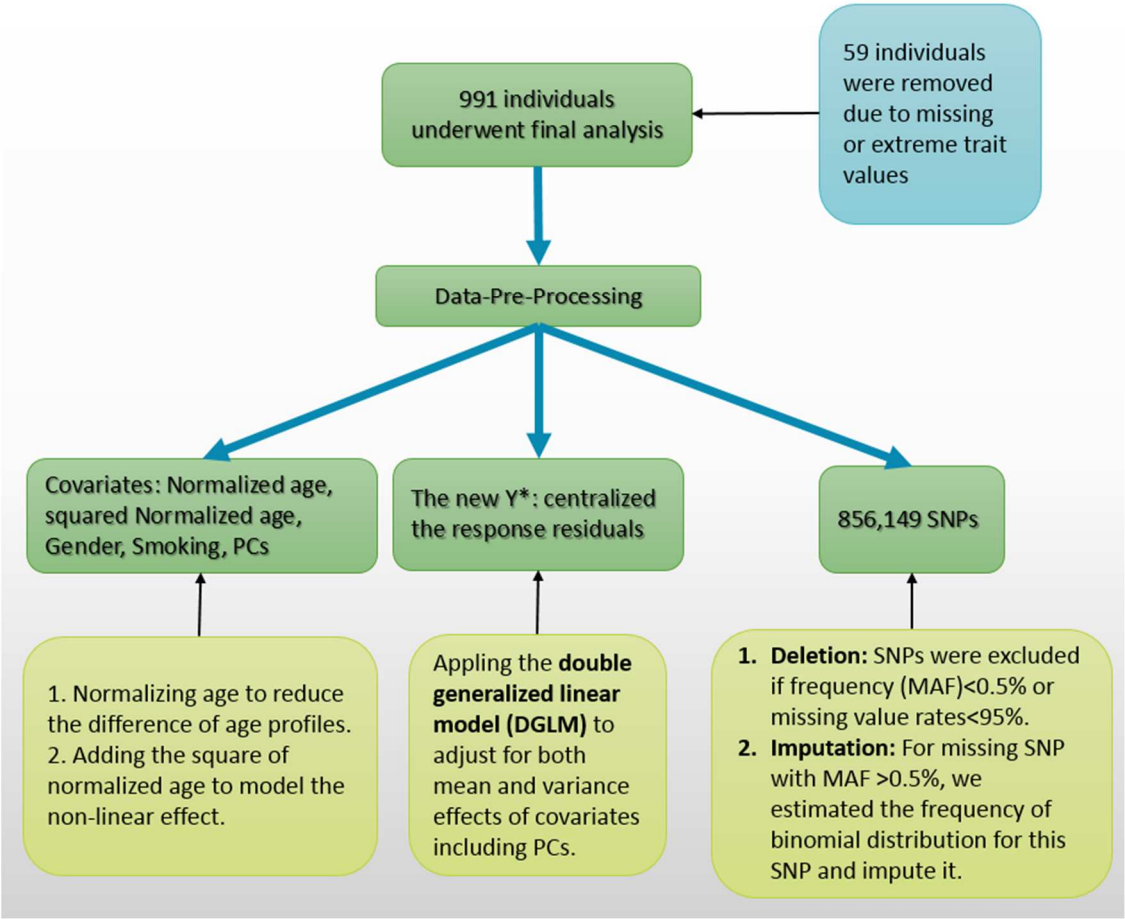

Figure 1-1: Flow Chart of Data Processing of COGA study

1.4 Results

1.4.1 Type I error rate of Single-Based Association tests

100,000 replicates were generated under the null model with no association ($\beta = 0, \alpha = 0, \gamma = 0$) for HSAT, JLS, LRT and MT at different nominal levels. The sample size is still 1000. Seen from **Figure 1-2,** MT, JLS and our HSAT generally controlled Type I error rates at different given nominal significance levels, while LRT is outside of the 95% concentration band and was a little inflated at larger nominal levels.

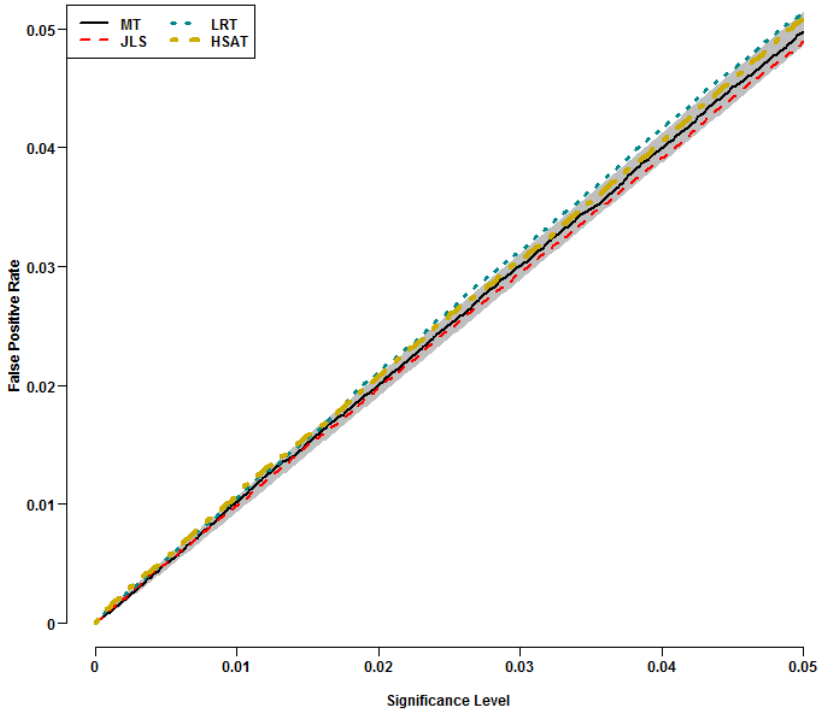

Figure 1-2: Comparison of false positive rates of eight methods under null hypothesis.

1.4.2 Empirical power comparisons of single-based Association Tests

To demonstrate that HSAT is not only robust but also more powerful than either the traditional mean test or other two orthogonal integrative methods, we investigated three simulation scenarios to evaluate the powers of MT, JLS, LRT and HSAT. We set sample size as 1000 that is close to real sample size of COGA study.

Figure 1-3 summarized the results for *Scenario I* at genome-wide nominal level $\alpha = 5 \times 10^{-8}$. This scenario does not favor the integration methods JLS, LRT and HSAT because there are no latent interactions or LD between test and causal locus that can bring about variance heterogeneity on trait value. Our proposed HSAT is the most powerful

among the four methods. Compared with the second most powerful test, MT, the power gain of HSAT is more noticeable for SNP with smaller MAF. In contrast, JLS and LRT is less powerful than MT under all situations. The power loss of JLS and LRT is due to the increase of degree of freedom in integrating mean and variance tests and orthogonal integration didn't bring about enough additional information.

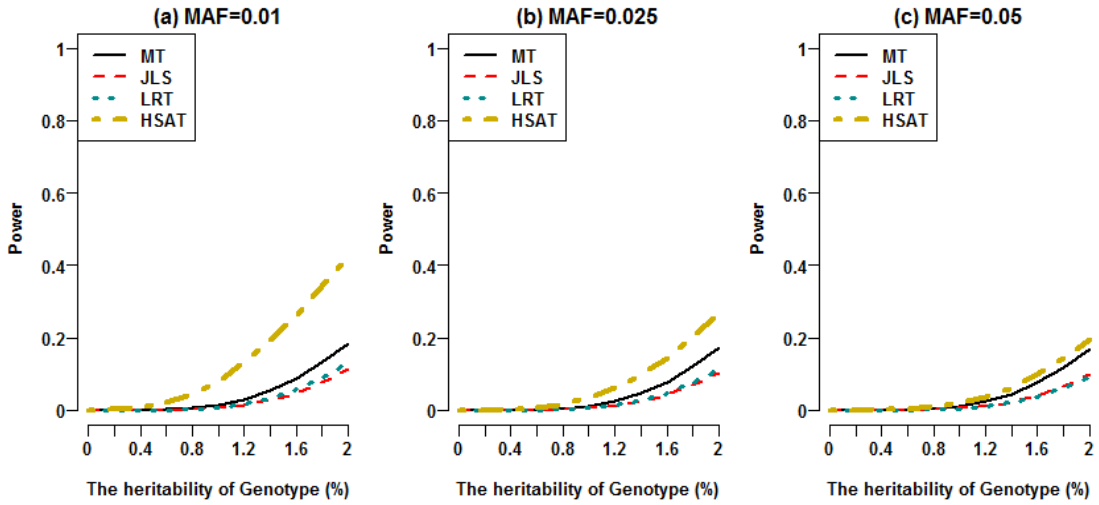

Figure 1-3: Power comparison of MT, JLS, LRT and HSAT under Scenario I at nominal level 5×10^{-8}.

We adopted simulation design by Soave et al[13] in *Scenario II*. There existed unobserved exposure variable (E) and the corresponding interaction between genotype and unobserved exposure variable ($G \times E$) in generating the trait. **Figure 1-4** displayed the power results for the four methods, in which HSAT is the most powerful method with different interaction effect sizes δs, followed by the second most powerful test, JLS for different genetic effect size βs.LRT is less powerful than JLS and is always superior to the traditional MT method. The MT has the least power due to its failure of capturing the

additional information from $G \times E$ interaction. The values of βs and δs are similarly determined as that by Soave et al.

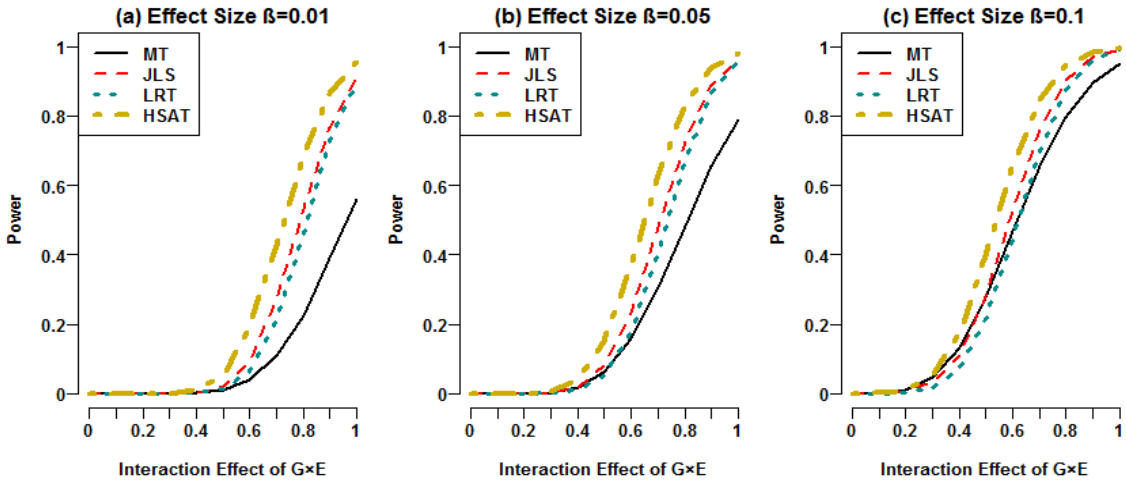

Figure 1-4: Power comparison of MT, JLS, LRT and HSAT under Scenario II at nominal level 5×10^{-8}.

For *Scenario III*, the simulation design framework mimics the situation of variance-heterogeneity loci (vQTL), which have different variances across genotypes. Such variance heterogeneity may be induced by LD with a functional causal variant or $G \times G$ interactions. The main effect size β is set to be 0.25 for common causal variants and 0.5 for low frequency causal variant in order to obtain reasonable power comparisons. Such setting is also consistent with the popular assumption that low and rare variants would have rarer causal loci have greater effects. The effect size γ is a measure of genetic effect on the variance of trait value. **Figure 1-5 (a)** displayed the power comparisons of the four methods for common variants with the main genetic effect size $\beta = 0.25$. When the effect of genotype on variance is small ($\gamma < 0.1$), the MT is slightly more powerful than our HSAT method, followed by JLS and LRT. When

increasing γ larger than 0.1, our HSAT outperformed MT and remained the most powerful. In addition, JLS and LRT methods is less powerful than MT and later outperformed MT, while the power of MT remained decreasing with the increase of variance heterogeneity (γ). **Figure 1-5 (b)** displayed the power comparisons of the four methods for low frequency variants with a larger main genetic effect $\beta = 0.5$. Our HSAT remained the most powerful with different variance effect sizes. While when the effect of genotype on variance $\gamma < 0.3$, the JLS and LRT is less powerful than MT. When increasing γ over 0.3, JLS and LRT methods is more powerful than MT. For these two situations, the power gains of our HSAT method over the traditional mean test (MT) appeared especially noteworthy with the increase of variance heterogeneity and it did not display severe power losses with trivial or no variance heterogeneity.

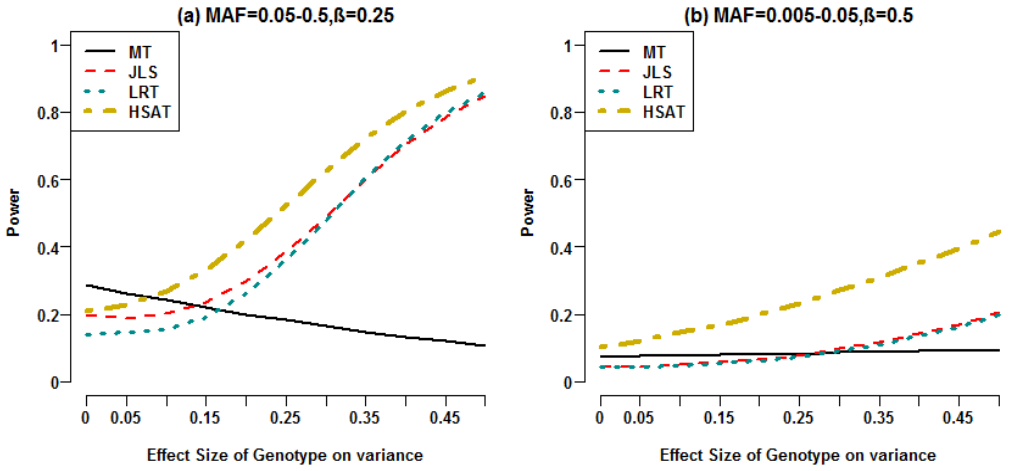

Figure 1-5: Power comparison of MT, JLS, LRT and HSAT under Scenario III at nominal level 5×10^{-8}.

1.4.3 Single-based association test in COGA study

Alcohol Dependence (AD) is a polygenic disorder that may be determined by effects of multiple variants. And there existed co-addiction between AD and other drug uses (e.g. nicotine and cocaine). For such phenotype, there is high possibility of latent $G \times E$, $G \times G$ interactions and causal biological network structures. Therefore, the effects of high-order information should not be ignored when analyzing AD and AD related diseases.

The QQ plot of mean test MT, HSAT, JLS and LRT is also presented in **Figure 1-6**, in which the inflation factor of HSAT test is 1.0121 that indicated no inflation.

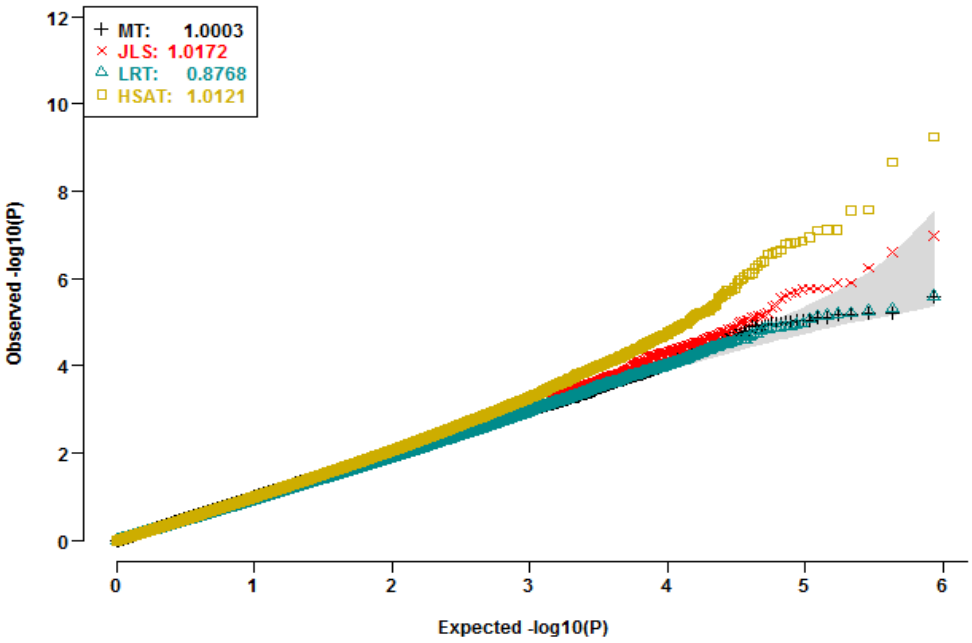

Figure 1-6: Q-Q plots of MT, JLS, LRT and HSAT.

As demonstrated by single-based association analysis, Traditional mean test ($p_{min} = 2.65 \times 10^{-6}$) did not yield a genome-wide bonferroni significant association

with nominal level $p < \frac{0.05}{856149} \approx 5.84 \times 10^{-8}$, whereas the HSAT did ($p_{min} = 5.72 \times 10^{-10}$). The Manhattan plot of MT is presented in **Figure A-1** and no association signal peaks were observed. Genome-wide significance was also not achieved by the JLS test ($p_{min} = 1.04 \times 10^{-7}$), the LRT test ($p_{min} = 2.50 \times 10^{-6}$). The Manhattan plot of HSAT is presented in **Figure 1-7**. Seen from **Figure 1-7**, we observed obvious signal peaks on chromosomes 2, 6, 7 and 19.

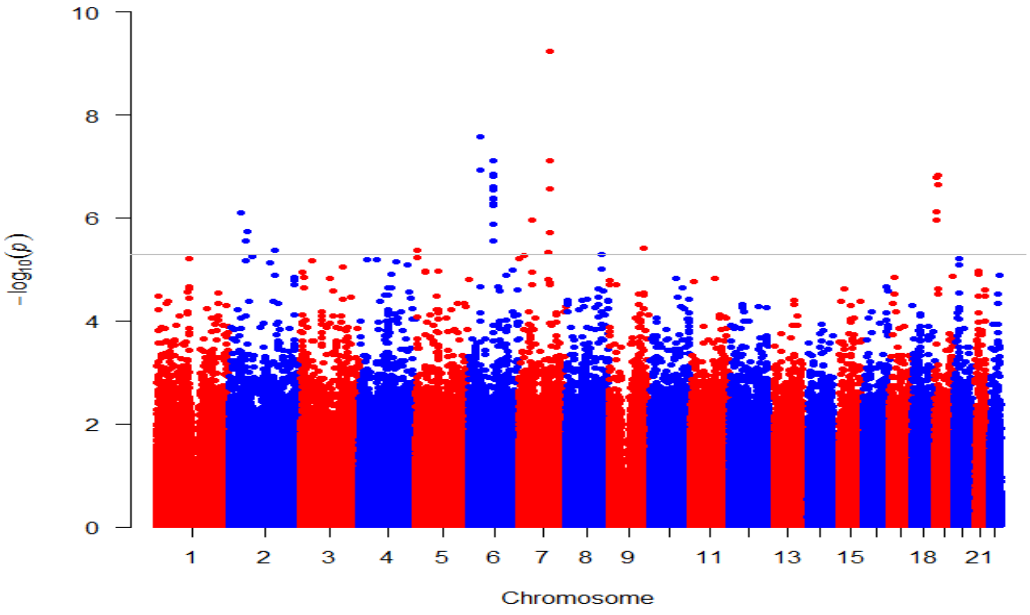

Figure 1-7: The Manhattan plot of HSAT.

Table 1-1 displayed 34 top-ranked significant SNPs selected by HSAT with suggestive genome-wide nominal level 5×10^{-6}. Three novel genes PIK3CG on chromosome 7, ZFAND3 on chromosome 6 and NFIX on chromosome 19 that contained multiple top-ranked significant SNPs are associated with AD. Specifically, the most significant detected SNP (rs849436) is only located about 32.761 kb from the PIK3CG

gene. Though PIK3CG has never been identified as candidate gene in genome association studies, its altered expression has been reported to be associated with AD[18] and cigarette smoking[19].

In addition, two previously reported genes HMGN3 and EHBP1 [20]were replicated. In particular, 11 out of the 34 top-ranked significant SNPs are located within or nearby the gene HMGN3 that is well reported in previous large-scale GWAS study [21], which lead to the strong signal peak on Chromosome 6 in **Figure 1-7**. **Table A-1** listed all detected SNPs with $p < 5 \times 10^{-5}$ and their corresponding genes, among which we reported previously replicated genes PTPRN[22-24], CNTN4[25, 26], PDLM5[27], CDH13[23, 28-30] that were related to AD in large-scale GWASs.

Table 1-1: Genome-wide Top-ranked Significant SNPs by the HSAT

Chr	rs	Position	Gene	Left Gene	Right Gene	HSAT	JLS	LRT	MT*
7	rs849436	106367588		PIK3CG	PRKAR2B	5.72E-10	1.04E-07	1.59E-05	6.77E-05
6	rs2842519	38042247	ZFAND3	MDGA1	BTBD9	2.70E-08	2.40E-04	1.32E-04	2.72E-05
6	rs1335535	79999203	HMGN3	PHIP	LOC100131959	7.86E-08	1.72E-06	3.53E-03	2.77E-02
6	rs9350803	79999595	HMGN3	PHIP	LOC100131959	7.86E-08	1.72E-06	3.53E-03	2.77E-02
7	rs849370	106307179	PIK3CG	FLJ36031	PRKAR2B	7.89E-08	1.23E-06	6.21E-05	2.68E-04
6	rs11963886	37970017	ZFAND3	MDGA1	BTBD9	1.16E-07	2.98E-04	1.70E-04	3.41E-05
6	rs1537740	80035233		HMGN3	LOC100131959	1.41E-07	2.08E-06	4.59E-03	3.16E-02
19	rs10402645	13058752	NFIX	DAND5	LYL1	1.51E-07	8.22E-05	4.91E-05	9.14E-06
6	rs7738508	80048256		HMGN3	LOC100131959	1.58E-07	1.71E-06	4.32E-03	2.04E-02
19	rs306045	2992700		TLE2	AES	1.62E-07	1.46E-05	3.85E-04	5.33E-04
19	rs11881808	13054782	NFIX	DAND5	LYL1	2.25E-07	1.98E-04	9.66E-05	1.77E-05
6	rs7763232	80030157		HMGN3	LOC100131959	2.49E-07	2.52E-06	5.85E-03	3.49E-02
7	rs849406	106320153	PIK3CG	FLJ36031	PRKAR2B	2.75E-07	7.04E-06	1.45E-04	4.23E-04
6	rs4706754	79969588	HMGN3	PHIP	LOC100131959	2.84E-07	1.75E-06	4.20E-03	8.01E-03
6	rs7772967	80051380		HMGN3	LOC100131959	4.05E-07	4.22E-06	6.73E-03	3.81E-02
6	rs10806163	79951373		PHIP	HMGN3	4.35E-07	4.46E-04	1.57E-03	7.59E-03
6	rs9343886	79983800	HMGN3	PHIP	LOC100131959	5.13E-07	2.14E-06	5.59E-03	9.96E-03
6	rs16890450	79949239		PHIP	HMGN3	5.86E-07	4.32E-04	1.59E-03	6.78E-03

19	rs1688114	2988787		TLE2	AES	7.56E-07	5.38E-05	1.08E-03	4.38E-03
2	rs17020307	37294768	CEBPZ	LOC100129418	C2orf56	8.01E-07	9.38E-02	0.1155	3.28E-02
2	rs28548299	37340278	PRKD3	C2orf56	QPCT	8.01E-07	9.38E-02	0.1155	3.28E-02
7	rs2453840	45920337	IGFBP3	IGFBP1	LOC100129619	1.08E-06	4.29E-05	1.77E-04	2.91E-04
19	rs1654678	2985077		TLE2	AES	1.11E-06	5.32E-05	1.22E-03	2.85E-03
6	rs2322219	80038110		HMGN3	LOC100131959	1.32E-06	8.68E-04	4.56E-03	3.22E-02
2	rs2421738	62877847	EHBP1	LOC100129162	LOC100132215	1.86E-06	2.71E-04	5.48E-03	0.6804
2	rs17027558	63065438	EHBP1	LOC100129162	LOC100132215	1.86E-06	271E-04	5.48E-03	0.6804
7	rs849408	106329620	PIK3CG	FLJ36031	PRKAR2B	1.88E-06	6.15E-06	1.25E-05	9.19E-04
7	rs849390	106296223	PIK3CG	FLJ36031	PRKAR2B	1.91E-06	2.51E-05	5.15E-04	1.41E-03
2	rs2871608	57499324		LOC647016	LOC100131953	2.73E-06	5.74E-05	9.10E-03	0.3063
6	rs1414283	80036646		HMGN3	LOC100131959	2.81E-06	4.63E-06	1.46E-02	2.56E-02
9	rs10982123	116050914	COL27A1	KIF12	ORM1	3.95E-06	8.18E-05	3.04E-04	1.10E-05
2	rs16829835	151831949		RBM43	NMI	4.22E-06	0.18820	0.1518	9.39E-02
5	rs159981	6042136		KIAA0947	LOC651419	4.28E-06	1.16E-04	2.68E-03	0.1612
7	rs4236534	96311548		SHFM1	LOC402679	4.75E-06	1.59E-04	4.28E-03	0.3449

* MT is the traditional mean test. The suggestive nominal level is 5×10^{-6}.

1.5 Conclusion and Discussion

Most famed marker-wise association tests are based on linear models (LMs), generalized linear models (GLMs) and generalized linear mixed models (GLMMs) with homoscedastic residuals. In such conventional homoscedastic working models, (variances of) residuals are assumed to be independent of the genetic predictors and environmental factors. The core idea of such association tests is to localize genetic determinants by exploiting linear trend (aka, correlation, association) between genetic variants and trait values. Very few exceptional methods (i.e., Soave et al.'s JLS, Cao et al.'s LRT) were developed under heteroscedastic regression models. The homoscedastic models are too simple to effectively capture high complex disease genetic mechanisms. If a test genomic marker harbors causal alleles of trait Y, the random error in the working model likely

follows a mixed distribution and is dependent on genotypic score $\boldsymbol{G}$ due to intra-marker LD and G×G, G×E interactions, etc.

Even for several existing methods considering variance effects, they integrate the association signals of mean test and variance test orthogonally and their test statistics of the mean test and variance test do not track each other at causal loci. In association analysis, integrating information will always increase the degree of freedom in the test. Failing to integrating the information efficiently would yield limited power gain and even be less powerful than traditional mean test (MT) when dispersion signal is weak relative to mean heterogeneity.

In this article, we offered novel paradigm applications for distilling and harmoniously integrating high-order information with mean effects while effectively calibrating major dispersion effects of confounders in single-based a studies. From extensive simulations above, our harmonious joint single-based method HSAT brought about dramatic association power gain in existence of low and rare frequency variants, $G \times G$ and $G \times E$ interactions and well controlled type I error rates at the same time. HSAT method includes the usual appealing features for data integration methods such as JLS and LRT and is even much more powerful than existent methods. Moreover, we have applied our method to the COGA on the alcohol addiction of 991 whites. The results demonstrate the noteworthy superiority of our method to existing tests in single-based association analyses. There are several advantages of our HSAT method.

High sensitivity: The HSAT is highly sensitive to association signals. In presence of latent $G \times E$ interaction and heteroscedasticity, it yields dramatic power gains over the famed combination tests and conventional mean heterogeneity test. In absence of real

dispersion heterogeneity, it overcomes the power loss of the famed combination methods compared to the conventional mean heterogeneity test. In particular, it displays power gains than its competitors to identify individual rare variants.

Broadness: The foundation of the HSAT is our large-sample theory. The joint asymptotic normality of its two test statistics does not require normality of the error terms. It only requires $E(e_i^4) < \infty$ and $E(e_i) = E(e_i^3) = 0$. This family covers very broad symmetric and asymmetric error distributions.

Robustness: The HSAT integrates two correlation tests, which fully inherit the core beauty of the robustness of least-squares estimates of slopes. The LRT does not apply when normality assumption on error term is severely violated (Data not shown here).

High accuracy: The test statistics have very fast rate to converge to the asymptotical joint normality. The Levene statistic adopted in JLS converges in distribution so slow that accurate approximate of its true p-value requires very large sample sizes.

Flexibility: The HSAT can be easily extended to quantitative biomarkers, i.e., gene expressions, methylations, imputation dosages, etc. It does not need any artificial partition of subjects. The LRT and the JLS rely on genotype categories to partition subjects. Partitioning subjects usually leads to power loss and other problems. For example, some categories can be too small or even empty in low-frequency and rare variant mapping.

Scalability: Our HSAT method is very efficient and stable in computation. It is hundreds fold faster than the LRT and JLS (data not shown here). This advantage is

crucial to large-sample whole-genome scan. Due to the iteration search, the LRT is not stable and may not converge to a meaningful point when the normality of residual is violated.

Lastly, we acknowledge that there existed situations where HSAT is less powerful than traditional mean test (MT) when the additional information to integrate cannot defeat the penalty induced by the increase of degree of freedom in the test. This would sometimes harm the power, as showed in Figure5 (a) as *Scenario III.* The development of more effective high-order effect integration methods requires further formal efforts. In addition, appropriate adjustment of background data structures and other hidden confounders are important for the success of effectively integrating informative high-order effects instead of spurious effects brought by environmental covariates. In real world, the high complexity of gene networks always exist in the majority of genetic datasets and the distribution of a phenotype can never be solely determined by its mean. Our HSAT method merely made one step further from existent traditional mean test and very few integrative methods. High-order effects are like hidden “gold mine” that are not exploited in existing genetic datasets and require particular methods to further distill it.

CHAPTER 2

INCOPERATING HIGH-ORDER EFFECTS CAN GAIN POWER IN GENE-BASED ASSOCIATION TESTS

2.1 Abstract

Previous studies have shown evidences that rare and common variants act collectively on disease risks. The increasing number of sequence-based association studies started to evaluate the cumulative effect of both rare and common variants on disease trait. Gene-wise association tests have been proposed to pool or collapse multiple variants in a group unit, such as a gene. Current prominent gene-based association methods of complex disease phenotypes are based on homoscedasticity working models that only aim to exploit the mean effects of variants on disease traits. As shown by real-world genetic data, the assumption of homoscedastic residuals is incompetent to account for phenotypic variation induced by the innate heterogeneous nature of the complex biological networks.

This chapter develops a harmonious novel gene-based association test (HGAT) framework of incorporating high-order effects of test markers while easily calibrating both the mean and variance effects of global covariates. High-order heteroscedasticity, which indicates genetic effects on the alteration of high-order moment of quantitative traits, may implicate potential interactions, causal networks, latent covariates, linkage

disequilibrium (LD) structure and admixture blocks that have influence on the distribution of trait values. In HGAT frame, the high-order effects of test markers are embedded as harmonious weights to better summarize the relative contribution of genes to the alteration of the distribution of phenotypes in the mean-only group association tests. By comprehensive simulation scenarios, HGAT can correctly control the type I error and outperform several existent popular gene-based association tests. We illustrate its application in homogeneous population and extend it to admixed population.

Key words:

Harmoniously integration, variance heterogeneity, $G \times E$ and $G \times G$ interactions, high-order effects, Gene-based method, admixed population

2.2 Introduction

The genome-wide association (GWAS) mainly focused on common variants and have been successful in identifying the associations of many common variants (say, MAF>5%) with human diseases such as type1 and type2 diabetes, rheumatoid arthritis, Crohn's disease and coronary heart disease[31-33]. Despite these, a large portion of the remaining heritability cannot be explained by common variants[34]. In recent years. With the advanced improvement in next-generation sequencing technology and the implementation of the 1000 Genome Project, large numbers of rare variants (SNPs) with MAF<5% have been identified accurately, which led to the consideration of rare variants as possible causal variants of human diseases to explain some missing heritability of common variants[35].

In recent years, several studies have shown evidences that rare and common variants act collectively on disease risks [36-38]. The increasing number of sequence-based association studies started to evaluate the cumulative effect of both rare and common variants on disease trait. Therefore, group-wise association tests, instead of single variant association tests, have been proposed to pool or collapse multiple variants in a group unit, such as a gene. Current prominent gene-based association methods of complex disease phenotypes are also based on homoscedasticity working models that only aim to exploit the mean effects of multiple variants on disease traits. All these models assume homoscedasticity that model residuals are independent of all predictors (covariates and variants). As discussed in Chapter 1, the assumption of homoscedastic residuals is incompetent to account for phenotypic variation induced by the innate heterogeneous nature of the complex biological networks.

Various gene-based methods have been developed. The CMC[39] method is one of the earliest and best-cited benchmarks. The phenotype is regressed on the collapsed variant score by all variants within the gene region. The most prominent sequence association tests are GLMMs based score tests, including the commonly used Sequence Kernel Association Test (SKAT)[40] and SKAT-O ('Optimal' SKAT)[41, 42]. Impelled by the assumption of rare or low frequency variants in explaining additional variability of the trait, these SKAT methods derived from burden tests and variance-component tests extensively employed a weighting scheme that up-weights the contribution of rare variants and down-weights the contribution of common variants by minor allele frequency (MAF). That is to say, such weight scheme mostly increases relative influence of rare or low frequency variants for any disease-related gene. Such a weighting scheme

can lead to loss of power when common variants in a region under investigation are also associated with disease trait. Another newly method is Mixed Effects Score Test (MiST)[43], which conducts a hierarchical model combining two independent test statistics of quantifying effect sizes of variants and annotation 'heterogeneity'.

In Chapter 2, we firstly develops a novel harmonious gene-based association test (HGAT) framework of incorporating the high-order effects of test markers within a gene region while easily calibrating both the mean and variance effects of environmental covariates in homogeneous population. High-order effect, as discussed in Chapter 1, may implicate potential high-order interactions, causal networks, latent covariates, linkage disequilibrium (LD) structure and admixture blocks among variants. Such high-order effects of test markers are embedded as better weights to summarize the relative contribution of the gene to the alteration of the distribution of disease trait beyond the change of the trait mean. By comprehensive simulation scenarios, our HGAT can correctly control the type I error and outperform several existent popular gene-based association tests. Then we employ HGAT to the same COGA on the alcohol addiction of 991 whites in Chapter 1. The results demonstrate the noteworthy superiority of our method to gene-based association tests in replicating and identifying novel susceptive genes.

Compared to homogeneous populations, much fewer association studies specifically focused on genetically admixed populations such as African Americans that comprise a substantial portion of the total population in United States. Genomes of admixed individuals derive from two or more distinct homogeneous ancestral populations. Such admixed genomes are mosaics of segments with various ancestries

(genetic origins)[44]. Local variation in ancestry (aka, local ancestry) usually indicates the number of alleles originating from reference ancestral population for each SNP of each admixed individual and reflects the region effect brought by ancestry mosaic structure. Genetic data of admixed individuals offer distinctive advantages for localizing admixture blocks that may harbor causal variants which exhibits substantially different frequencies between ancestral populations and unravel the ethnicity-specific patterns of disease prevalence. Therefore, local ancestry, in most of the time, represents the accumulating effects over the entire ancestral block in which may include certain number of variants to impact the distribution of disease traits. Statistically significant differences among high-order moment of phenotypes under different local ancestry groups may also implicate potential interactions (e.g., Ancestry×Gene and Ancestry×Ancestry), latent causal relationship among local ancestry, genotype and phenotype. In terms of admixed populations, ancestry-driven high-order effects would be non-ignored and provide additional information in traditional gene-based studies.

Therefore, we also extended our HGAT to admixed populations. High-order effect of local ancestry is included in our HGAT framework as a new weight to better summarize the relative contribution of the ancestry block to the alteration of the distribution of disease trait in admixed population. We applied HGAT to reanalysis an existing well-characterized sample of 1334 unrelated Africans selected from the Study of Addiction: Genetics and Environment (SAGE). Based on our findings, we underscore the importance of incorporating high order effects of both genotypes and local ancestry in data analysis for admixed populations.

2.3 Materials and methods

2.3.1 Model Notation and Construction of Harmonious Gene-Based Association Test (HGAT)

For the i^{th} subject, Y_i is the trait value, $\boldsymbol{X}_i$ is the $p \times 1$ vector of environmental covariate(s), $\boldsymbol{X}_i'$ is its transpose, $\boldsymbol{G}_i = (g_{i1}, g_{i2}, \dots, g_{im})'$ is the vector of copy numbers of minor alleles of m markers at the test gene. We propose a novel linear mixed model framework that harmoniously incorporate high order effects of test markers. Firstly we derive the mean-only association test as followings:

$$mo: \begin{cases} Y_i = \boldsymbol{X}_i'\boldsymbol{\gamma_1} + \boldsymbol{G}_i'\boldsymbol{\beta}_1 + e_i, \\ \boldsymbol{\beta}_1 \sim \mathcal{N}_m(0, \tau \boldsymbol{W}) \end{cases} \quad \textbf{Eq. 2-1}$$

where e_i is the random error term and $\boldsymbol{W} = diag(w_1, \dots, w_m)$ are the weights of test markers to represent the relative contribution of test markers to disease trait. $\boldsymbol{\beta}_1 = (\beta_{11}, \dots, \beta_{1m})$ are the vector of effect sizes of genotypes. $\boldsymbol{\beta}_1$ are random effects and assumed to follow multivariate normal distribution with mean zero vector and diagonal covariance-variance matrix $\tau\boldsymbol{W}$, where τ is a variance component and $\boldsymbol{\gamma_1} = (\gamma_{11}, \gamma_{12}, \dots \gamma_{1p})$ are the vector of coefficients of environmental covariates and are fixed effects. The null hypothesis for mo is $H_{0.mo}: \tau = 0 (\Leftrightarrow \boldsymbol{\beta}_1 = \boldsymbol{0})$. According to Wu et al.[40], we can obtain a score test statistic for testing $H_{0.mo}$.

$$S = \left(Y - \hat{Y}\right)'\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'\left(Y - \hat{Y}\right), \quad \textbf{Eq. 2-2}$$

where $\hat{Y}$ is the estimated value of Y under $H_{0.mo}$, in which **Eq. 2-1** collapsed to a general linear model and $\boldsymbol{G} = (\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_n})'$ is $n \times m$ genotype matrix. The score statistic S follows a mixture of chi square distributions, and p-values can be computed analytically by Davies' method under $H_{0.mo}$.

Then we adopted the following axillary high order test to calculate the weight of the k_{th} marker at the test gene as followings:

$$ho: Y_i^{*2} = G_{ik}^2 \beta_{2k} + e_i', \quad \textbf{Eq. 2-3}$$

where e_i' is the random error term, $\beta_{2\mathrm{k}}$ is the effect coefficient of the squared genotype G_{ik}^2 $(k = 1,2, \dots m)$. Y_i^* is the trait residual after calibrating both the mean and variance effects of environment covariates and β_{2k} is the effect size of genotype G_{ik} on Y_i^{*2} for each marker. The null hypothesis for ho is $H_{0.ho}: \beta_2 = 0$. The association statistic of testing $H_{0.ho}$ is

$$T = \frac{\sqrt{n-2}\, r(Y^{*2}, G^2)}{\sqrt{1 - r^2(Y^{*2}, G^2)}}, \quad \textbf{Eq. 2-4}$$

where $r(Y^{*2}, G^2)$ is the sample pearson coefficient of correlation between G^2 and Y^{*2}. The highlight of ho is to capture the second-order moment information by regressing the squared trait residual Y_i^* on the square of genotype G_i.

For a non-causal gene, all the markers within the gene region would have no effect on the distribution of disease trait. Therefore, the null hypothesis of our HGAT method is $H_{0.mh}: \tau = 0, \beta_2 = 0$. An appealing property of our HGAT method is that the independence between test statistic of mo and that of do statistic under $H_{0.md}$ and the dependence between mo and do under alternative hypothesis. As defined in Chapter 1, we call such pair of test statistics as harmonious. The proof of the null independence between mean test and high order test is provided in **Appendix A.1** and **Appendix A.2**. The advantage of such null independence in gene-based association test is that we can incorporate additional high-order effect of each marker as the new weight in mean association test **Eq. 2-1** and still control the type I error rate of the test. In addition, high-

order effect as the new weight would not increase the degrees of freedom in testing $H_{0.md}$. In this *do* test, we scale each marker in the test gene by the weight $\kappa_j = -\left(\frac{\log_{10}(p_j)}{\sum_{j=1}^{m}\log_{10}(p_j)}\right)$, where p_j is the p-value of testing $H_{0.ho}$ obtained from **Eq. 2-4**. The weight matrix for the test gene is defined as $\boldsymbol{K} = diag(\kappa_1, \dots, \kappa_m)$. The weighting scheme is rooted from our notion on harmonious statistics in Chapter 1. Each weight κ_j reflects the relative contribution of the *j*-th marker to the score statistic of the gene. Different from adopting the traditional *MAF* or external biological function information from other datasets as weight scores, we adopted the high order effect of the marker as our weight score to measure the relative importance of the marker to the alteration of trait distribution.

Then we can obtain a score test statistic for testing $H_{0.mh}$ as

$$S_{HGAT} = \left(Y^* - \widehat{Y^*}\right)'\boldsymbol{GKG}'\left(Y^* - \widehat{Y^*}\right), \qquad \textbf{Eq. 2-5}$$

where Y^* is the trait residual after adjusting for both the mean and dispersion effects of environmental covariates and $\widehat{Y^*}$ is the predicted value of Y^* obtained from **Eq. 2-1** under $H_{0.mh}$. The test statistic S_{HGAT} also follows a mixture of chi square distributions, and p-values can be computed analytically by Davies' method[45] under $H_{0.mh}$.

2.3.2 Extension of HGAT to Admixed Populations

Current gene-based statistical methods solely consider global population structure while the effect of local ancestry on analysis of rare and common variants are being ignored, especially in admixed populations (e.g., African Americans and Latinos). In admixed populations, At least two ancestral populations have been mixing for short generations to form a new population with the ancestry of each individual explained by

different proportions of the original populations. Due to recombination events, within the chromosomes of a single individual, different regions of the genome could stem from different ancestral populations. The genome of each individual can be regarded as a "mosaic" structure with segments from different ancestry[46, 47]. 1000 Genome Project[48] has shown that frequency of variants differs dramatically among different populations. Thus local ancestry in certain genomic regions in admixed populations is likely to provide additional useful information in association analysis. **Figure 2-1** indicates the relationship among local ancestry $\boldsymbol{A}$, causal and non-causal variants $\boldsymbol{G}$ ($\boldsymbol{G_{Non-causal}}$ and $\boldsymbol{G_{causal}}$) and trait value $\boldsymbol{Y}$. $\boldsymbol{X}$ is added as covariates including population structure and other environmental covariates. Single-arrow line represents possible causal direction (i.e. $\boldsymbol{A}$ to $\boldsymbol{G}$, $\boldsymbol{G}$ to $\boldsymbol{Y}$) and double-arrow line indicates correlation. Under $H_{0.mh}$, local ancestry **A** would also have no effect on the trait value Y.

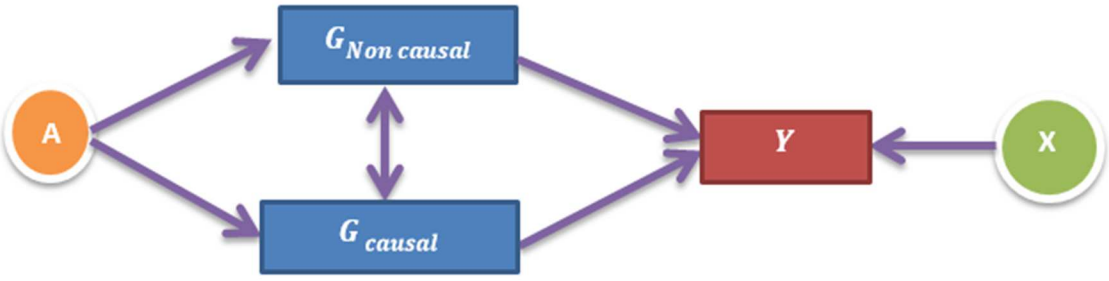

Figure 2-1: Causal graph among trait value Y, gene data X and local ancestry A.

HGAT can be easily extended to admixed populations. For recently admixed population such as African Americans, variants within one extended admixture block mostly have the same local ancestry due to sparse switch points in ancestry across a chromosome. Local variation in ancestry (aka, local ancestry) usually indicates the

number of alleles originating from reference ancestral population for each SNP of each admixed individual. Specifically, for African Americans in our study, the local ancestry is coded as 0,1 and 2 to indicate the number of alleles originating from CEPH Europeans from Utah (CEU). We let $\boldsymbol{A} = (a_1, a_2, \dots, a_n)'$ represent n × 1 vector of local ancestry for the *m* markers within the test gene. In admixed population, the score test statistic S_{HGAT_Adm} of HGAT_Adm for testing $H_{0.mh}$ is written as

$$S_{HGAT\_Adm} = \left(Y^* - \widehat{Y^*}\right)' \boldsymbol{G_{Adm}} \boldsymbol{K_{Adm}} \boldsymbol{G_{Adm}}' \left(Y^* - \widehat{Y^*}\right), \qquad \textbf{Eq. 2-6}$$

where Y^* is the same as in **Eq. 2-5** and $\boldsymbol{G_{Adm}} = (\boldsymbol{G}, \boldsymbol{A}) = \begin{pmatrix} \boldsymbol{G_1} & a_1 \\ \boldsymbol{G_2} & a_2 \\ \vdots & \vdots \\ \boldsymbol{G_n} & a_n \end{pmatrix}$ is the new $n \times (m+1)$ matrix with an adding column **A** to represent the cumulating effect of local ancestry for the test gene**.** The new weight matrix is $\boldsymbol{K_{Adm}} = diag(\kappa_1, \dots, \kappa_m, \kappa_A)$, where $\kappa_A = -\left(\frac{\log_{10}(p_A)}{\sum_{j=1}^{m} \log_{10}(p_j) + \log_{10}(p_A)}\right)$. p_A is the p-value of testing $H_{0.ho}$ obtained from **Eq. 2-3** and **Eq. 2-4** in which the genotype $\boldsymbol{G}$ is replaced by local ancestry $\boldsymbol{A}$. The test statistic S_{HGAT} also follows a mixture of chi square distributions, and p-values can be computed analytically by Davies' method under $H_{0.mh}$.

2.3.3 Simulation Configurations for Gene-Based Analysis

We considered three different experiment simulations for gene-based studies. The first two simulations come from the genetic additive model framework with different methods of determining the effect sizes of genotypes. In the last simulation, we used the previously reported gene to mimic the scenario when latent gene by environmental or gene by gene interaction existed in the dataset.

2.3.3.1 **List of Methods for Comparisons**

We conducted extensive simulations to evaluate the performance of HGAT and compare it with current commonly used methods, including CMC, SKAT, SKAT_O and MiST. The choice of MiST was in part based on the simulation results that MiST had similar better power than the burden test and SKAT-type tests under wide range of scenarios. Here we also consider DGAT which is short for Dispersion Gene Association Test in comparison. In DGAT, we also scale each marker in the test gene by the weight $\psi_j = -\left(\frac{\log_{10}(p'_j)}{\sum_{j=1}^{m}\log_{10}(p'_j)}\right)$ and corresponding weight matrix $\boldsymbol{\Phi} = diag(\psi_1, \dots, \psi_m)$, where p'_j is the p-value obtained from dispersion test in double generalized linear model (DGLM) instead of high-order model **Eq. 2-4**. In addition, we considered a weighted version for DGAT and HGAT in simulation scenarios when rarer causal variants have greater effects than common ones, and they are denoted by wDGAT and wHGAT. We adopted the default suggested weight scheme in SKAT $\omega_j = \frac{beta(f_j,1,25)}{\sum_{j=1}^{m} beta(f_j,1,25)}$ and the corresponding weight matrix $\boldsymbol{\Omega} = diag(\omega_1, \dots, \omega_m)$, where f_j is the *MAF* for the jth causal variant. For wDGAT and wHGAT, the score test statistic for testing $H_{0.mh}$ as

$$S_{wDGAT} = \left(Y^* - \widehat{Y^*}\right)' \boldsymbol{G}(\boldsymbol{\Phi} + \boldsymbol{\Omega})\boldsymbol{G}'\left(Y^* - \widehat{Y^*}\right), \qquad \textbf{Eq. 2-7}$$

$$S_{wHGAT} = \left(Y^* - \widehat{Y^*}\right)' \boldsymbol{G}(\boldsymbol{K} + \boldsymbol{\Omega})\boldsymbol{G}'\left(Y^* - \widehat{Y^*}\right), \qquad \textbf{Eq. 2-8}$$

2.3.3.2 **Simulation of LD structure of Genotypes**

For marker $s \in \{1, \dots, m\}$ at test gene, we generate minor allele frequency (*MAF*) $f_l \sim$unif(0.005,0.5) and compute the f_s-percentile c_s of the standard normal distribution $\mathcal{N}(0,1)$. And the LD between SNP *j* and *k* is generated based on the well-known exponential decay model $\rho_{jk} = \exp(-d|j-k|)$ [49-51].We then generate an $m \times m$ positive definite Toeplitz matrix $\boldsymbol{P} = (\rho_{\mathrm{jk}})$ to represent the LD structure of *m* makrers within the test gene. A vector $(z_1, \dots, z_m)'$ generated from the m-variate normal distribution $\mathcal{N}_m(\mathbf{0}, \boldsymbol{P})$ defines a haplotype $\boldsymbol{h} = (\delta(z_1 < c_1), \dots, \delta(z_m < c_m))'$, where $\delta(.)$ is the indictor of an underlying event, e.g., $\delta(z_1 < c_1) = 1$ if $z_1 < c_1$, and = 0 if $z_1 \geq c_1$. Two haplotypes $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ generated in such a way compose a genotypic vector $\boldsymbol{g} = \boldsymbol{h}_1 + \boldsymbol{h}_2 = (\boldsymbol{g_1}, \dots, \boldsymbol{g_m})'$ for the test gene.

2.3.3.3 **Homogeneous polygenic (HP) model framework**

As the foundation of genetic association studies, linkage disequilibrium (LD) occurs among tightly linked genomic markers and decays along the physical distance. LD may extend from a few kilo-bases (kb) to greater than 100kb[52-54]. Although omnipresent, the LD-driven higher-order moment information (beyond mean heterogeneities) have not been well acknowledged and exploited. Therefore we first demonstrate the noteworthy benefits and methods of exploiting LD-driven high order effects in gene-based tests. We firstly conducted simulations from a homogeneous polygenic (HP) model. The trait Y_i of subject *i* is generated by a homoscedastic residual

e_i that follows standard normal distribution, dichotomous covariate X_{1i} (i.e. $X_{1i} \sim Binom(1,0.5)$) to mimic gender, continuous covariate X_{2i} (i.e. $X_{2i} \sim \mathcal{N}(0,1)$) to mimic normalized age and genotypic scores $(G_{1i}^c, \ldots, G_{li}^c)$ at *l* causal SNPs, which randomly reside in causal gene with *m* (>*l*) markers. Following Kruglyak[55], Zhang and Stram[49], we set $d = 0.3$ to mimic moderate correlated SNPs in genome-wide SNP data. The HP model is as followings:

$$Y_i = 0.5X_{1i} + 0.5X_{2i} + \sum_{j=1}^{l} G_{ji}^c \beta_j + e_i. \qquad \textbf{Eq. 2-9}$$

Based on **Eq. 2-9**, we set $m = 50$ and $l = 10$ for 10000 replicates of 1000 subjects. The *l* causal SNPs are randomly chosen from the *m* test markers. At each causal SNP *j*, the effect size is determined by minor allele frequency MAF_j and the direction is determined by parameter c_j where $\Pr(c_j = 1) = 1 - \Pr(c_j = -1) = \pi$. "+1" indicates the positive effects and "-1" is the negative effect of the variants. The effects are determined as such because a complex trait is influenced by common and rare variants with effects of diverse sizes and directions. The effect size is written as

$$\beta_j = -0.15 c_j \log_{10}(f_j), \qquad \textbf{Eq. 2-10}$$

where f_j is the MAF of jth markers. f_j is generated by the strategy introduced in section **2.3.3.2** above.We employ the setting of β_j in **Eq. 2-10** in favor of the SKAT and SKAT-O. The 10 causal markers explain about 5% of total heritability. We adopted different π (i.e. $\pi = 1, 0.8, 0.5, 0.2$) to represent diverse directions of the markers.

2.3.3.4 **Fisher's Model Framework**

In this section, we generated the simulation model according to Fisher's theory[56] instead of the HP models that are commonly used in SKAT papers, because adopting a different simulation set up can provide us with a clear understanding of the robustness of various methods under different scenarios. We generated SNP-specific effects for the same given total heritability $h^2$of l causal SNPs. Most causal markers were of small marginal effects whereas only a small portion of causal markers were of relatively larger marginal effects. Individual trait values were generated from the following linear model

$$Y_i = \sum_{j=1}^{l} c_j G_{ji}^{c} \beta_j + e_i, \qquad \textbf{Eq. 2-11}$$

where $var(Y) = 1, e_i \sim N(0, 1 - h^2)$. β_j is set to be $\sqrt{h_j^2 / var(\boldsymbol{G_j})}$, in which $h_j^2 = h^2/l$ and $\boldsymbol{G_j}$ was genotypic vector of SNP *j*, and $var(\boldsymbol{G_j})$ was estimated from the genotype data. The genotype data is generated based on the simulation of section 2.3.3.1. c_j is the direction parameter defined in **Section 2.3.3.3**. We adopted different π (i.e. $\pi = 1, 0.8, 0.5, 0.2$) to represent diverse directions of the markers. We set $m = 50$ and $l = 10$ for 10000 replicates of 1000 subjects. In addition, we let $h^2 = 2\%$ to represent the raw proportion of phenotype that is explained by genotypes.

2.3.3.5 **Latent $\boldsymbol{G \times E}$ and $\boldsymbol{G \times G}$ interaction**

This set of simulation for power comparison mimicked a real data situation. We used the genotype data from the COGA study. 991 individuals underwent final analysis. As reported in several previous researches, OPA3 is a well replicated gene that is related

to alcohol dependence [23, 57, 58]. Therefore, we used 23 SNPs with MAF>0.005 within OPA3 as our reference to generate the dataset. Among 23 SNPs, rs811589, a well replicated SNP in previous researches, is chosen as causal SNP to generate the trait value Y as followings:

$$Y_i = 0.5 * X_{1i} + 0.5 * X_{2i} + 0.3 * X_{3i} + G_i\beta + G_iX_{3i}\delta + e_i, \qquad \textbf{Eq. 2-12}$$

where the error term e_i follows a standard normal distribution. X_{1i} is continuous normal distributed covariate (e.g. $X_{1i}\sim N(0,1)$), X_{2i} follows binomial distribution with frequency 0.5 that mimics the binary covariate such as gender. The unobserved exposure variable X_{3i} was binary variable with frequency 0.3 (e.g. $X_{3i}\sim B(2,0.3)$). The main mean genetic effect β was set to be 0.25, and the interaction effect δ of $G_i \times X_{3i}$was varied between 0 and 0.5 by a grid of 0.05. This simulation mimic the situation when the potential latent $G \times E$ interaction exist in genetic dataset. 100,00 replicates were simulated using the genotype structure of OPA3 in COGA study with sample size 991.

As illustrated in previous research, potential latent $G \times G$ interactions would lead to variance heterogeneity[11]. Next we still used rs811589 as a variance-heterogeneity quantitative trait loci (vQTL) to generate the trait value as followings:

$$Y_i = 0.5 * X_{1i} + 0.5 * X_{2i} + G_i\beta + e_i, \qquad \textbf{Eq. 2-13}$$

where $e_i\sim N(0, \exp(G_i\gamma))$, in which γ is the effect size of genotype on variance and $exp(.)$ is the exponential function which guarantees that variance of the normal distribution is always positive. In other words, γ is a parameter to measure the magnitude of effects of latent $G \times G$ interactions. β is the effect size of genotype G_i and is still set to be 0.25. X_{1i} and X_{2i} are the same as defined in **Eq. 2-12**. 100,00 replicates were simulated using the genotype structure of OPA3 in COGA study with sample size 991.

2.4 Results

2.4.1 Type I Error Control of Competitors

We compared six methods: CMC, SKAT_O, SKAT, Mist, DGAT, wDGAT, HGAT and wHGAT to evaluate the type I error control. 100,000 replicates were generated under the null model with no genetic association ($\boldsymbol{\beta} = 0$). Seen from **Figure 2-2**, CMC, SKAT, MiST, DGAT, wDGAT, HGAT and wHGAT methods generally controlled type I error rates at different given nominal significance levels, while SKAT_O is always outside of the 95% concentration band and was a little inflated at larger nominal levels. The sample size is set to be 1000.

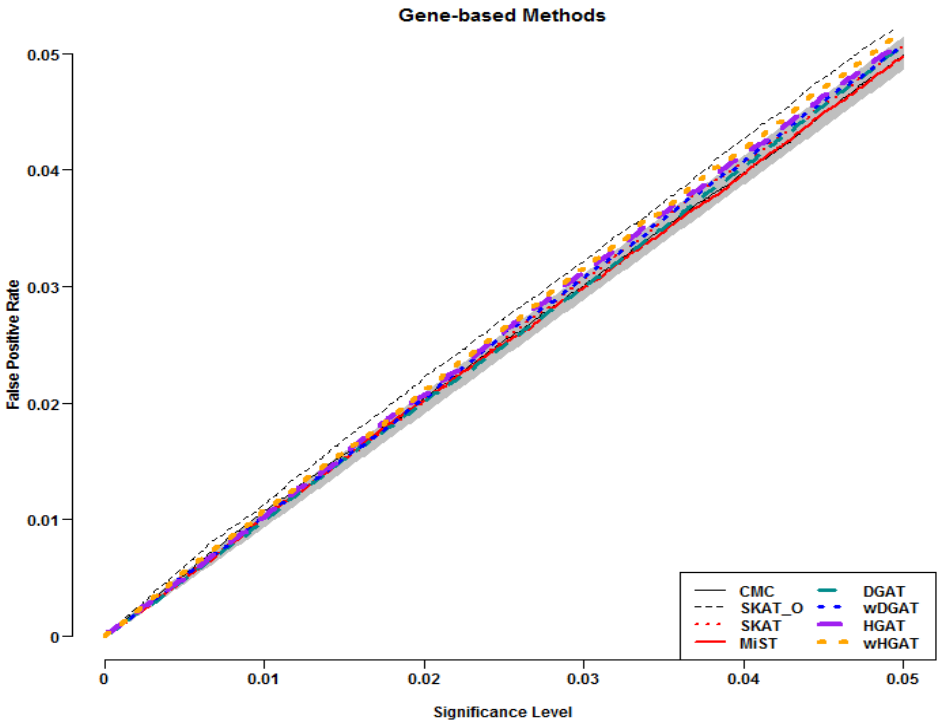

Figure 2-2: Comparison of false positive rates of eight methods under different nominal levels.

2.4.2 Empirical Power Comparisons of competitors

2.4.2.1 Power Comparisons under HP model framework

We considered HP model framework for variant effects favoring SKAT and SKAT_O methods. This scenario was to assume that rarer variants had stronger effects. The empirical power of each method was estimated by the proportion of p-values surpassed by the specified nominal significance level under alternative hypothesis among 10000 simulated data sets of 1000 unrelated individuals. All the m=50 markers in the test gene with given LD structure were genotyped according to section 2.3.3.2 and the coefficients of 10 causal loci set according to 2.3.3.3. **Figure 2-3** illustrate power comparison among different methods.

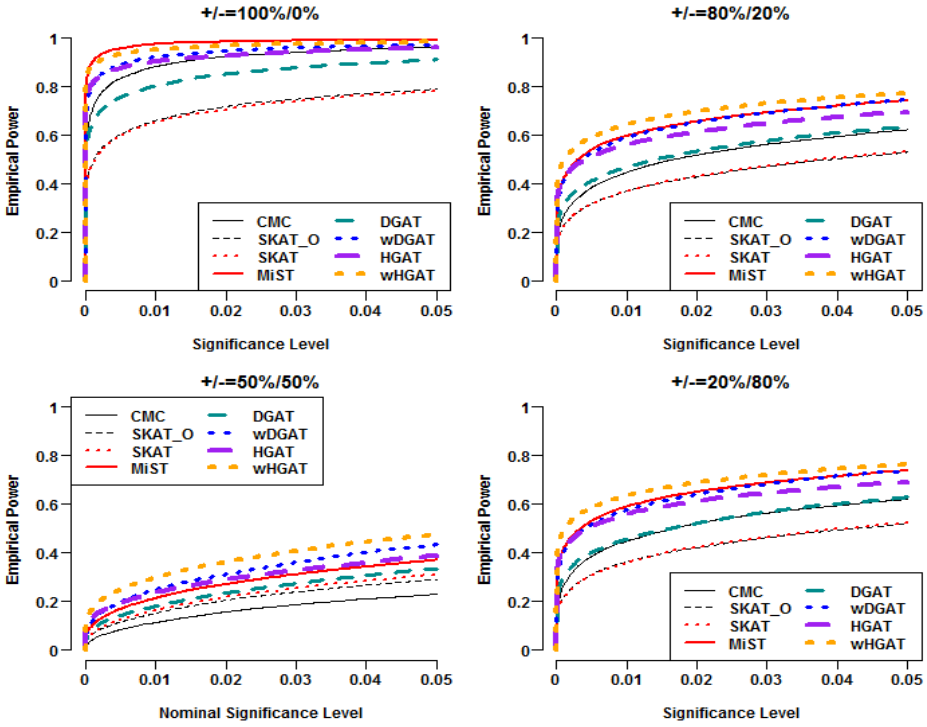

Figure 2-3: Comparison of empirical powers of eight methods at different nominal levels under HP model.

We simulated scenarios in which the effects of 10 causal SNPs are in the same or opposite directions with positive/negative=100%/0%, 80%/20%, 50%/50% and 20%/80%. The notable power gains of wHGAT and HGAT was observed in all four scenarios compared to CMC, SKAT, SKAT_O, DGAT and wDGAT. Since in this simulation rarer causal SNP have greater effects, all weighted tests wHGAT and wDGAT are more powerful than their respective unweighted tests HGAT and DGAT. When all causal loci SNPs are in the same position (i.e. +/-=100%/0%), wHGAT and MiST almost have the similar power at different nominal levels. And HGAT and wDGAT almost have the similar powers, followed by CMC method, DGAT, SKAT and SKAT_O. When the majority of causal loci SNPs are in the same position (i.e. +/-=80%/20%, 20%/80%), wHGAT is the most powerful method at different nominal levels,followed by wDGAT and MiST. HGAT is a little less powerful than wDGAT and MiST, followed by CMC. SKAT-type methods perform poorly compared to CMC because the situation of all or most of causal loci in same direction are in favor of CMC. When 10 causal SNPs are in opposite directions (i.e. +/-=50%/50%), the power of CMC drops to the lowest and SKAT, SKAT_O are more powerful then CMC. wHGAT is still the most powerful method followed by wDGAT and HGAT. MiST is a little less powerful than HGAT. For all simulation scenarios, wHGAT (or HGAT) is always more powerful than wDGAT (or DGAT), which indicate that incorporating high-order effects instead of variance effects would lead to more power gain.

2.4.2.2 **Power Comparisons under Fisher's model framework**

To show that HGATs are not only robust but also more powerful than other methods tests, we conducted another set of simulation experiments using Fisher's model frame work. In this set, we again assume that rarer variants had stronger effects but adopt a different setting for the effect size βs of genotypes. **Figure 2-4** illustrates power comparison among different methods.

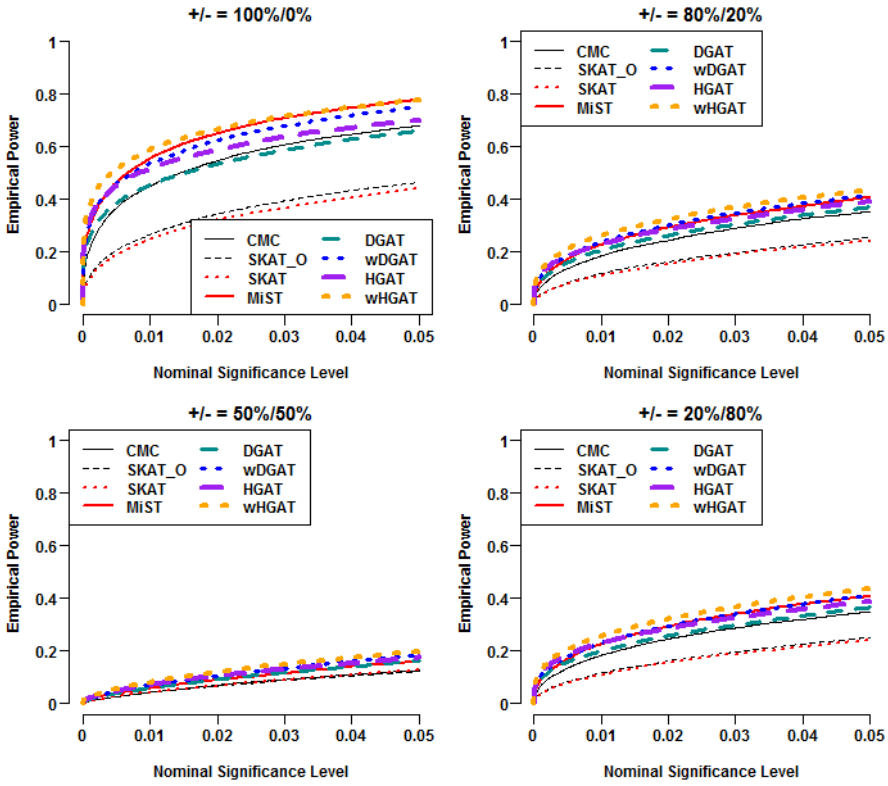

Figure 2-4: Comparison of false positive rates of eight methods at different nominal levels under Fisher's model framework.

For this simulation set that is different from HP model, SKAT-type methods have the least powers. wHGAT is still the most powerful method, followed by MiST, wDGAT and HGAT. These three methods almost have the similar powers when different directions of causal SNPs exist. When all causal SNPs are in the same direction, MiST is

a little more powerful than wDGAT and HGAT. the weighted tests (wHGAT and wDGAT) are still more powerful than their unweighted counterparts (HGAT and DGAT) because the rare or low frequency causal variants on average have small variance of genotype an this leads to larger effect size than common variants. For all simulation scenarios, wHGAT (or HGAT) is still always more powerful than wDGAT (or DGAT), which again indicates that incorporating high-order effects instead of variance effects would lead to more power gain.

2.4.2.3 **Power Comparisons with Latent G×E and G×G Interactions**

For simulation designs of G×E interactions, we utilized the real gene OPA3 in the COGA dataset to generate the trait value. We simulated scenarios in which well replicated rs811589 was chosen as the causal SNP. We compared the eight methods at nominal level 5×10^{-3} and 5×10^{-4} respectively in **Figure 2-5**. HGAT always outperformed other methods followed by DGAT. Similarly, among weighted tests, wHGAT is more powerful than wDGAT. Since in this simulation the effect size of causal SNP is not related to MAF, all weighted tests wHGAT and wDGAT are less powerful than their respective unweighted tests HGAT and DGAT. The power of the MiST is slightly higher than the wHGAT and wDGAT with small G×E interaction. Then it is surpassed by wHGAT and wDGAT with the increase of effect size δ of G×E interaction. The SKAT test has the least power and is less powerful than SKAT_O and CMC. The simulation results showed that latent G×E interaction would lead to the augmentation of high-order effects. The power gain of HGAT comes from the integration of high-order effects.

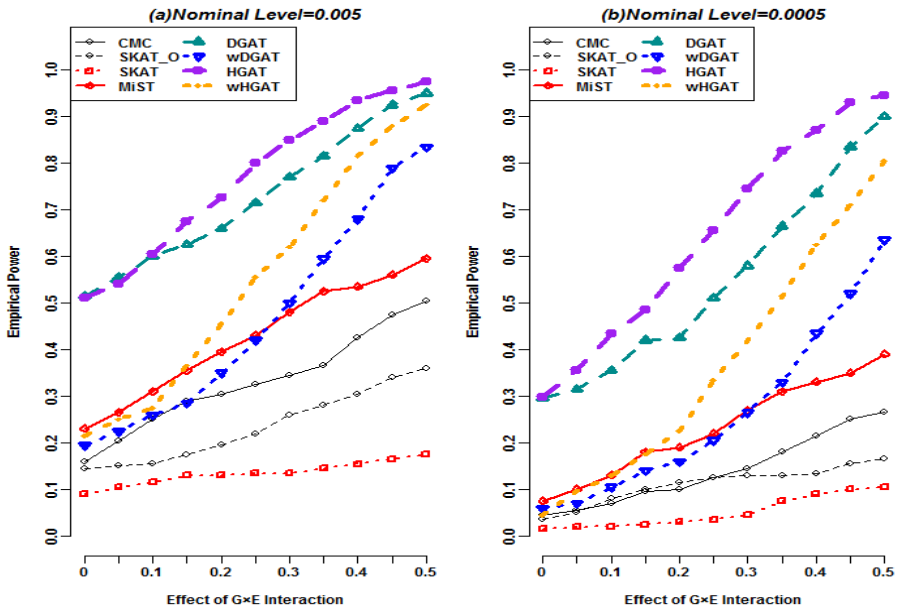

Figure 2-5: Comparison of empirical power of eight methods levels when latent G×E interaction exists at nominal level 0.005(a) and 0.0005(b).

For simulation designs of G×G interactions, we utilized the real gene OPA3 in the COGA dataset to generate the trait value. The well replicated rs811589 was chosen as the causal vQTL that has both mean and variance effect on trait value. We compared the eight methods at nominal level 5×10^{-3} and 5×10^{-4} respectively in **Figure 2-6**. HGAT still always outperformed other methods followed by DGAT. And among weighted tests, wHGAT is more powerful than wDGAT. For mean-only association methods CMC, SKAT, SKAT_O and MiST, their power would decrease with the increase of γ in **Eq. 2-13**. The SKAT test still has the least power and is less powerful than SKAT_O and CMC. The simulation results showed that latent G×G interaction would also lead to the augmentation of high-order effects. Therefore integrating high-order effect in HGAT can gain the power for detecting whether the gene has influence on the alteration of distribution of disease trait.

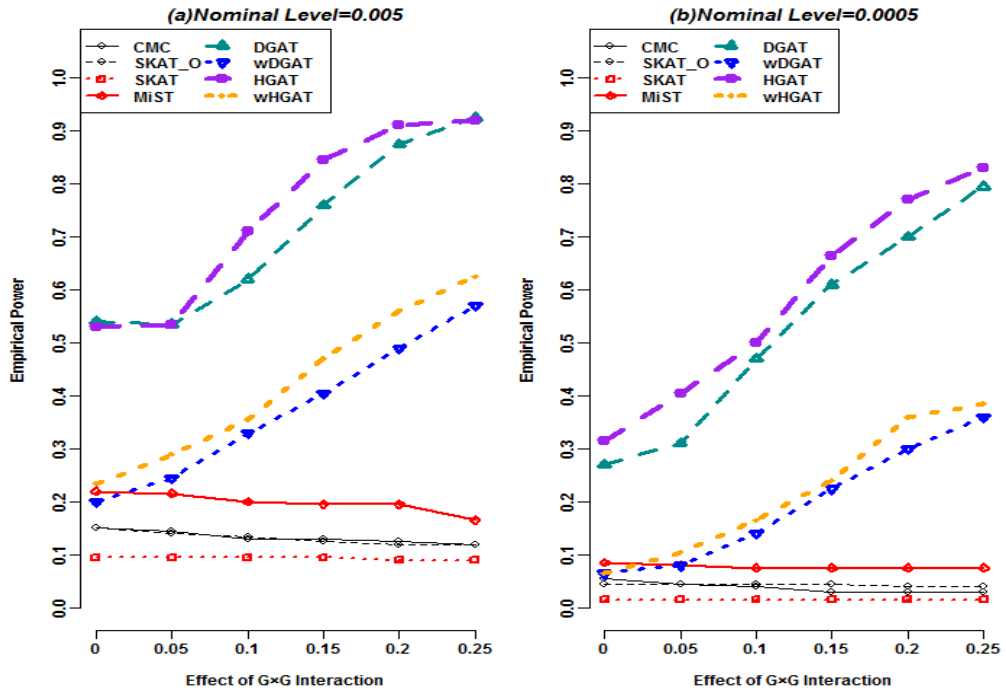

Figure 2-6: Comparison of empirical power of eight methods levels when latent G×G interaction exists at nominal level 0.005(a) and 0.0005(b).

2.4.3 Real Data Analysis on Genetics of Alcoholism (COGA) Study

We compared our HGAT and wHGAT with several popular gene-based method CMC, SKAT, SKAT_O and MiST that only focus on mean associations. In addition, we compare HGATs with DGATs that only incorporate dispersion effect instead of high-order effects. We conducted our analyses to SNPs within the 16346 gene regions. The Q-Q plots for the eight methods were shown in **Figure 2-7**, in which the inflation factor of HGAT and wHGAT are 1.0307 and 1.0552 respectively. HGATs methods indicated no inflation. Due to the relatively small sample size (n=991), none of the methods reached Bonferroni gene-wise statistical significance ($p < \frac{0.05}{16346} = 3.06 \times\times 10^{-6}$).Setting

suggestive gene-wise nominal level (i.e.5×10^{-4}), we identified 17 significant genes (i.e either p_{HGAT} or p_{wHGAT} <5×10^{-4}), among which PTPRN[22-24] and PDLIM5[27] are well replicated genes related to AD in previous GWAS studies. Among the 17 top ranked genes, gene expression of ACTN2[59] was associated with alcohol-related traits and IGFBP3[60] was related to alcohol-induced liver disease. In addition, the gene expression of UEVLD[61] is related to alcohol exposure. EMILIN2[62] and DEFA4[63] have association with smoking that has high correlation with alcohol dependence. The top-ranked genes are listed in **Table 2-1**.

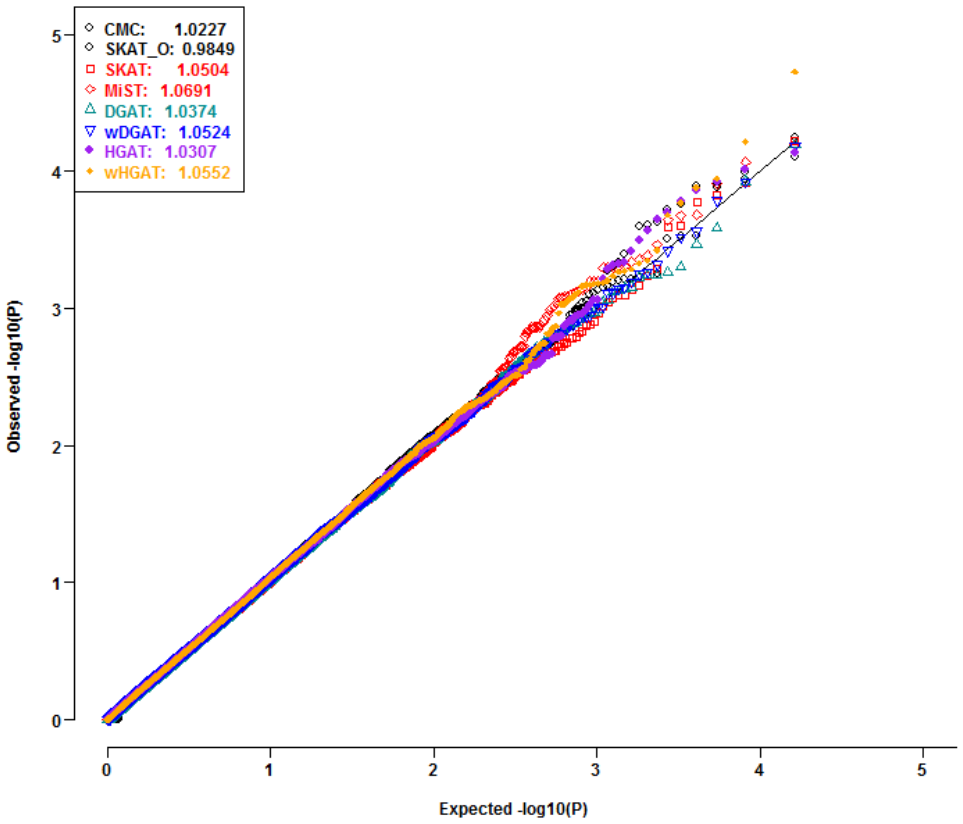

Figure 2-7: Q-Q plots of eight gene-based methods

Table 2-1: Top-ranked Significant Genes by HGAT or wHGAT

Chr	Gene	HGAT	wHGAT	DGAT	wDGAT	MiST	SKAT	SKAT_O	CMC
2	PTPRN	1.50e-05	1.35e-05	3.43e-04	2.69e-03	6.42e-05	0.049	0.063	4.03e-03
1	STXBP3	7.34e-05	6.10e-05	1.43e-02	6.47e-05	4.40e-04	5.95e-05	5.66e-05	0.603
8	DEFA4	9.69e-05	5.11e-03	3.43e-03	0.025	0.098	0.043	0.045	0.136
4	LSM6	1.19e-04	1.37e-03	0.154	0.039	0.034	0.039	0.036	0.045
21	SFRS15	1.37e-04	1.13e-04	0.117	0.052	2.11e-04	0.022	0.036	0.016
6	ZFAND3	1.64e-04	1.32e-04	0.025	9.03e-03	4.06e-03	2.09e-03	3.70e-03	0.491
7	IGFBP3	2.02e-04	2.08e-04	1.62e-03	1.52e-03	0.086	0.431	0.456	0.049
19	NFIX	2.26e-04	4.47e-04	0.073	0.036	6.33e-04	3.06e-03	3.63e-03	0.385
9	BSPRY	2.72e-04	6.24e-04	2.6e-04	1.19e-03	0.019	0.188	0.312	0.015
1	ACTN2	3.21e-04	6.77e-04	0.024	0.030	0.020	0.135	0.196	0.014
11	TSG101	3.85e-04	9.35e-04	5.75e-04	1.43e-03	0.030	0.665	0.356	0.013
11	UEVLD	4.67e-04	5.79e-04	5.91e-04	7.72e-04	0.282	0.897	1	0.119
18	EMILIN2	4.76e-04	1.78e-03	8.95e-04	9.94e-04	0.252	0.227	0.351	0.266
4	PDLIM5	4.83e-04	6.47e-04	3.32e-05	4.55e-05	0.100	0.879	1	0.024
7	GUSB	5.15e-04	1.88e-05	6.37e-05	1.63e-04	9.64e-04	2.55e-04	2.52e-04	0.036
2	DNAJB2	6.68e-03	3.74e-04	0.017	5.64e-04	8.45e-04	2.50e-04	2.49e-04	3.00e-04
12	MBD6	0.0127	1.69e-04	1.18e-04	1.19e-04	4.10e-04	1.21e-04	1.14e-04	0.012

The suggestive nominal level is 5×10^{-4}.

We selected 24 previous reported genes that have been implicated as candidate genes related to AD by more than one GWAS or sequencing paper. The results are shown in **Table 2-2.** For the majority of previous reported genes, our HGATs methods (HGAT or wHGAT) obtained smaller p-values compared to other methods.

Table 2-2: P values of 24 previous replicated genes in COGA dataset

Chr	Gene	HGAT	wHGAT	DGAT	wDGAT	MiST	SKAT	SKAT_O	CMC
1	OLFM3[64-66]	0.386	0.353	0.276	0.264	0.347	0.342	0.496	0.387
1	TNN[64, 67]	0.387	0.702	0.444	0.750	0.972	0.961	0.846	0.772
1	NRD1[58, 64, 65]	0.497	0.245	0.629	0.303	0.177	0.048	0.063	0.848
2	THSD7B[68-70]	6.38e-03	0.010	0.033	0.015	0.211	0.269	0.425	0.186

3	CNTN4[25, 71, 72]	7.23e-03	2.10e-03	0.088	0.031	0.079	0.072	0.129	0.125
4	ADH1C[73-75]	0.151	0.033	0.206	0.077	0.0748	0.0343	0.0318	0.414
5	DOCK2[21, 30]	0.075	0.059	0.336	0.176	0.377	0.145	0.238	0.818
5	PPP2R2B[64, 65, 76, 77]	0.567	0.476	0.811	0.667	0.529	0.307	0.271	0.785
6	SYNE1[57, 68, 78, 79]	0.336	0.491	0.344	0.499	0.703	0.673	0.666	0.527
7	CNTNAP2[30, 71]	0.594	0.374	0.632	0.375	0.262	0.100	0.160	0.800
8	CSMD1[21, 29, 77, 78, 80]	0.196	0.402	0.336	0.491	0.887	0.875	0.742	0.682
10	KCNMA1[30, 57, 81]	0.646	0.725	0.722	0.795	0.916	0.832	0.948	0.768
10	HTR7[30, 57]	0.273	0.134	0.367	0.173	0.074	0.063	0.110	0.293
11	TTC12[57, 82, 83]	0.011	0.176	0.022	0.205	0.884	0.785	0.483	0.806
11	PKNOX2[68, 84]	0.709	0.371	0.725	0.312	0.100	0.089	0.131	0.088
11	NAP1L4[57, 64]	0.686	0.611	0.818	0.711	0.593	0.506	0.308	0.502
11	GRM5[64, 85]	0.318	0.304	0.405	0.384	0.211	0.328	0.397	0.122
12	ITPR2[21, 72]	0.587	0.901	0.495	0.878	0.952	0.941	0.832	0.745
12	SOX5[23, 57]	0.304	0.706	0.451	0.812	0.905	0.961	0.982	0.621
12	ALDH2[86, 87]	0.508	0.628	0.575	0.696	0.994	0.926	0.869	0.897
12	SLC2A14[57, 88]	0.663	0.838	0.724	0.870	0.973	0.814	0.743	0.734
13	SLC10A2[21, 57, 68]	0.378	0.467	0.365	0.445	0.855	0.640	0.832	0.909
18	CCBE1[23, 57, 89]	0.088	0.102	0.141	0.182	0.628	0.428	0.453	0.619
19	OPA3[23, 57, 58, 64]	2.40e-03	0.085	4.65e-03	0.094	0.908	0.813	0.743	0.733

2.4.4 Real Data Analysis on Study of Addiction: Genetics and Environment (SAGE)

In this section, we applied HGAT_Adm method to re-analyzed a large, well-characterized sample of 1334 unrelated individuals from the Study of Addiction: Genetics and Environment (SAGE). It contains 942048 SNPs. Positions of all SNPs are genome build 36.3. The primary phenotype is DSM-IV AD.

2.4.4.1 **Genotype Quality Control, Local Ancestry Inference and Estimation of Global Ancestry**

SAGE dataset contains 942048 SNPs. Positions of all SNPs are genome build 36.3. The primary phenotype is DSM-IV AD. SNPs were excluded if minor allele frequency (MAF)$<5\%$ or call rates$<95\%$, leaving 917,681 SNPs after genotype quality control. Among 1334 individuals, 69 were excluded due to missing or extreme trait values. After data cleaning, 1265 individuals (48.6% males; 39.9 ± 7.3 years) underwent final analysis. We didn't remove SNPs that violates Hardy-Weinberg Equilibrium (HWE) because the presence of admixture often violates the assumptions of HWE. Simply removing such SNPs would lose the ancestry information.

We inferred local ancestries of 1265 unrelated African American genomes at non-overlapped adjacent windows using the ELAI package[90]. Reference panels of HapMap West African Yoruban (YRI) and CEPH Europeans from Utah (CEU) genotypes (http://hapmap.ncbi.nlm.nih.gov/) are download from International HapMap Project. Processed by Plink, genotypes were coded as 0, 1 and 2 to represent the count number of the minor allele. The ancestry states (dosages) of each SNP of an admixed individual obtained by ELAI are then recorded to 0, 1 and 2 to indicate the number of alleles originating from CEU. After local ancestry inference, 860,427 SNPs are maintained with both local ancestry and genotype information in final analysis.

One additional important component we need to consider in admixed population is global ancestry that represents ancestral proportions averaged across the whole genome of an admixed individual. It is usually adjusted as covariate representing population stratification. Suggested by Price et al.[91], the top eigenvectors are shown to be effective

in capturing the demographic uniqueness of a population. Thus global ancestries are often estimated by top principal components (PCs) of genotypic matrix of a subset of all genotyped markers. To verify our local ancestry estimation, we also estimated the global ancestry of a subject by averaging the inferred local ancestries using ELAI across the genome of the subject and calculated the correlation between estimated global ancestry and the first principal component (PC). The accuracy of local ancestry inference is verified by the extremely high correlation ($r^2 > 0.9974$) between estimated global ancestry and the first PC calculated from genotypes of admixed individuals. This indicates the high accuracy of inference of local ancestry using ELAI. Estimated global ancestry as covariate was included in our final data analysis to capture the population stratification.

2.4.4.2 **Adjustment of Covariates**

Following genotype quality control and local ancestry inference above, we applied the double generalized linear model (DGLM) to adjust for both mean and variance effects of covariates. The DGLM is implemented in R package dglm. The covariates to adjust for in analysis are gender (1=Male, 2=Female), smoking (0~7), normalized age, squared-normalized age and estimated global ancestry. Since age ranges from 18 to 64, normalizing age can reduce the difference of age profiles. Adding the square of normalized age allows you to model the effect of age that may have a non-linear relationship with the phenotype AD. The inclusion of smoking was to remove possible spurious results caused by effects of smoking considering the moderate relationship ($r^2 = 0.4874$) between drinking and smoking. From **Table 2-3**, we observed

significant (P<0.05) dispersion effects of gender, smoking, normalized age (Age*), squared-normalized age and estimated global ancestry (Global), which implicate the necessity of correcting for the effects of heteroscedasticities for covariates.

Table 2-3: Separate analyses of drinking symptom

Mean			Dispersion		
Effects	Estimate	p-values	Effects	Estimate	p-values
Intercept	21.1041	7.4e-28	Intercep	6.0555	5.0e-28
Smoke	3.4242	7.8e-50	Smoke	0.2968	2.1e-54
Gender	-8.2967	1.6e-19	Gender	-0.9339	3.1e-31
Age*	-0.0829	0.8166	Age*	-0.0905	0.0236
$(Age^*)^2$	-0.2848	0.1826	$(Age^*)^2$	-0.1046	3.1e-05
Global	5.0718	0.2778	Global	1.5023	8.9e-04

After adjusting for both mean and variance effects, covariates have no significant effects on AD (**Table 2-4**). Covariates adjustment does not remove the ancestry and SNP information on both the mean and variance of phenotype. And hence we can estimate local ancestry effects and genotypic effects on this adjusted trait residual after removing the effects of covariates on both mean and variance.

Table 2-4: Separate analyses of drinking symptom after adjustment

Mean			Dispersion		
Effects	Estimate	p-values	Effects	Estimate	p-values
Intercep	0.0315	0.7903	Intercep	-0.0129	0.9387
Smoke	0.0017	0.8983	Smoke	-0.0016	0.9317
Gender	-0.0351	0.5368	Gender	0.0090	0.9112
Age*	0.0107	0.7066	Age*	0.0005	0.9906
$(Age^*)^2$	-0.0114	0.5210	$(Age^*)^2$	0.0009	0.9713
Global	0.1648	0.6065	Global	-0.0035	0.9938

2.4.4.3 **Replication of previous highlighted genes for alcohol dependence**

We compare our methods HGAT_Adm and its weighted version wHGAT_Adm with the commonly used mean-only gene based association methods CMC, SKAT,

SKAT_O and SKAT. The statistic S_{wHGAT_Adm} of wHGAT_Adm has the same setting as defined in **Eq. 2-8** by replacing K with K_{Adm}. In addition, we also consider CMC_adj, SKAT_adj and SKAT_O_adj that adjust for gene-wise local ancestry as covariate. In SAGE data analysis, we selected 26 previous reported genes that have been implicated as candidate genes related to AD by more than one GWAS or sequencing paper. The results of the 26 replicated genes are listed in **Table 2-5.**

Table 2-5: P values of 26 previous reported genes in SAGE dataset

Chr	Gene	HGAT_Adm	wHGAT_Adm	SKAT	SKAT_adj	SKAT_O	SKAT_O_adj	CMC	CMC_adj
1	OLFM3	0.047	0.024	0.131	0.232	0.203	0.311	0.569	0.373
1	TNN	0.118	0.273	0.899	0.959	0.936	0.968	0.394	0.505
1	NRD1	0.244	0.054	0.016	0.013	0.028	0.024	0.743	0.686
2	THSD7B	5.82e-04	6.23e-03	0.715	0.672	0.218	0.064	0.759	0.843
2	MREG	8.47e-03	6.16e-03	0.240	0.209	0.374	0.336	0.399	0.386
3	BBX	0.666	0.869	0.981	0.958	0.799	0.551	0.582	0.422
3	CNTN4	1.60e-03	4.84e-03	0.953	0.931	0.700	0.927	0.453	0.854
4	ADH1C	0.049	0.131	0.419	0.405	0.289	0.268	0.154	0.152
5	DOCK2	7.14e-04	3.11e-04	0.025	0.134	0.038	0.233	0.074	0.806
5	PPP2R2B	0.025	0.093	0.609	0.549	0.809	0.762	0.138	0.133
6	SYNE1	0.085	0.124	0.799	0.640	0.448	0.015	0.356	0.173
7	CNTNAP2	4.51e-06	1.23e-04	0.739	0.584	0.835	0.816	0.759	0.848
8	CSMD1	3.40e-07	6.26e-06	0.969	0.949	0.979	0.954	0.566	0.555
10	KCNMA1	2.04e-03	1.59e-03	0.326	0.248	0.491	0.398	0.610	0.540
10	HTR7	0.066	0.056	0.471	0.460	0.146	0.229	0.228	0.218
11	TTC12	0.038	0.094	0.589	0.536	0.784	0.740	0.989	0.817
11	PKNOX2	6.44e-03	3.78e-03	0.175	0.071	0.270	0.127	0.694	0.603
11	NAP1L4	0.064	5.47e-03	1.75e-03	8.92e-04	3.34e-03	1.74e-03	0.287	0.213
11	GRM5	0.145	0.183	0.633	0.574	0.836	0.794	0.479	0.472
12	ITPR2	3.55e-04	6.21e-04	0.239	0.141	0.359	0.235	0.111	0.477
12	SOX5	0.070	0.048	0.309	0.206	0.471	0.356	0.532	0.689
12	ALDH2	0.032	0.028	0.256	0.205	0.105	0.114	0.048	0.056
12	SLC2A14	0.028	0.071	0.723	0.777	0.839	0.957	0.239	0.154
13	SLC10A2	0.034	0.051	0.115	0.083	0.194	0.142	0.104	0.174
18	CCBE1	0.145	0.439	0.918	0.811	0.984	0.627	0.623	0.418

19	OPA3	0.046	0.049	0.304	0.292	0.415	0.395	0.543	0.841

We investigated the genetic variants in these 26 well replicated genes previously shown to be associated with AD. For the majority of previous replicated genes, our HGAT_Adm's methods obtained smaller p-values compared to other methods. The better performance of our proposed method came from integrating the additional high-order effects of local ancestry in the gene-based association test. Utilizing local ancestry instead of calibrating it can provide additional useful additional information regarding the source of cumulative effect of admixture block region on disease trait.

2.5 Conclusions and Discussions

High-order effect, as discussed in Chapter 1, may implicate potential high-order interactions, causal networks, latent covariates, linkage disequilibrium (LD) structure and admixture blocks among variants. The novel principle of harmonious tests have also been introduced in detail in Chapter I. For Chapter II, we apply such harmonious principle to propose the novel HGAT and HGAT_Adm method for distilling and harmoniously integrating high-order information of genotype and local ancestry in gene-based studies. Such high-order effects of test markers are embedded as better weights to summarize the relative contribution of the gene to the alteration of the distribution of disease trait beyond the change of the trait mean.

There are several advantages to HGAT modeling. The statistic that we developed for HGAT has the appealing features of the score test in linear mixed models such as calculating p-value analytically and saving computation time compared to permutation. The axillary high order test is generated to capture the high-order effects. Due to the

independence of *mo* and *ho* test under null hypothesis, our HGATs methods can control the type I error. Due to the dependence of *mo* and *ho* test under alternative hypothesis, HGATs outperformed commonly used existent popular gene-based association tests.

For admixed population, local ancestry offer additional information resource in terms of the ethnicity-specific patterns of disease prevalence. In other words, local ancestry represents the accumulating effects over the entire ancestral block in which may include certain number of variants to impact the distribution of disease traits. Therefore, statistically significant differences among high-order moment of phenotypes under different local ancestry groups may also implicate potential interactions (e.g., Ancestry×Gene and Ancestry×Ancestry), latent causal relationship among local ancestry, genotype and phenotype. Therefore, we also extended our HGAT to HGAT_Adm in admixed populations by including high-order effect of local ancestry in HGAT framework as a new weight to better summarize the relative contribution of the ancestry block to the alteration of the distribution of disease trait in admixed population.

By application to COGA and SAGE datasets, we demonstrate the noteworthy superiority of HGAT methods to existent gene-based mean-only association tests in replicating and identifying novel susceptive genes. The development of more effective high-order effect integration methods requires further formal efforts. In addition, appropriate adjustment of both mean and variance effects of covariates are important for the success of effectively integrating informative high-order effects instead of spurious effects brought by environmental covariates.

CHAPTER 3

INTEGRATING MEAN AND HIGH-ORDER HETEROGENEITIES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES

3.1 Abstract

Identifying differentially expressed (DE) genes with distinct mean expression levels between different experimental conditions is a main challenge in functional genomics studies. Mean heterogeneity, namely the difference between condition-specific means of gene expression levels, reflects one aspect of its distribution alteration. If a DE gene is a genuine functional gene that involve in gene-gene co-expression and interaction networks related to the disease, its distribution change in the expression level cannot be solely completely determined by mean heterogeneity. Higher-order heterogeneities, namely the difference between condition-specific high-order moment beyond the first order moment (i.e. the mean), can provide extra valuable information for describing the distribution change of expression levels. There are two parts in this chapter. For Part I, I firstly introduced our published integrative mean-variance test (IMVT) that combined gene-wise mean heterogeneity and variance heterogeneity. For moderate samples, the IMVT well controlled type I error rates and outperformed its competitors under comprehensive simulations of normality and Laplace settings. In presence of variance heterogeneity, the IMVT appeared noticeably more powerful than all the mean heterogeneity tests. For Part II, a novel double Welch *t* test (DWT) was proposed to capture both mean heterogeneity and second-order heterogeneity instead of variance heterogeneity. The DWT outperformed our earlier IMVT method and also well controlled type I error rates. Both IMVT and DWT methods were applied to the gene profiles of peripheral

circulating B. After adjusting for background data structure, IMVT replicated previous discoveries and identified novel experiment-wide significant candidate functional DE genes. And we also compared the results of IMVT and DWT in replication of previous reported genes. Our results indicate tremendous potential gain of integrating informative high-order heterogeneity after adjusting for global confounders and background data structure. Therefore, particular attention should be paid to explicitly exploit the high-order heterogeneity induced by condition change in functional genomics analysis.

Key words: Functional genomics studies; DE genes; High-order heterogeneities; Latent confounders; Latent biomarkers

3.2 Part I: Integrating Mean and Variance Heterogeneities to Identify Differentially Expressed Genes

3.2.1 Introduction

Typically the core challenge in comparative microarray experiments is to identify statistically significant genes of biologically meaningful changes in expression levels under different conditions. Differentially expressed genes may help identify disease biomarkers that are important for the diagnosis of multiple diseases [92, 93]. There are several existent mean heterogeneity tests for identifying differentially expressed genes. The Student *t* test (ST) has been widely applied as a standard routine for identifying mean differentially expressed (MDE) genes in two-condition experiments [94]. The null hypothesis of this test is mean homogeneity H_{01}: the testing gene has identical mean expression level under the two conditions. It assumes variance homogeneity H_{02}: the testing gene has identical variance in expression level under the two conditions. The necessity of H_{02} for the ST was

formally examined under normality setting [95]. It tends to inflate type I error rate for rejecting mean equality if the smaller sample is from the population with the larger variance. In contrast, it tends to be conservative if the larger sample is from the population with smaller variance. The WT [96] is an adaptation of the ST to allow for potential variance heterogeneity between two experimental conditions. This test calibrates potential variance heterogeneity as an impediment to identify differentially expressed genes. Demissie et al. developed the MWT [97] to obtain more stable estimates of the error variance of a gene in a low-replicate microarray experiment. The MWT outperformed the Welch test to allow for variance heterogeneity. All aforesaid tests either simply ignore or take the variance heterogeneity as an impediment and calibrate it when identifying differentially expressed genes.

For a gene in a complex network, its distribution heterogeneity of expression levels can include heterogeneities in mean, variance, and even higher-order mathematical characteristics. Thus far, researchers have been conventionally focusing on exploiting mean heterogeneity, simply ignoring or adjusting for overall intra-condition variance heterogeneity. Herein, we distinguish ‘informative component’ from ‘impediment component’ of the overall variance heterogeneity. Specifically, we call the variance heterogeneity due to condition change as ‘informative variance heterogeneity’; and call variance heterogeneity due to environmental covariates and latent factors (i.e., background data structure) as impediment variance heterogeneity. However, informative variance heterogeneity has not been well recognized and exploited. Informative variance heterogeneity of a susceptible gene can capture extra information conveyed by complicated biological networks. High gene-gene correlations are common in co-expression networks

of differentially expressed genes [98, 99]. Genes can interact with each other and/or interact with environmental factors. Therefore, the alteration of expression distribution of a susceptible gene cannot be completely determined by its mean heterogeneity. Heterogeneities of high-order characteristics, e.g., variance and kurtosis, can provide extra valuable information. Exploiting informative mean heterogeneity of gene expression level alone would be incompetent to extract the information of the second-order moment (i.e., the variance). Existent methods cannot explicitly integrate the informative variance heterogeneity of gene expressions due to condition change; and little has been done to distill informative variance heterogeneity.

In Part I, we put forth mean-variance differentially expressed (MVDE) gene as a novel concept. The family of MVDE genes is broader than that of conventional MDE genes. It goes one step closer to our generic concept of a susceptible gene − a gene displays reliable changes in any aspects of the entire distribution of its expression level with the change in condition. A MVDE gene may display different means and/or variances of expression levels between two different conditions. The proper null hypothesis of testing MVDE is $H_{03} = H_{01} \cap H_{02}$: the gene has equal mean and equal variance of expression levels between the two conditions. We reject the dual null hypothesis (H_{03}) and claim the testing gene. Under normality setting, the two-sample *F*-test is the most powerful procedure for exploiting variance heterogeneity. But the *F*-test is very sensitive to the violation of normality [100]. Beyond normality setting, the Levene test [101] and the Brown–Forsythe test [102] are two popular alternatives for inspecting variance heterogeneity.

We mathematically proved and empirically illustrated that testing statistics of mean heterogeneity and variance heterogeneity are independently distributed under H_{03}. This null independence is not well-known to many, but is crucial to assure the type I error rate control of the IMVT using Fisher's method [102]. Under comprehensive simulations, the IMVT appeared noticeably more powerful than existent mean heterogeneity tests (i.e., WT, MWT and STSD) as well as the LRT and the SMVT for identifying MVDE genes. In particular, the IMVT appeared strikingly more powerful than the mean heterogeneity tests to identify genes with variance heterogeneity. To illustrate the practical utility of our IMVT, we reanalyzed the gene profiles of peripheral circulating B cells [103] after adjusting for global confounders and background data structure. Our IMVT replicated previous discoveries and identified novel genes that were missed by existent mean heterogeneity tests.

3.2.2 Methods

Let the dataset contain expression levels of M gene probes of n_c unrelated subjects from condition c (i.e., $c = 1$ for control group, and $c = 2$ for treatment group). To be specific, let G_{ijc} be the expression level of gene probe i $(= 1,2, \dots, M)$ on subject j $(= 1,2, \dots, n_c)$ under condition c, and let $n = n_1 + n_2$ be the total sample size. Let μ_{ic} and σ_{ic}^2 be the gene-specific mean and variance of the expression levels of gene probe i under condition c, respectively. The standard unbiased estimators of μ_{ic} and σ_{ic}^2 are given by $\hat{\mu}_{ic} = \bar{G}_{ic} = \sum_{j=1}^{n_c} G_{ijc} / n_c$ and $\hat{\sigma}_{ic}^2 = \sum_{j=1}^{n_c} \left(G_{ijc} - \bar{G}_{ic}\right)^2 / (n_c - 1)$, respectively.

3.2.2.1 **Concept of MDE genes and mean heterogeneity tests**

Researchers conventionally focus on identifying MDE genes. A MDE gene displays mean differentials between the expression levels under two experimental conditions ($\mu_1 \neq \mu_2$). The ST has been widely used routine to identify MDE genes. This mean heterogeneity test rejects the null hypothesis $H_{01}: \mu_1 = \mu_2$ if the Student statistic of the testing gene departures from zero significantly. A default assumption behind the ST is variance equality$H_{02}: \sigma_1^2 = \sigma_2^2$ at the testing gene. Specifically, for the i^{th} gene, let $\boldsymbol{G_1} = \left(G_{i11}, G_{i21}, \dots, G_{in_1 1}\right)'$ and $\boldsymbol{G_2} = \left(G_{i12}, G_{i22}, \dots, G_{in_2 2}\right)'$ be the expression levels of two independent random samples from normal populations $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$ and $\mathcal{N}(\mu_{i2}, \sigma_{i2}^2)$, respectively. The ST on $H_{01}^{(i)}: \mu_{i1} = \mu_{i2}$ assumes variance homogeneity ($H_{02}^{(i)}: \sigma_{i1}^2 = \sigma_{i2}^2$) between the two conditions, and defines the test statistic as

$$\hat{t} = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-\frac{1}{2}} (\hat{\mu}_{i1} - \hat{\mu}_{i2})}{\sqrt{\hat{\sigma}_p^2}},$$

where $\hat{\sigma}_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i1}^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i2}^2$ is the pooled sample variance estimator of the common variance σ^2. If $H_{03}^{(i)} = H_{01}^{(i)} \cap H_{02}^{(i)}$ is true, then the testing statistic $\hat{t}$ follows the centralized Student *t* distribution with $(n_1 + n_2 - 2)$ degrees of freedom ($\hat{t} \sim t_{n_1 + n_2 - 2}$). It is well known that violating the assumption of variance homogeneity would result in type I error inflation or power loss of the ST [20].

The WT, as an adaptation of the ST, is more reliable when the two-group samples have unequal variances and unequal sample sizes. The Welch statistic is defined by

$$\widehat{WT} = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\frac{\hat{\sigma}_{i1}^2}{n_1} + \frac{\hat{\sigma}_{i2}^2}{n_2}}}.$$

This statistic calibrates the impact of potential variance heterogeneity between two conditions. For a gene with equal means between two conditions (regardless of variance heterogeneity), $\widehat{WT}$ approximately follows a *t*-distribution with the Welch–Satterthwaite degree of freedom:

$$\nu = \frac{\left(\frac{\hat{\sigma}_{i1}^2}{n_1} + \frac{\hat{\sigma}_{i2}^2}{n_2}\right)^2}{\left(\frac{\hat{\sigma}_{i1}^4}{n_1^2(n_1 - 1)} + \frac{\hat{\sigma}_{i2}^4}{n_2^2(n_2 - 1)}\right)}.$$

To calibrate unequal variances, another alternative is the MWT [97], which would yield reliable condition-specific variance estimators for low-replicate experiments. For large-sample experiments, one can perform Student t test on standardized data (STSD), where the gene expression levels are divided by condition-specific sample standard deviations respectively.

3.2.2.2 **Concepts of MVDE genes and variance heterogeneity tests**

A gene is called to be susceptible if the change in condition can alter arbitrary aspects of the entire distribution of its expression level, i.e., mean, variance, kurtosis and/or even higher-order characteristics. The term MVDE gene is adopted to describe a gene whose mean and/or variance in expression level is sensitive to the change in condition. Formally, a MVDE gene has different means ($\mu_1 \neq \mu_2$) and/or variances ($\sigma_1^2 \neq \sigma_2^2$) of expression levels between two conditions. This concept of MVDE genes goes one step closer to our general concept of a susceptible gene and is more reasonable than the

conventional concept of MDE genes, which confines to differential mean expression levels only. In gene co-expression networks, genes work together and the expression levels are correlated. Some susceptible genes may also interact with other susceptible genes and/or environmental factors. Such correlations and interactions among biological networks are very common and are major drivers for the variance heterogeneity of a test susceptible gene. Variance heterogeneity, to some extent, indicates how a gene involve in complex networks. Therefore, we argue that variance heterogeneity should be as equally important as mean heterogeneity for identifying differentially expressed genes. To identify susceptible genes, one crucial step is to extract summary statistics containing potential information about variance heterogeneity, i.e., the *p* values computed from some appropriate test statistic on the null hypothesis $H_{02}^{(i)}$ (variance homogeneity).

For a random gene, if its (transformed) expression levels follow normal distribution, then the classical two-sample *F*-statistic

$$\hat{F} = \frac{\hat{\sigma}_{i1}^2}{\hat{\sigma}_{i2}^2}$$

follows the centralized *F*-distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom ($\hat{F} \sim F_{n_1-1,n_2-1}$) since $H_{02}^{(i)}$ is true. Under normality setting, the *F*-test is the most powerful test for exploiting variance heterogeneity. Nevertheless, the *F*-test is very sensitive to the violation of normality. Therefore, it may claim random genes to be spuriously significant if their (transformed) expression levels do not strictly follow normal distributions. Actually, the two-sample *F* test is more suitable for testing normality other than variance heterogeneity [100].

As a robust alternative, the Brown–Forsythe statistic is the *F*-ratio that stems from applying the ordinary one-way analysis of variance on the absolute deviations from the median:

$$\widehat{BF} = \frac{(n_1 + n_2 - 2)\sum_{c=1}^{2} n_c(\bar{Z}_{ic} - \bar{Z}_i)^2}{\sum_{c=1}^{2}\sum_{j=1}^{n_c}\left(Z_{ijc} - \bar{Z}_{ic}\right)^2},$$

where $Z_{ijc} = \left|G_{ijc} - \tilde{G}_{ic}\right|$, $\bar{Z}_{ic} = \frac{1}{n_c}\sum_{j=1}^{n_c} Z_{ijc}$, $\bar{Z}_i = \frac{1}{n_1+n_2}\sum_{c=1}^{2}\sum_{j=1}^{n_c} Z_{ijc}$, and $\tilde{G}_{ic} = median(\boldsymbol{G_c})$. When $H_{02}^{(i)}$ is true, the distribution of $\widehat{BF}$ follows approximately the *F*-distribution with degrees of freedom 1 and $(n_1 + n_2 - 2)$.

Another alternative, the Levene test, uses the mean instead of the median:

$$\widehat{LF} = \frac{(n_1 + n_2 - 2)\sum_{c=1}^{2} n_c(\bar{Z}_{ic} - \bar{Z}_i)^2}{\sum_{c=1}^{2}\sum_{j=1}^{n_c}\left(Z_{ijc} - \bar{Z}_{ic}\right)^2},$$

where $Z_{ijc} = \left|G_{ijc} - \bar{G}_{ic}\right|, \bar{Z}_{ic} = \frac{1}{n_c}\sum_{j=1}^{n_c} Z_{ijc}, \bar{Z}_i = \frac{1}{n_1+n_2}\sum_{c=1}^{2}\sum_{j=1}^{n_c} Z_{ijc}$ and $\bar{G}_{ic} = mean(\boldsymbol{G_c})$. If $H_{02}^{(i)}$ is true, then $\widehat{LF}$ follows approximately the *F* distribution with degrees of freedom 1 and $(n_1 + n_2 - 2)$.

For each gene, the optimal test for variance heterogeneity depends on the underlying gene expression distribution. According to Brown and Forsythe's Monte Carlo studies [102], the Levene test provided the best power for symmetric, moderate-tailed distributions; whereas the Brown–Forsythe test performed best when the underlying data followed heavily skewed distributions.

3.2.2.3 **Integrating mean and variance heterogeneities**

One most commonly used method to integrate two independent pieces of information is Fisher's linear combination. For a testing gene, let $p_{WT}, p_F, p_{BF}, p_{LF}$ denote

the *p*-values of the Welch statistic, the *F* statistic, the Brown-Forsythe statistic and the Levene statistic, respectively. We recommend using $\widehat{IMVT} = -2(\log(p_{WT}) + \log(p_{LF}))$ to integrate mean and variance heterogeneities. Another two alternatives are $\widehat{FWT} = -2(\log(p_{WT}) + \log(p_F))$ and $\widehat{BFWT} = -2(\log(p_{WT}) + \log(p_{BF}))$. Each of the three Fisher linear combinations follows approximately the χ^2- distribution with 4 degrees of freedom, provided that the *p*-values of mean heterogeneity tests are independent of the *p*-values of variance heterogeneity tests under joint null H_{03}.

3.2.2.4 **Alternative tests for the joint null hypothesis of mean and variance equalities**

To test H_{03}, a framework of separate mean and variance tests (SMVT) can also be conducted. This framework applies WT on H_{01} (mean equality) at nominal level α_1 and Levene test on H_{02} (variance equality) at nominal level α_2, respectively. H_{03} is rejected if H_{01} or H_{02} or both are rejected. By our proposition on the null independence, type I error rate of this framework is given by $\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$. It is intractable to choose universal optimal α_1 and α_2 for all genes. To control the overall type I error rate at nominal level α, one typical choice is setting $\alpha_1 = \alpha_2 = 1 - \sqrt{\alpha}$. Similar as Fisher's linear combination, the SMVT gives equal weight to mean heterogeneity and variance heterogeneity.

The two-sample *LRT* is another alternative to test H_{03}, assuming the (transformed) expression levels follow normal distributions. Specifically, the LRT statistic is given by

$$\widehat{LRT} = \frac{\left(\frac{n_1-1}{n_1}\hat{\sigma}_{i1}^2\right)^{\frac{n_1}{2}}\left(\frac{n_2-1}{n_2}\hat{\sigma}_{i2}^2\right)^{\frac{n_2}{2}}}{\left(\frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1}(G_{ij1}-\hat{\mu})^2+\sum_{j=1}^{n_2}(G_{ij2}-\hat{\mu})^2\right)\right)^{\frac{n_1+n_2}{2}}},$$

$\hat{\mu} = \frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1} G_{ij1} + \sum_{j=1}^{n_2} G_{ij2}\right)$ (See **APPENDIX** B for mathematical derivation of the LRT statistic). Under normal setting with H_{03}, $\hat{\chi}_2^2 = -2\ln(\widehat{LRT})$ follows χ^2- distribution with 2 degrees of freedom asymptotically for large sample sizes.

3.2.3 Results

3.2.3.1 The null independence between the mean and variance heterogeneity tests

It's commonly believed that testing statistics of mean and variance heterogeneities are dependently distributed, even if the data forming them are from an identical normal population. For example, both Student's *t*-statistic and the *F*-statistic are defined in terms of sample variances. In fact, all aforesaid testing statistics of mean heterogeneity are independent of all aforesaid testing statistics of variance heterogeneity under H_{03}. This null independence lays the foundation of type I error rate control of the integrative heterogeneity tests. Herein, we formally formulate the finite-sample null independence by the following proposition:

Proposition: *Student t statistic and Welch t statistic are independent of the F-, Levene and Brown-Forsythe statistics if the finite samples (**G**$_1$, **G**$_2$) forming them jointly follow an arbitrary spherically symmetric distribution.*

The proposition formulates the finite-sample null independence under a broader distribution family, including normality as a special member (see the APPENDIX B for mathematical proofs). Its typical members include multivariate Gaussian, Student, Kotz, exponential power, Laplace distributions with spherically symmetric variance-covariance matrices [100]. Many researchers are familiar with and usually adopt normality assumption

on (transformed) gene expression levels. By this proposition, if the normality assumption is met, the proposed integrative heterogeneity tests can properly control the type I error rate. However, the normality assumption is often violated more or less by real-world gene expression data. Rigorously speaking, no transformation of gene expression data can assure exact normality. Therefore, it is necessary and useful to extend the null independence to broader distribution families, e.g., spherically symmetric family.

To empirically illustrate the proposition, we generated 100000 replicates of two-group samples from the standard normal distribution with sample size $n_1 = n_2 = 40$. As anticipated by the proposition, the majority of replicate-specific pairs of Welch t statistic ($\widehat{WT}$) and Levene statistic ($\widehat{LF}$) randomly concentrates around (0, 1) (**Figure 3-1 (a)**) and so do the replicate-specific Welch t statistic and *F* statistic pairs (**Figure 3-1 (b)**). Under this simulation design, Welch t and Student t statistics ($\widehat{WT}$, $\hat{t}$) appeared equivalent (**Figure 3-1 (c)**). The correlation between Levene statistic ($\widehat{LF}$) and Brown-Forsythe statistic ($\widehat{BF}$) turned to be 0.9894 (**Figure 3-1 (d)**). The scatterplots of ($\hat{t}, \widehat{LF}$), ($\hat{t}, \widehat{BF}$), and ($\hat{t}, \hat{F}$) are qualitatively the same as those of ($\widehat{WT}, \widehat{LF}$) (Results not shown here). Under the normality setting with smaller sample sizes, we also obtained the corresponding figures for some other sample sizes (**Figure B-1**-**Figure B-7**), which revealed very similar patterns to **Figure 3-1**. Standard multi-variate normal distribution is a typical member in the family of spherically symmetric distributions. These simulation results illustrate the null independence within the family of all spherically symmetric distributions.

As explorations outside of the spherically symmetric family, we performed comprehensive simulations by generating the data from the standard Laplace distribution. Univariate Laplace distribution is a typical member of the family of symmetric

distributions. However, the joint distribution of independent univariate Laplace variables is outside of the spherically symmetric distribution family. Under the standard Laplace setting, we obtained the corresponding scatterplots and observed similar patterns of the joint distributions of the mean and variance test statistics (**Figure S2.1-Figure S2.4**). These empirical results illustrate the robustness of the null independence between mean and variance tests for the data from the family of symmetric distributions.

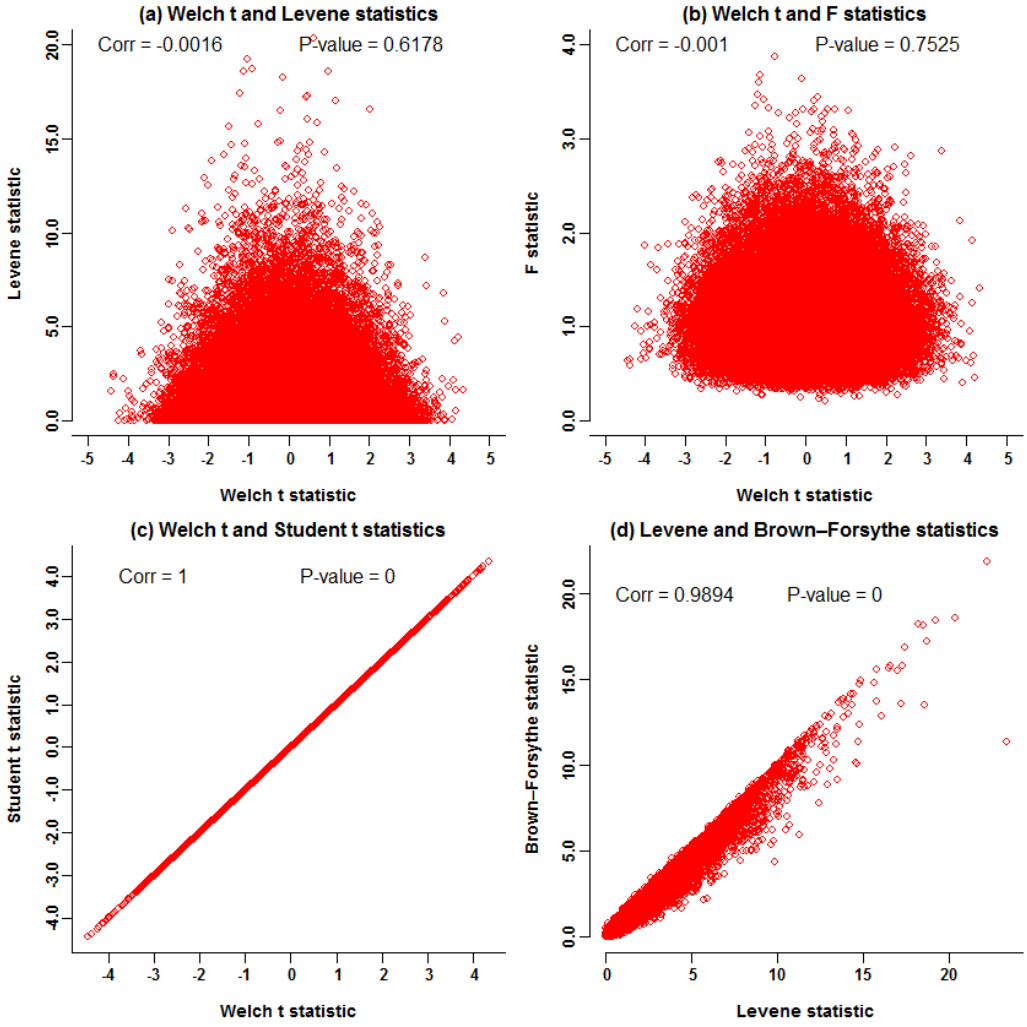

Figure 3-1: Null joint distributions of the test statistics on mean and variance heterogeneities under normality setting.

3.2.3.2 **Type I error rates control of the competitors**

Under normality setting. With extremely small samples, none of the eight competitors could properly control type I error rates (**Figure 3-2 (a)**). The LRT and the STSD severely inflated type I error rates. The IMVT and the SMVT appeared equally anti-conservative; both were much less anti-conservative than the LRT and the STSD. The MWT performed the best to control type I error rates; it was slightly conservative. The WT and the FWT appeared equally conservative; both were clearly more conservative than the MWT. The BFWT appeared severely conservative. The LRT inflated the type I error rates because the χ_2^2 distribution could not well approximate the exact distribution of the LRT statistic. The anti-conservative of the STSD stemmed from the variability of condition-specific data standardization. Specifically, sample standard deviations of small samples could not precisely estimate the standard deviation. The conservativeness of the BFWT stemmed from the well-known conservativeness of the Brown-Forsythe test [104, 105]. For larger sample sizes (**Figure 3-2(b-d)**), the LRT, the STSD, the SMVT and the IMVT appeared less anti-conservative, and the MWT, the WT, the FWT and the BFWT became less conservative. When sample sizes reached 40, the IMVT and the SMVT as well as the WT, the MWT and the FWT properly controlled the Type I error rates (**Figure 3-2(d)**).

Under the Laplace setting, the LRT and the FWT appeared severely anti-conservative (**Figure 3-3 (a-d)**). Their inflations in type I error rate appeared even severer as the samples increased. The LRT had inflated type I error rates because it was derived from normality assumption of gene expression levels. The FWT had inflated type I error rates because the *F* test statistic is very sensitive to the non-normality of the samples [100].

The other tests displayed similar patterns to those under normality setting. For extremely small sample sizes, the STSD, the IMVT and the SMVT appeared successively less anti-conservative; whereas the MWT, the WT and the BFWT appeared successively more conservative (**Figure 3-3 (a)**). Their magnitudes of inflations and deflations in type I error rate appeared to vanish as the sample sizes increased (**Figure 3-3 (b-d)**).

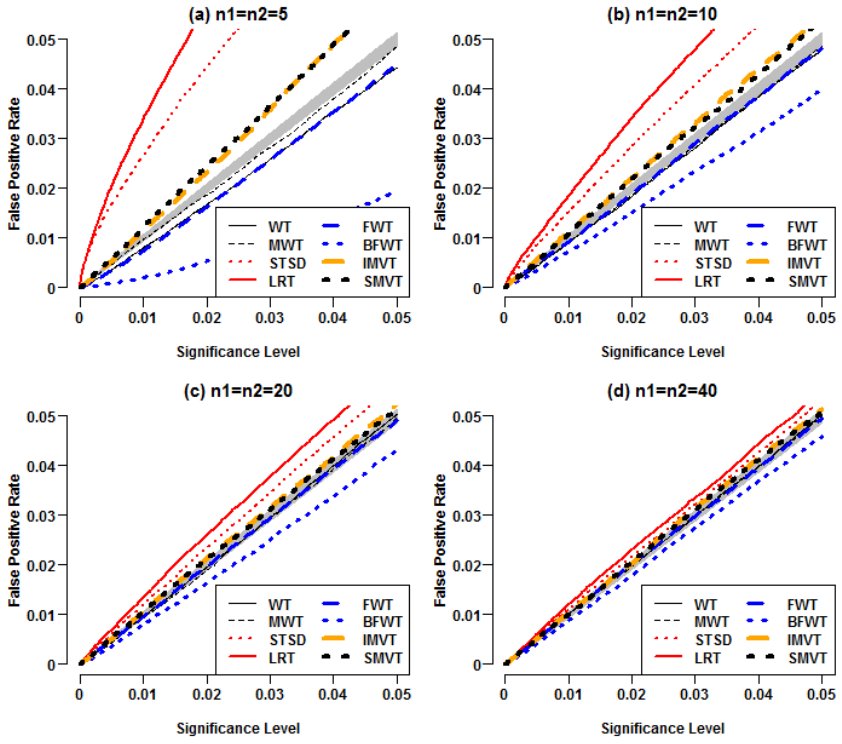

Figure 3-2: Comparison of false positive rates of eight methods under standard normality setting.

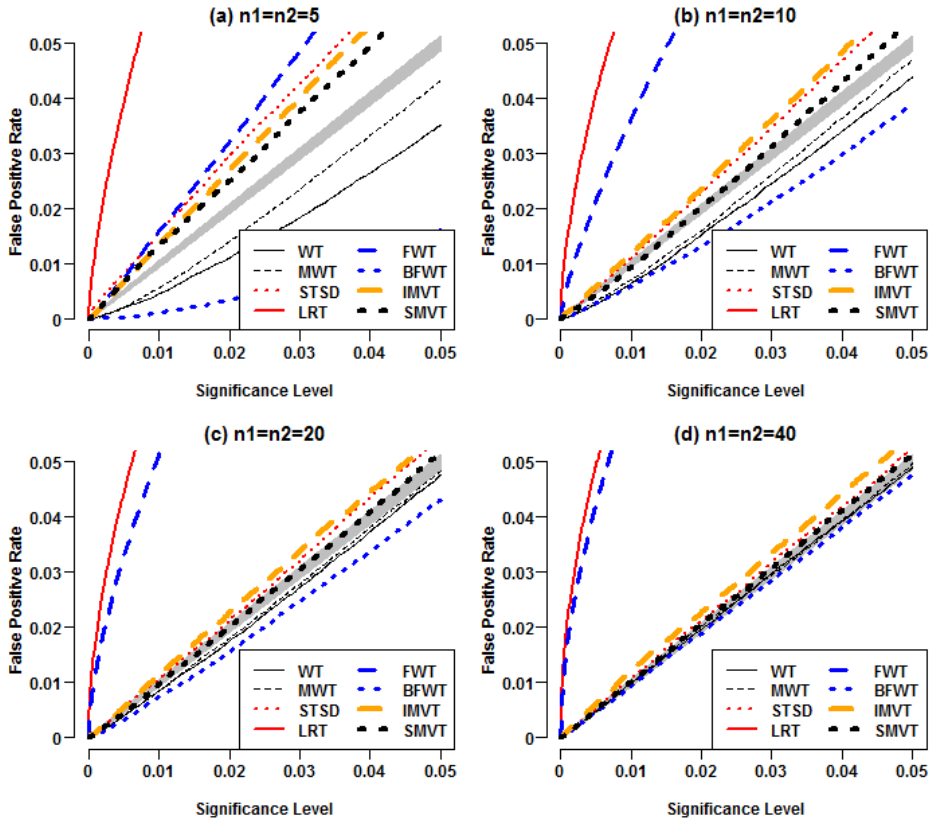

Figure 3-3: Comparison of false positive rates of eight methods under standard Laplace setting.

3.2.3.3 **Empirical power comparisons under normality setting and non-normality setting**

For power comparisons, we investigated three kinds of scenarios under both normality setting and Laplace setting: (1) unequal mean and equal variance, (2) equal mean and unequal variance and, (3) unequal mean and unequal variance. For sample sizes as large as $n_1 = n_2 = 40$, the proposed and existent tests well controlled type I error rates under normality and Laplace setting. And the sample size is very close to those of the gene

expression files of Pan et al. [103]. We thus presented here the power comparisons with the sample sizes $n_1 = n_2 = 40$.

Under normality setting, Herein, the parameters r and s represent the magnitudes of mean and variance heterogeneities, respectively. When $s \neq 0$, the IMVT and the FWT displayed the highest powers, followed by the SMVT; and all the three joint heterogeneity tests outperformed the three mean heterogeneity tests, i.e., the WT, the MWT and the STSD (**Figure 3-4 (a-b)**). The power gains of the joint heterogeneity tests over the mean heterogeneity tests appeared especially noteworthy when $s \neq 0$ and $r = 0$ (**Figure 3-4(b)**). The joint heterogeneity tests did not display severe power losses even for the theoretical scenarios favoring the mean tests (**Figure 3-4(c)**). In addition, the FWT slightly outperformed the IMVT because the F test statistic is the optimal test statistic for variance heterogeneity under normality setting. Here, we did not compare the powers of the LRT and the BFWT since they could not control type I error rates.

Under Laplace setting, we simulated independently 10000 replicates of $n_1 = 40$ data points from standard Laplace distribution $Laplace\ (0,1)$ and $n_2 = 40$ data points from $Laplace\ (r, (1+s)^2)$ for each (r, s) pair. Again, the parameters r and s represent the magnitudes of mean and variance heterogeneities, respectively. Under the Laplace setting, we observed qualitatively the same patterns as those under the normality setting. When $s \neq 0$, the IMVT outperformed the SMVT; and both the joint heterogeneity tests outperformed the three mean heterogeneity tests, i.e., the WT, the MWT and the STSD (**Figure 3-5(a-b)**). The power gains of the joint heterogeneity tests over the mean heterogeneity tests appeared especially noteworthy when $s \neq 0$ and $r = 0$ (**Figure 3-5(b)**). The joint heterogeneity tests did not display severe power losses even for the

theoretical scenarios favoring the mean heterogeneity tests (**Figure 3-5(c)**). Here, we did not compare the powers of the LRT, the FWT and the BFWT since they could not control type I error rates under non-normality setting.

These results formally demonstrate the importance of integrating informative variance heterogeneity. In general, the power gains of the IMVT over its competitors are solid. For the scenarios of mean heterogeneity only, the IMVT would have small power losses. All in all, the IMVT displayed valuable merits over its competitors. At least, the IMVT is an admissible procedure. It should be useful to improve the power to identify susceptible genes involved in co-expression networks. By its robustness to non-normality data, we recommend the IMVT as a powerful alternative to exploit microarray profiles.

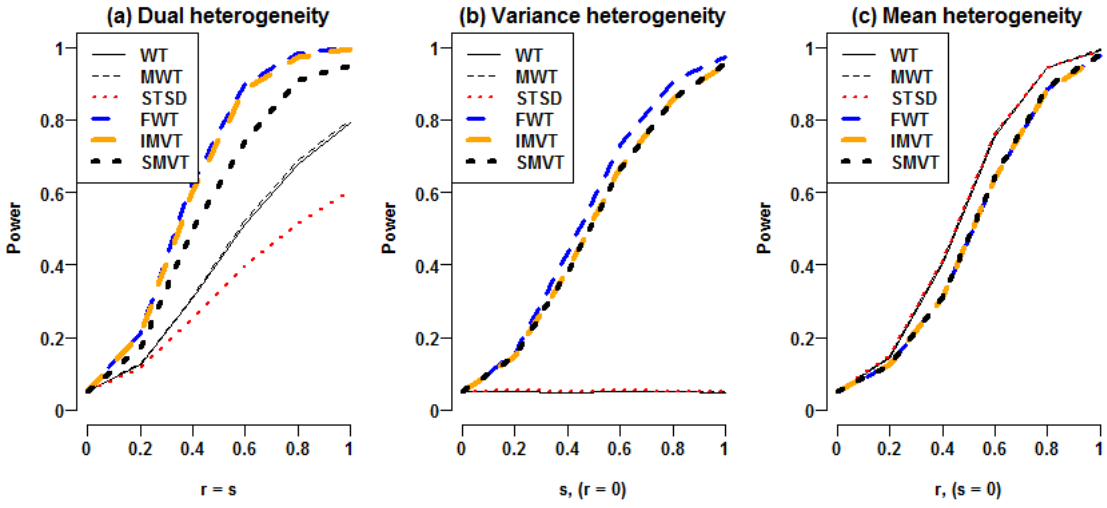

Figure 3-4: Power comparison of six methods under two-condition normality setting.

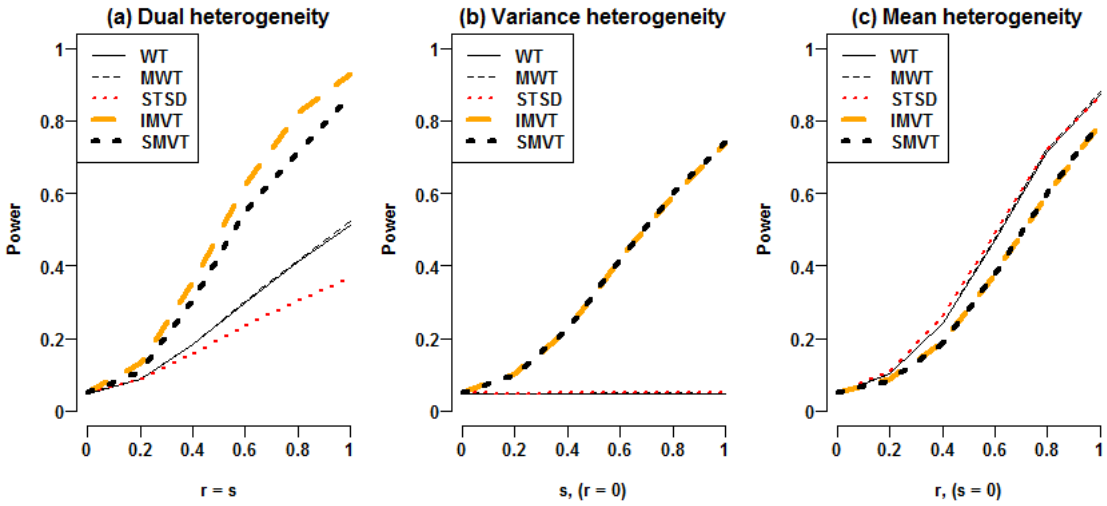

Figure 3-5: Power comparison of six methods under two-condition Laplace setting.

3.2.3.4 Re-analyzing the gene expression profiles of peripheral circulating B Lymphocytes

Pan *et al.* [103] compared the gene expressions profiles of peripheral circulating B cells between 39 smoking and 40 non-smoking healthy US white women. Using MAS5 software, they normalized the expression levels of 7215 selected probes out of all the 22,283 experiment-wide probes. They applied traditional *t* tests to the normalized expression levels and report 125 promising DE genes. The authors justified why they did not adjust for menopausal status and age. However, they neglected the latent background data structure. Using the MAS5 software, we normalized the raw expression levels of all the 22283 experiment-wide gene probes. For the normalized data, we computed the probe specific test statistics and *p* values of five competitors. The genomic inflation factors [106] of these heterogeneity tests would be close to 1 if they could properly control type I error rates. However, all the tests displayed huge genomic inflation factors, especially the STSD

(**Figure 3-6**). All the *Q*-*Q* plots climbed quickly above the upper limit of the 95% concentration band (the gray band). The severe genomic inflations indicated that some major latent factors would confound all the competitors. Thus, the *t* tests performed by Pan *et al*. [103] would be confounded since they did not adjust for any background factors.

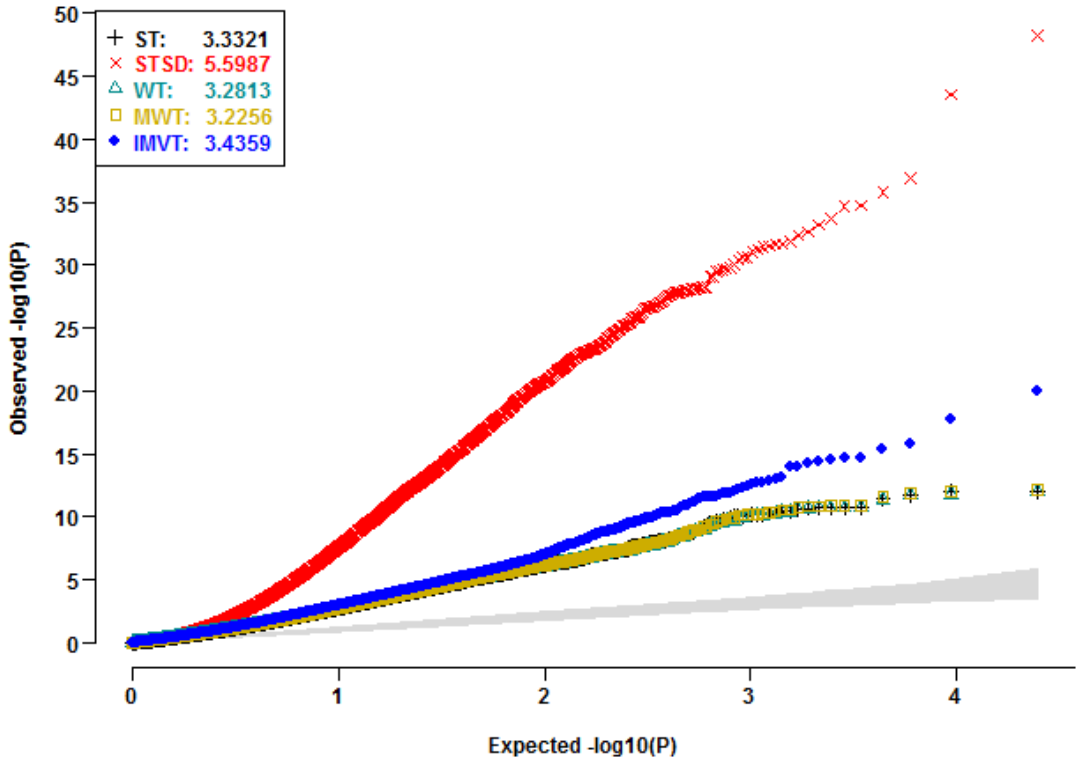

Figure 3-6: Q-Q plots of the five competitors without adjusting for latent data structure and covariates.

To reveal latent data structure, we first conducted PCA of the MAS5 normalized expression levels of all the 22283 experiment-wide gene probes (**Figure 3-7, Table B1**). PC1 was the unique major PC, accounting for 98.24% of the total variation (**Figure 3-7 (a)**). PC2 merely accounted for 0.32% of total variation. Neither PC1 nor PC2 displayed

mean heterogeneity or variance heterogeneity between the smokers and nonsmokers (**Figure 3-7 (b)**). PC4 displayed strikingly significant mean heterogeneity ($p_{WT} = 1.91 \times 10^{-15}$), even if it only accounted for 0.13% of the total variation. PC6 displayed very significant variance heterogeneity ($p_{LF} = 3.2 \times 10^{-4}$) even if it accounted for 0.07% of the total variation only. PC4 and PC6 distinguished the smokers and the nonsmokers (**Figure 3-7 (c)**). **Table B-1** listed the first 2 and all the global PCs with significant mean and/or variance heterogeneities. These significant global PCs did not distinguish informative heterogeneities and impediment heterogeneities. They were so significant in that they would account for portions of informative mean and variance heterogeneities of DE genes in addition to background heterogeneities. As shown in **Figure 3-8**, naively adjusting for the significant global PCs of all gene probes would result in severe power loss (genomic deflation).

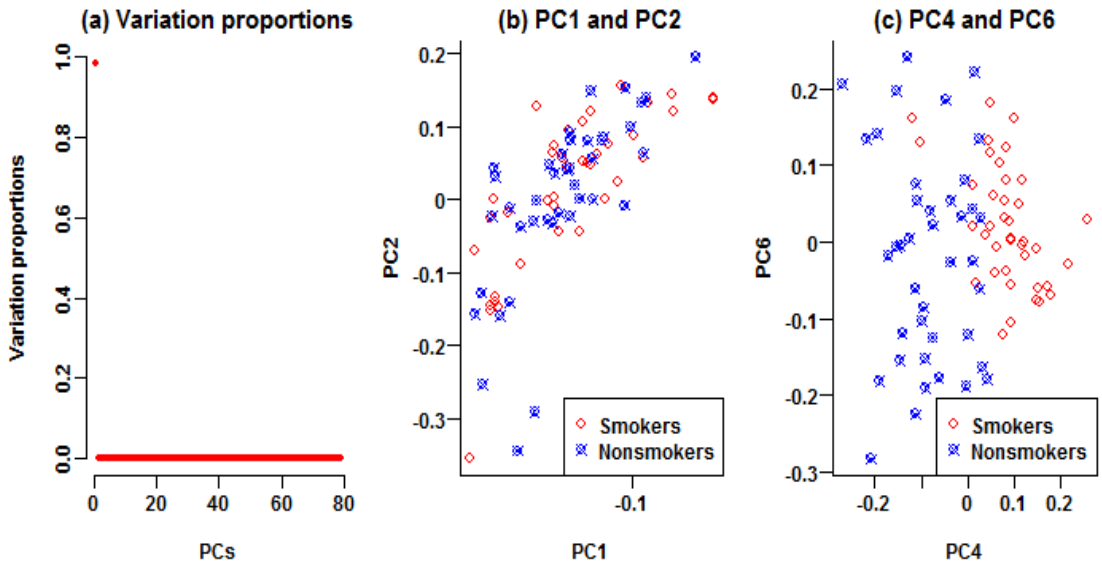

Figure 3-7: Global data structure of all the experiment-wide gene expression levels.

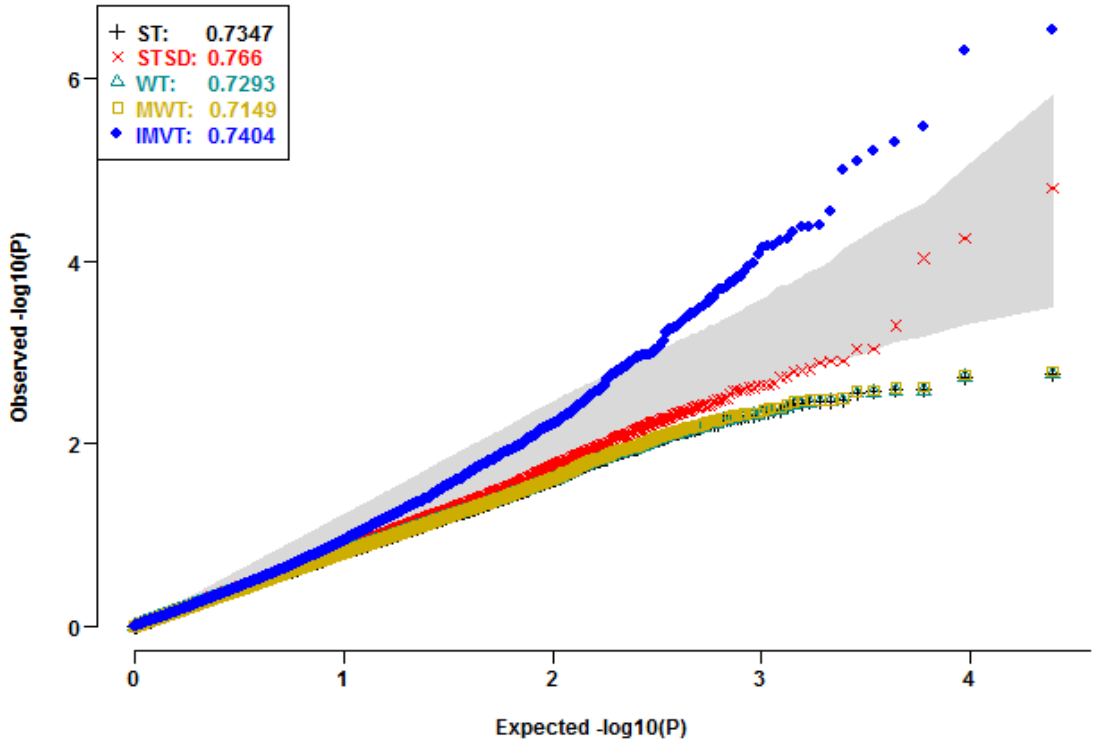

Figure 3-8: Deflations due to the over adjustment of the experiment-wide data structure.

To prevent false positives and false negatives, we selected 13415 'robust' gene probes to capture the background data structure. The spirit here is similar to the use of control genes to account for unwanted variation [107]. None of the robust gene probes displayed mean heterogeneity or variance heterogeneity, before and after calibrating the significant background PCs, age and menopausal status. We conducted PCA of the MAS5 normalized data of the 'robust' gene probes (**Figure 3-9, Table B-2**). PC1 alone accounted for 98.35% of the total variation and was the unique major PC. PC2 merely accounted for 0.37% of total variation (**Figure 3-9 (a)**). Neither PC1 nor PC2 displayed mean

heterogeneity or variance heterogeneity (**Figure 3-9 (b)**). PC14 displayed the most significant mean heterogeneity ($p_{WT} = 0.0036$), even if it only accounted for 0.03% of the total variation. PC28 displayed the most significant variance heterogeneity ($p_{LF} = 0.0069$) even if it only accounted for 0.01% of the total variation. PC14 and PC28 displayed clear stratification of the smokers and the nonsmokers (**Figure 3-9 (c)**). In addition, **Table B-2** listed the first 2 and all the background PCs with significant mean and/or variance heterogeneities. After adjusting for these significant background PCs, age and menopausal status, the *Q*-*Q* plots of all the five tests climbed above the diagonal (**Figure 3-10**). Especially, the *Q*-*Q* plot of the IMVT climbed above the upper limit of the 95% concentration band. All the five tests displayed reasonable inflation factors. The mild inflation might be due to weak differentials or residual correlations between DE genes.

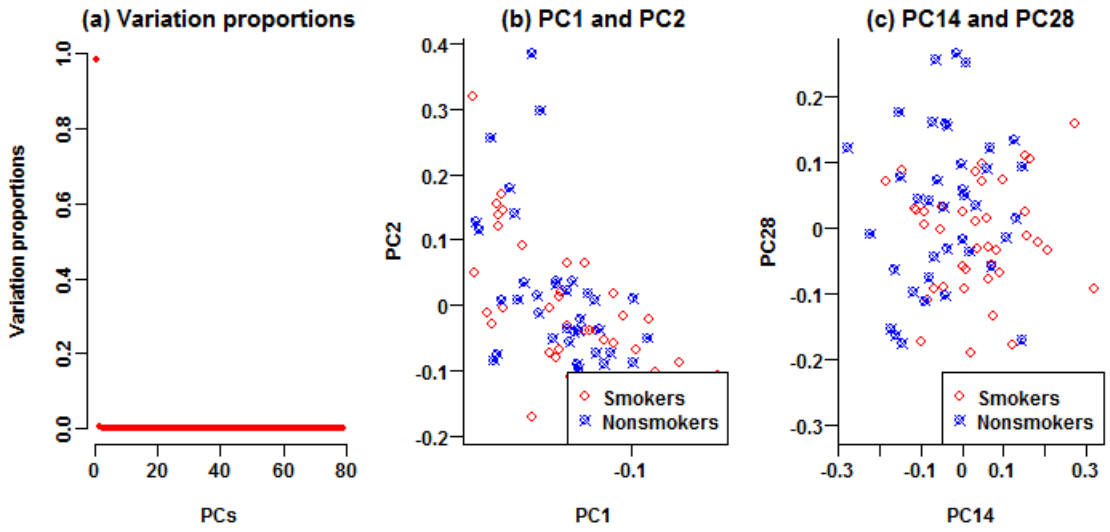

Figure 3-9: Background data structure of the expression levels of robust gene probes.

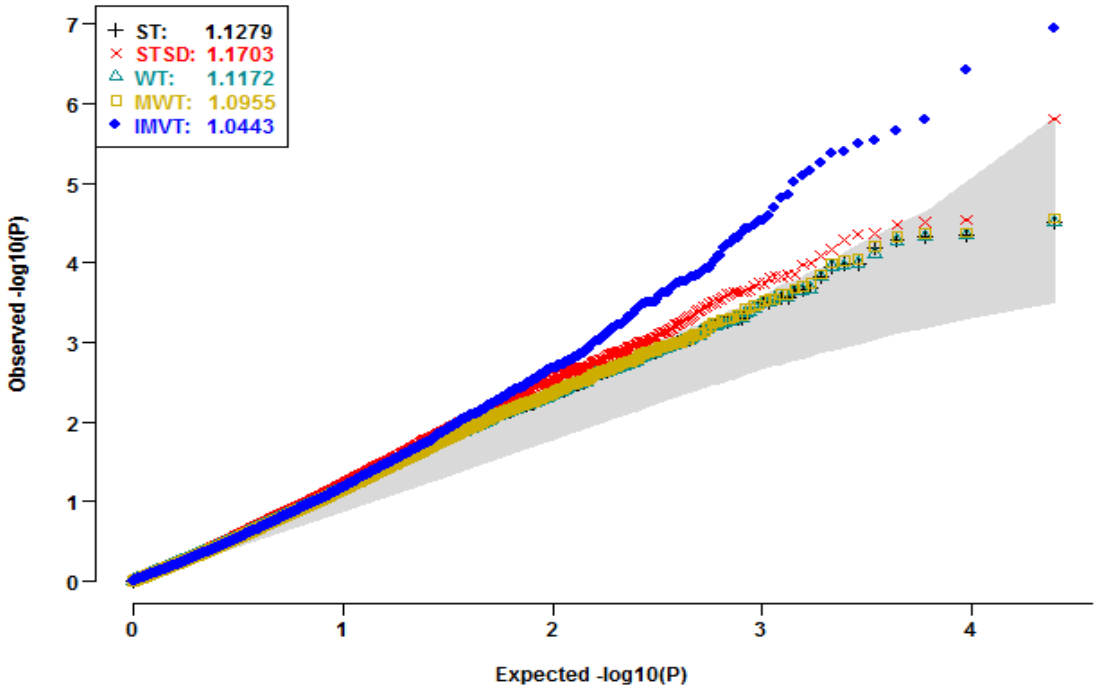

Figure 3-10: Q-Q plots of the five competitors after adjusting for background data structure and covariates.

Applied to the calibrated expressions, our IMVT identified *CUL7*, *RBMY1J*, *RDH5* and *SOCS3* to be experiment-wide significant (**Table 3-1**), i.e., $p_{IMVT} < 0.05/22283 = 2.24 \times 10^{-6}$. The STSD only identified *CUL7* as experiment-wide significant gene; while the WT and the MWT failed to identify any experiment-wide significant genes. The experiment-wide minimum p value of the WT and the MWT turned to be 2.73×10^{-5}, much larger than 2.24×10^{-6}. The SMVT failed to identify any gene to be experiment-wide significant. At *DDX3X*, the WT reached the experiment-wide minimum $p_{WT} = 3.10 \times 10^{-5}$. For SMVT, both p_{WT} and p_{LF} must be smaller than threshold $1 - \sqrt{1 - 0.05/22283} = 1.12 \times 10^{-6}$ to control overall experiment-wide type I error rate at 0.05. Therefore, our analysis of the real data provided solid evidence for the superiority of

the IMVT over the SMVT. Without adjusting for the data structure and covariates, Pan et *al.* [103] did not report any of the four genes although their results were severely inflated. *SOCS3* was reported to be related to tobacco smoking by independent studies [108-111]. Per the database of cancer gene networks (TCNG; http://tcng.hgc.jp/index.html), *CUL7* [112-114], *RBMY1J* [112, 114] and *RDH5* [112-117] were reported to involve in function gene networks related to smoking. All the four experiment-wide significant gene probes displayed both mean and variance heterogeneities (**Figure 3-11**). In addition to the four experiment-wide significant genes, our IMVT identified 16 genes that testified to be involved in functional networks by Pan et *al.* [103] at nominal level 0.05 (**Table 3-2**). For a test gene within a network of functional genes, incorporating its informative variance heterogeneity proved one effective way to exploit extra information as provided by the other function genes in the same network.

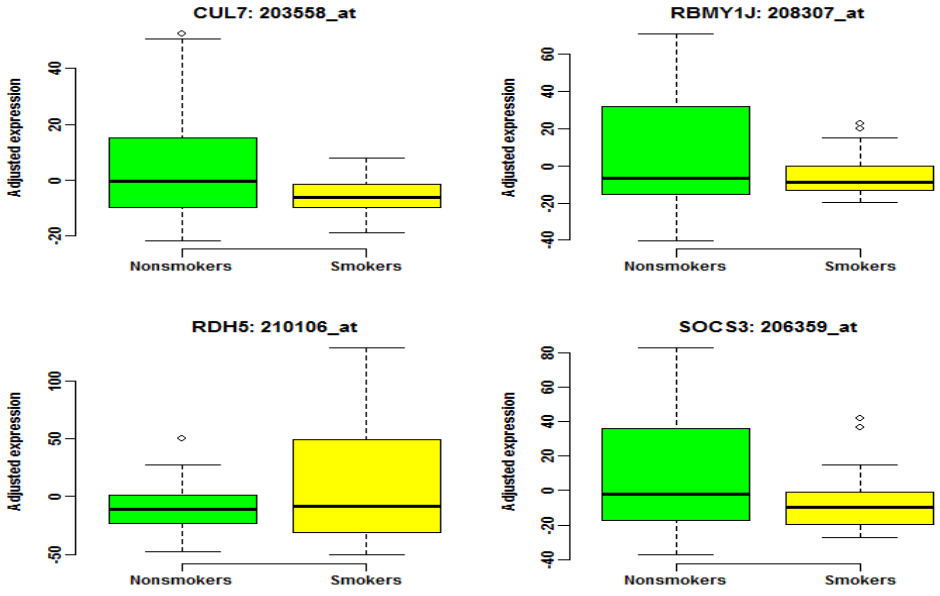

Figure 3-11: Boxplots of four experiment-wide significant gene probes.

Table 3-1: Experiment-wide significant discoveries by the IMVT*

AffyID	Gene	IMVT	STSD	MWT	WT
203558_at	*CUL7*	1.12E-07	1.55E-06	0.0034	0.0024
208307_at	*RBMY1J*	3.82E-07	0.0051	0.0422	0.0398
210106_at	*RDH5*	1.56E-06	0.0059	0.0295	0.0302
206359_at	*SOCS3*	2.22E-06	0.0014	0.0081	0.0078

* All the probe-specific p_{IMVT} values reported here are smaller than $0.05/22283 = 2.2438\times10^{-6}$. The STSD identified CUL7 with much weaker evidence while the WT and MWT did not identify any gene probe to be experiment-wide significant.

Table 3-2: The overlap of the discoveries of our IMVT and the genes which were testified to be involved in functional networks

AffyID	Gene	Adjusted MAS5*				MAS5**
		IMVT	STSD	MWT	WT	ST
201085_s_at	*SON*	0.0075	0.0021	0.0021	0.0023	2.15E-14
203868_s_at	*VCAM1*	0.0030	0.0004	0.0005	0.0005	2.03E-07
204524_at	*PDPK1*	0.0470	0.0328	0.0337	0.0346	7.12E-11
204600_at	*EPHB3*	0.0178	0.0165	0.0207	0.0213	2.83E-04
205008_s_at	*CIB2*	0.0387	0.0122	0.0117	0.0123	1.25E-06
205099_s_at	*CCR1*	0.0058	0.0104	0.0160	0.0165	6.55E-11
206788_s_at	*CBFB*	0.0003	4.34E-05	4.28E-05	4.71E-05	<1.00E-17
207961_x_at	*MYH11*	0.0001	0.0139	0.0370	0.0383	8.11E-06
208164_s_at	*IL9R*	0.0311	0.0074	0.0072	0.0077	4.05E-05
209876_at	*GIT2*	0.0024	0.0040	0.0053	0.0057	1.20E-08
211197_s_at	*ICOSLG*	0.0448	0.0423	0.0479	0.0487	3.28E-05
211699_x_at	*HBA1*	0.0455	0.3238	0.3632	0.3667	2.70E-03
212514_x_at	*DDX3X*	0.0002	3.06E-05	2.73E-05	3.10E-05	2.22E-16
213446_s_at	*IQGAP1*	0.0082	0.0306	0.0400	0.0413	8.37E-10
217557_s_at	*CPM*	0.0347	0.2422	0.2678	0.2701	1.61E-03
219599_at	*EIF4B*	0.0006	0.0005	0.0018	0.0019	5.80E-14

***These raw *p* values of the heterogeneity tests based on the calibrated expression levels after adjusting for age, menopausal status, and the background structure.**

****These raw *p* values of Student *t* tests in Pan et al. [103] based on the MAS5 normalized data before adjusting for any of age, menopausal status, and the background structure.**

The false discovery rate (FDR) would be a more appropriate error rate to control than the familywise error rate in microarray studies; and several standard FDR controlling procedures have been widely practiced [118-121]. We did identify more promising gene probes when applying the most widely used FDR controlling procedure to the *p* values generated by our IMVT. For example, controlling FDR at the stringent level 0.05, our IMVT identified 24 out of the experiment-wide 22283 gene probes. Controlling FDR at the same level, the STSD only identified *CUL7*, while both the WT and the MWT missed all promising gene probes (**Table B-3**). Controlling FDR at level 0.1, our IMVT claimed 55 gene probes, while all the three mean heterogeneity tests discovered no additional gene probes. These results have well demonstrated noteworthy gains of explicitly exploiting informative variance heterogeneity. Without adjusting for background data structure, Pant *et al.* claimed 125 gene probes with local FDRs < 0.05. Their published list of promising gene probes displays huge discrepancies to ours. Such discrepancies stemmed from the severe inflation in their *t* tests (**Figure 3-6**). Judiciously calibrating background data structure is thus necessary for accurately prioritizing gene probes.

3.3 Part II: Novel Double Welch *t* test to Identify Functional Differentially Expressed Genes

3.3.1 Introduction

Part I presented the limitations of only exploiting mean heterogeneity and proposed an integrative IMVT method to combine the mean and variance heterogeneities from Welch *t* test and Levene test, respectively. The proper null hypothesis of testing MVDE gene is $H_{03} = H_{01} \cap H_{02}$: the gene has equal mean and equal variance of expression levels

between the two conditions. As demonstrated in discussion, the main disadvantage of integrating mean and variance heterogeneities is the independence of mean test and variance tests, either under null or alternative hypothesis. Therefore to overcome the disadvantage of integrating mean and variance heterogeneities, we put forth a more powerful novel method DWT to integrate mean and high-order heterogeneities in Part II. It goes one step closer to detecting MVDE gene - a gene displays reliable changes in any aspects of the entire distribution of its expression level with the change in condition than any other existent methods. If a gene is not a function gene related to corresponding disease, it would display mean equality (H_{01}) and second-order moments equality (H_{04}) of expression levels between two different conditions. The proper null hypothesis of testing functional MVDE gene is $H_{05} = H_{01} \cap H_{04}$: the gene has equal mean and equal second order moment of expression levels between the two conditions. This null hypothesis is equivalent to equal mean and equal variance hypothesis since the second order moment is the summation of the square of mean and variance. If H_{05} is rejected, then we claim the testing gene as candidate MVDE gene. To capture the high-order heterogeneity, we constructed a welch t test statistic for testing H_{04}. Under H_{05}, the testing statistics of detecting mean heterogeneity and high-order heterogeneity are asymptotically independently distributed. This null independence is crucial for controlling the type I error rate control of DWT. While under alternative hypothesis, the two test statistics are dependent. Therefore the DWT appeared more powerful than our earlier IMVT method integrating mean and variance signals to identify genes with or without variance heterogeneity. We also reanalyzed the gene profiles of peripheral circulating B cells [16] after adjusting for global confounders and background data structure. Our DWT replicated

more reported genes that involve in networks and had better performance than IMVT method. Our results highlighted the importance of exploiting informative high-order heterogeneity, which is a rich resource about the biology mechanism of gene expressions beyond the mean heterogeneity.

3.3.2 Methods and Materials

3.3.2.1 The double welch t test (DWT) to integrate mean and second-order heterogeneities

Let the dataset contain expression levels of M gene probes of n_1 unrelated subjects from control groups and n_2 unrelated subjects for treatment group, respectively. For a specific gene, let $X_1 = \left(X_{11}, X_{12}, \dots, X_{1n_1}\right)$ be the expression level of gene probes under control group and $X_2 = (X_{21}, X_{22}, \dots, X_{2n_2})$ be the expression level of gene probes under treatment group. The total sample size is $n = n_1 + n_2$ be. Let $\mu_1(X_c)$ and $\sigma^2_{X_c}$ be the gene-specific mean and variance of the expression levels of gene probe under condition c (i.e., $c = 1$ for control group, and $c = 2$ for treatment group). And let $\mu_2(X_c) = E(X_c^2)$ be the second-order moment of X_c. According to the definition of second order moment, $\mu_2(X_c) = \left(\mu_1(X_c)\right)^2 + \sigma^2_{X_c}$.

Without loss of generality, we assume $\overline{X_1} < \overline{X_2}$. Define $Y_{1i} = X_{1i} - \overline{X_1}$ and $Y_{2j} = X_{2j} - \overline{X_1}$, where $i = 1,2, \dots, n_1, j = 1,2, \dots, n_2, \overline{X_1} = \sum_{i=1}^{n_1} X_{1i}$ and $\overline{X_2} = \sum_{j=1}^{n_2} X_{2j}$. Firstly, we constructed the first welch *t* test statistic to capture the mean heterogeneity between two groups.

Primary Test

$$T_{w_1} = \frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\frac{S_{Y_1}^2}{n_1} + \frac{S_{Y_2}^2}{n_2}}}$$

where $\overline{Y_1} = \sum_{i=1}^{n_1} Y_{1i}$, $\overline{Y_2} = \sum_{j=1}^{n_2} Y_{2j}$ and $S_{Y_1}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y_1})^2$, $S_{Y_2}^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(Y_{2j} - \overline{Y_2}\right)^2$. T_{w_1} is equivalent to $\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{S_{X_1}^2}{n_1}+\frac{S_{X_2}^2}{n_2}}}$. Next, we constructed another welch *t* type test statistic to capture the second-order heterogeneity between two groups as auxiliary test below.

Auxiliary Test

$$T_{w_2} = \frac{\overline{Y_1^2} - \overline{Y_2^2}}{\sqrt{\frac{S_{Y_1}^2}{n_1} + \frac{S_{Y_2}^2}{n_2}}}$$

where $\overline{Y_1^2} = \sum_{i=1}^{n_1} Y_{1i}^2$, $\overline{Y_2^2} = \sum_{j=1}^{n_2} Y_{2j}^2$ and $S_{Y_1^2}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(Y_{1i}^2 - \overline{Y_1^2}\right)^2$, $S_{Y_2^2}^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(Y_{2j}^2 - \overline{Y_2^2}\right)^2$. Under H_{05}: $\mu_1(X_1) = \mu_1(X_2)$ and $\mu_2(X_1) = \mu_2(X_2)$, we demonstrated the asymptotical independence of T_{w_1} and T_{w_2} as

$$\begin{pmatrix} T_{w1} \\ T_{w2} \end{pmatrix} \xrightarrow{a.d.} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

This conclusion can be mathematically proved when $n_1 = n_2$ and $n_1, n_2 \longrightarrow \infty$ (See **Appendix B.4**). When H_{05} is false, T_{w_1} and T_{w_2} are dependent. This property guaranteed the control of Type I error rate under null hypothesis and the potential power gain under alternative hypothesis. Based on the null independence, we adopted Fisher's method to define the DWT statistic as

$$DWT = -2\left(\log\left(p_{w_1}\right) + \log\left(p_{w_2}\right)\right)$$

Where p_{w_1} and p_{w_2} are the p value for the Welch statistics T_{w_1} and T_{w_2}. DWT follows approximately the χ^2- distribution with 4 degrees of freedom when n_1 and n_2 are large sample sizes. In reality, n_1 and n_2 are usually limited sample sizes. In next result section of Type I error rate, we would show that DWT method can also be applied to the moderate sample sizes and can still generally control Type I error rates under H_{05}.

3.3.3 Results

3.3.3.1 Type I error rate controls of competitors

Under normality setting, we generated 100000 replicates of two-group samples from the standard normal distribution with sample size $n_1 = n_2 = 5,10,20,40$. WT, MWT, STSD, IMVT, SMVT and DWT are the competitors here. With extremely small samples, none of the six competitors could properly control type I error rates (**Figure 3-12 (a)**). The STSD severely inflated type I error rates. And the IMVT, SMVT and DWT appeared anti-conservative; all the three methods were much less inflated than the STSD. The MWT, MT and WT performed slightly conservative equally. The serious inflation of the STSD is due to the variability of condition-specific data standardization. In **Figure 3-12 (b)**, DWT and STSD are still a little inflated while the other four methods are slightly anti-conservative. For moderate sample sizes (**Figure 3-12 (c-d)**), all the six methods seemed generally controlled the Type I error rate.

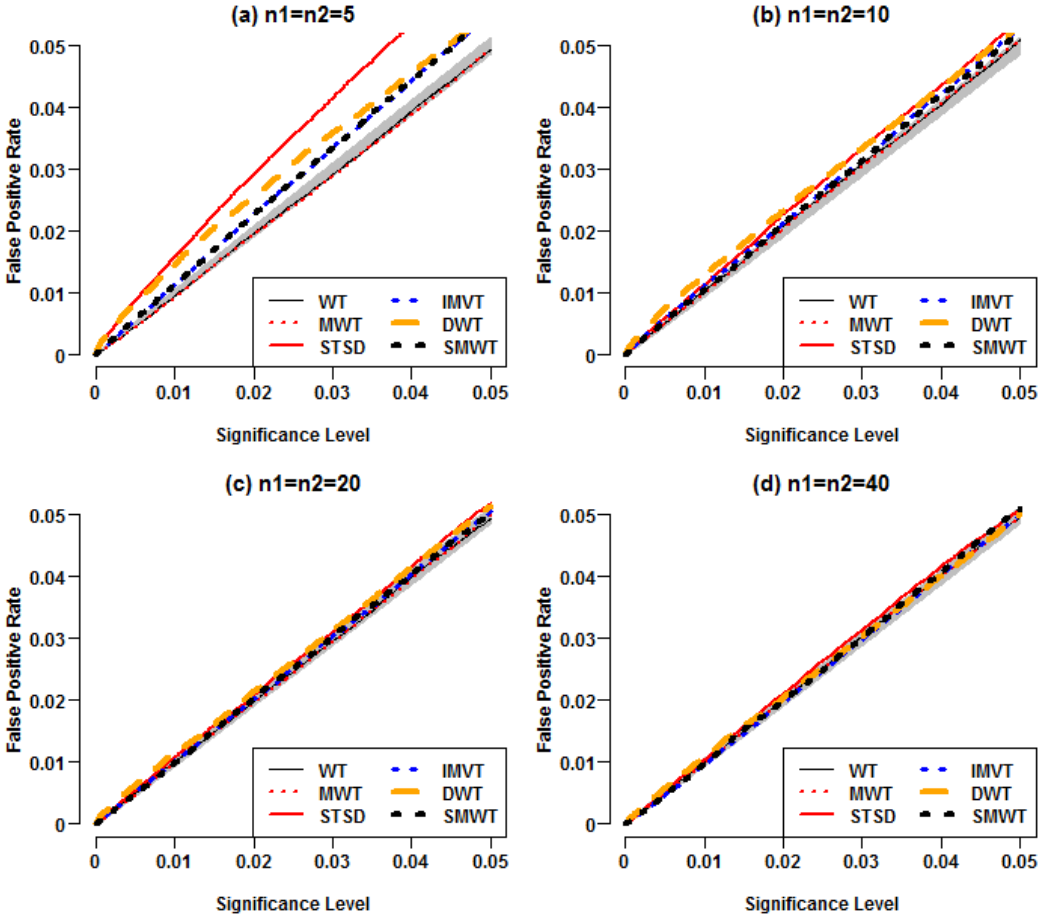

Figure 3-12: Comparison of false positive rates of six methods under standard normality setting.

3.3.3.2 **Empirical power comparisons**

Similar as that in Part I, we simulated independently 10000 replicates of $n_1 = 40$ data points from normal distribution $\mathcal{N}(0,1)$ and $n_2 = 40$ data points from $\mathcal{N}(r, (1+s)^2)$ for each (r, s) pair. The parameters r and s represent the magnitudes of mean and variance heterogeneities, respectively. In gene co-expression networks, few genes work independently and genes can interact with each other and/or interact with environmental factors. Susceptible genes can co-express as indicated by gene-gene correlations. Such

correlations and interactions among biological networks are very common and are major drivers for the high-order heterogeneity of testing susceptible gene. Therefore, for power comparisons in Part II, we investigated three kinds of scenarios with different mean heterogeneities levels ($r = 0.25, 0.5, 0.75$). For each scenario, we presented the power comparisons of six methods with different variance heterogeneities ($s = 0, 0.1, 0.2, 0.3, 0.4, 0.5$). The nominal level α is set to be 5×10^{-3} to obtain the reasonable powers.

When no mean heterogeneity existed ($r = 0$), the DWT displayed the highest powers, followed by the IMVT and SMVT which presented the similar powers; and all the three methods outperformed the three mean heterogeneity tests, i.e., the WT, the MWT and the STSD (**Figure 3-13 (a)**). When the mean heterogeneity existed ($r \neq 0$), WT, MWT and STSD are slightly more powerful than DWT method, followed by IMVT and SMVT with small variance heterogeneity. With the increase of variance heterogeneity, the DWT is always the first to surpass the three mean heterogeneity tests and remained the most powerful compared to IMVT and SMVT (**Figure 3-13 (b-c)).** In addition, the power of WT, MWT and STSD remained decreasing with the increase of variance heterogeneity (**Figure 3-13 (b-c)**). For these three situations, the power gains of our DWT method over the three mean heterogeneity tests appeared especially noteworthy with the increase of variance heterogeneity and it did not display severe power losses with trivial or no variance heterogeneity. In addition, DWT is always more powerful than IMVT and SMVT methods.

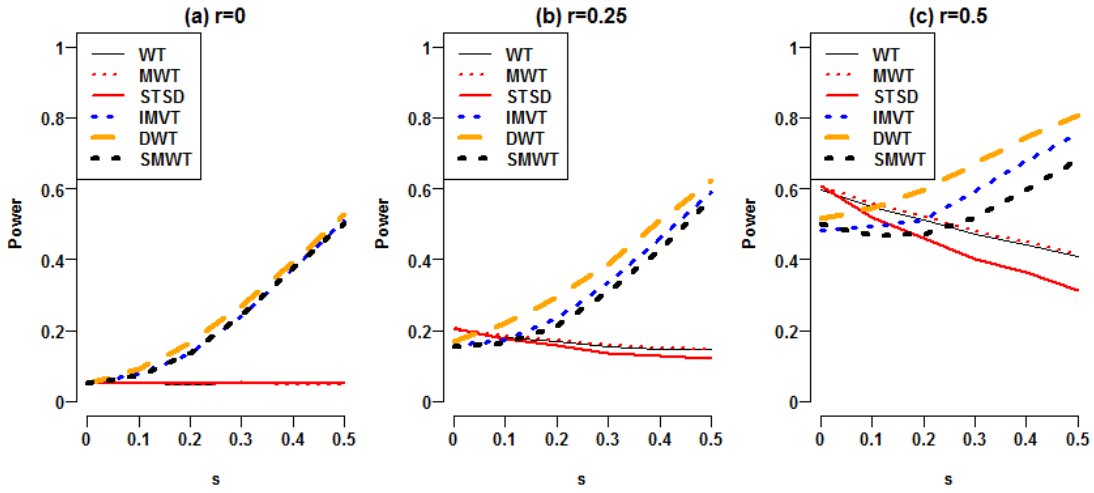

Figure 3-13: Power comparison of six methods with different mean heterogeneities levels at nominal level 0.05

3.3.3.3 Advantage of DWT over IMVT

As discussed in Part I, the mean heterogeneity tests and variance heterogeneity tests are always independent under both null hypothesis and alternative hypothesis. When trivial or no variance heterogeneity existed, IMVT integrating variance heterogeneity were not capable of overcoming the penalty of increasing the degree of freedom of integrative test. Unlike IMVT that utilized the variance heterogeneity, DWT integrates second-order heterogeneity that is made up of both mean and variance heterogeneity. The auxiliary test we constructed to detect second-order heterogeneity is independent of mean heterogeneity test under null hypothesis to guarantee the control of type I error rates. In contrast, it is dependent of mean heterogeneity test under alternative hypothesis. This alternative dependence between mean heterogeneity test and second-order heterogeneity test of DWT would lead to more power gain than IMVT even without variance heterogeneity. When no variance heterogeneity existed ($s = 0$), the DWT is always more

powerful than IMVT with different mean heterogeneity ($r = 0.1, 0.2, 0.3, \ldots, 1$) at different nominal levels (**Figure 3-14**)

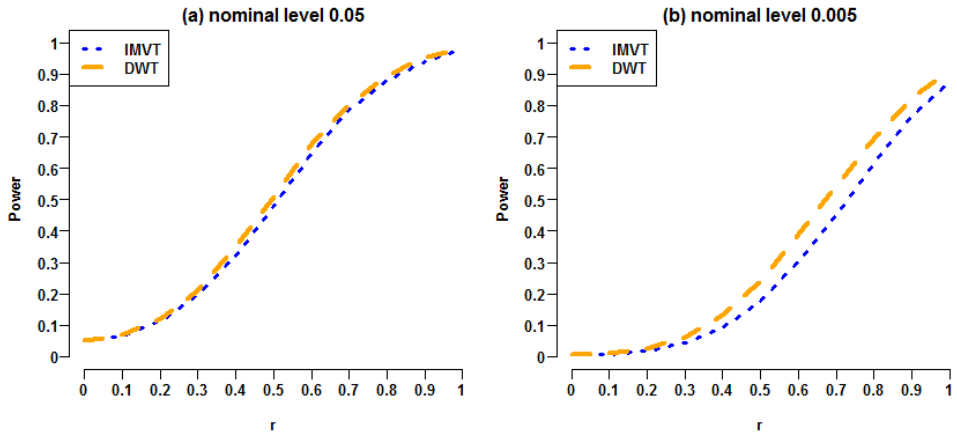

Figure 3-14: Power comparison of DWT and IMVT at nominal level 0.05 and 0.005, respectively

3.3.3.4 **Replication of previously reported gene probes that involve in functional network**

The expression distribution of a gene involve in network cannot be solely determined by its mean. Therefore Integrating informative high-order heterogeneity is a more powerful method to identify genes that involve in gene-gene co-expression and interaction networks then existent mean heterogeneity methods. To illustrate DWT's performance in detecting function genes, we still used the gene expressions profiles of peripheral circulating B cells between 39 smoking and 40 non-smoking healthy US white women by Pan *et al*. The same data processing procedures were conducted to adjust for background data structure.

Pan et *al*. reported 33 gene probes to involve in constructed functional network. We applied DWT method to the calibrated expressions and replicated 19 out of the 33 reported gene probes that involved in network. DWT obtained smaller p values compared to IMVT in the majority of the 19 gene probes.

Table 3-3: The overlap of the discoveries of DWT and the genes which were testified to be involved in functional networks

AffyID	Gene	Adjusted MAS5*				
		DWT	**IMVT**	**MWT**	**WT**	**STSD**
201085_s_at	*SON*	0.0090	0.0075	0.0021	0.0023	0.0021
203868_s_at	*VCAM1*	0.0005	0.0030	0.0005	0.0005	0.0005
204600_at	*EPHB3*	0.0025	0.0178	0.0207	0.0213	0.01654
205008_s_at	*CIB2*	0.0209	0.0387	0.0117	0.0123	0.0122
205099_s_at	*CCR1*	0.0033	0.0058	0.0160	0.0165	0.0105
206788_s_at	*CBFB*	4.63E-05	0.0003	4.28E-05	4.71E-05	4.34E-05
207961_x_at	*MYH11*	0.0038	0.0001	0.0370	0.0383	0.0139
208164_s_at	*IL9R*	0.0052	0.0311	0.0072	0.0077	0.0074
209876_at	*GIT2*	0.0002	0.0024	0.0053	0.0057	0.0040
211197_s_at	*ICOSLG*	0.0138	0.0448	0.0479	0.0487	0.0423
212514_x_at	*DDX3X*	3.92E-05	0.0002	2.73E-05	3.10E-05	3.06E-05
213446_s_at	*IQGAP1*	0.0289	0.0082	0.0400	0.0413	0.0360
217557_s_at	*CPM*	0.0357	0.0347	0.2678	0.2701	0.2422
219599_at	*EIF4B*	0.0003	0.0006	0.0018	0.0019	0.0005
208224_at	*HOXB1*	0.0093	0.0603	0.0437	0.0446	0.0365
215530_at	*FANCA*	0.0358	0.0829	0.0244	0.0251	0.0245
207844_at	*IL13*	0.0174	0.1102	0.0311	0.0318	0.0307
216647_at	*TCF3*	0.0399	0.0711	0.0144	0.0150	0.0150
210883_x_at	*EFNB3*	0.0107	0.0707	0.0284	0.0291	0.0275

***These raw *p* values of the heterogeneity tests based on the calibrated expression levels after adjusting for age, menopausal status, and the background structure.**

3.4 Conclusion and Discussion

In Part I, we illustrated that integrating informative variance heterogeneity holds tremendous potential to identify novel genes which involve in gene-gene co-expression and interaction networks. Susceptible genes can co-express as indicated by gene-gene correlations [98, 99]. Genes can interact with each other and/or interact with environmental factors. For example, Pan et *al*. [103] reported 33 gene probes to involve in constructed functional network. Among which, independent studies reported *MYH11*, *HOXB1*, *GIT2*, *VCAM1*, *CCR1*, *IQGAP1*, *PDPK1*, *HBA1 HBA2*, *SON*, and *CPM* to involve in networks related to lung cancer and smoking [112-117]. Within a complex network, the distribution change in the expression level of a single susceptible gene cannot determined by its mean heterogeneity completely. Higher-order heterogeneities can provide extra valuable information for the distribution change. This is why the IMVT led to smaller *p* values than did existent mean heterogeneity tests in our data analyses. In conclusion, integrating informative variance heterogeneity proved an effective step to better capture the latent information conveyed by the co-expression and interaction networks of susceptible genes. It represents one efficient way to extract the inherent higher-order information as induced by complex networks of multiple biomarkers.

The IMVT aims to identify genes whose expression distributions are susceptible to the change in condition. It does not distinguish informative variance heterogeneity from mean heterogeneity. Before applying the IMVT, background data structures must be calibrated to prevent false positive discoveries and power loss. Data structure can be a major confounder for differential analyses, as illustrated by our reanalysis of Pan *et al.*'s gene profiles [103]. The discrepancy between Pan et al.'s and our discoveries showed the

severe confounding impact of the global data structure on differential analyses. In a judicious data calibration, the data structure should be computed from random genes to prevent power loss due to over adjustment.

The IMVT and the SMVT as well, inherit the advantages and disadvantages of the Levene test and the WT. The Levene test is a robust non-parametric method. The exact distribution of the Levene statistic is intractable, and thus its *p*-value must be evaluated by its asymptotic distribution. The condition-specific variance estimators in the Welch statistic could not be accurate for small samples. Thus, the current IMVT is suitable for large samples other than small samples. By our simulation studies and the work of Demissie *et al*. [97], the MWT could outperform the WT, especially for extremely small sample sizes. Novel parametric methods, i.e., the LRT, are needed to mine expression files of low-replicate experiments. However, the test statistic and its exact null distribution of a parametric test statistic depend on the exact distributions of the (transformed/calibrated) gene expression levels. It is intractable to learn the exact distributions of gene expressions from small samples. Model miss-specifications can mess up differential analyses, as showed by the severe inflations in type I error rate of the normality-based LRT under the Laplace settings. The development of effective small-sample tests requires further formal efforts. In addition, appropriate adjustment of background data structures and other hidden confounders are important for the success of effectively integrating informative variance heterogeneity instead of spurious variance heterogeneity.

Lastly, we acknowledge that there is no need to consider variance heterogeneity in case the distribution of the expression measure of a gene can be determined by a single parameter, i.e., its mean. In such a case, the IMVT can be less powerful than the Welch

test. However, single-parameter distribution cannot well fit real-world expression levels in general.

In part II, we put forward an alternative of our IMVT method - DWT that integrated the mean heterogeneities and high-order heterogeneities. Utilizing second-order heterogeneities instead of variance heterogeneities would further improve the detecting power of candidate MVDE genes that have a high possibility of involving in gene networks. Due to the high complexity of gene networks, the expression distribution of a gene cannot be solely determined by its mean. Distribution heterogeneity is a much bigger umbrella than mean heterogeneity. The proposed IMVT and DWT methods merely made one step further from traditional mean heterogeneity tests. High-order heterogeneities are quite common and require particular exploitation methods.

APPENDIX A

SUPPLEMENTARY OF HARMONIOUS SIGNAL AUGMENTATION SCHEMES IN ASSOCIATION TESTS OF DNA SEQUENCE

A.1 Proof of Proposition about asymptotic joint distribution of $\mathbf{T_1}$ and $\mathbf{T_2}$

Proposition: *Under primary model, if* $E(e_i^4) < \infty$ *and* $E(e_i) = E(e_i^3) = 0$, *then* $T_1 - \delta_1$ *and* $T_2 - \delta_2$ *converge in distribution to a bivariate normal distribution with unit variance and correlation coefficient* $\rho = \beta_1\delta_1^{-1}\delta_2^{-1}[\sigma_e^2(3\mu_4 - 3\mu_2^2 + \mu_1^2\mu_2 - \mu_1\mu_3) + \beta_1^2(\mu_6 - 2\mu_2\mu_4 + \mu_1\mu_5 - \mu_1^2\mu_4 + \mu_2^3 - \mu_1\mu_2\mu_3 + \mu_1^2\mu_2^2)]$, $in\ which\ \mu_k \overset{\text{def}}{=} E(G^k) for\ integer\ k\ and\ var(e_i) \overset{\text{def}}{=} \sigma_e^2$.

$$\begin{pmatrix} T_1 - C_1 \\ T_2 - C_2 \end{pmatrix} \xrightarrow{a.d.} \boldsymbol{\mathcal{N}}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where C_1 *and* C_2 *are function of* β_1 *and* β_2, *respectively.*

Proof:

Without loss of generality, let Y_i be the trait residual of individual i after adjusting for global covariates. Let individual i have G_i copies of the minor allele at a single test marker. In single-SNP association analysis, Y_i relates to G_i by linear model

$$Y_i = G_i\beta_1 + e_i, \qquad \textbf{Eq. A-1}$$

where β_1 is the regression coefficient, and e_i is regression residual such that $E(e_i^4) < \infty$ and $E(e_i) = E(e_i^3) = 0$. Based on **Eq. A-1**, we have

$$Y_i^2 = (G_i\beta_1 + e_i)^2 = \sigma_e^2 + G_i^2\beta_2 + \varepsilon_i, \qquad \textbf{Eq. A-2}$$

where $\beta_2 = \beta_1^2$ and $\varepsilon_i = 2\beta_1 G_i e_i + e_i^2 - \sigma_e^2$ has a mixed distribution with mean 0 and variance $\sigma_\varepsilon^2 = 2\left(1 + 2\beta_1^2 E\left(G^2\right)\right)\sigma_e^2$. Under HWE, $E(G^2) = 2(1+f)f$, where f is the minor allele frequency at the SNP. If G is associated with Y ($\beta_1 \neq 0$), then G^2 is also associated with Y^2 ($\beta_2 \neq 0$). The association statistic for testing $H_{01}: \beta_1 = 0$ is given by **Eq. 1-2**. The association statistic for testing $H_{02}: \beta_2 = 0$ is given by **Eq. 1-3**.

Substituting equations **Eq. A-1** and **Eq. A-2** into the definition of $\hat{\sigma}_{Y,G}$, we derive

$$\begin{aligned}
\hat{\sigma}_{Y,G} &= \frac{1}{n}\sum_{i=1}^{n}[\beta_1(G_i - \bar{G}) + (e_i - \bar{e})](G_i - \bar{G}) \\
&= \beta_1\left[\frac{1}{n}\sum_{i=1}^{n} G_i^2 - (\bar{G})^2\right] + \frac{1}{n}\sum_{i=1}^{n} G_i e_i - \bar{G}\bar{e} \\
&= \beta_1\left(\overline{G^2} - E(G^2) + E(G^2)\right) \\
&+\left(\overline{Ge} - E(Ge) + E(Ge)\right) \\
&-\beta_1\left((\bar{G})^2 - \bar{G}E(G) + \bar{G}E(G)\right) \\
&-\left(\bar{G}\bar{e} - \bar{G}E(e) + \bar{G}E(e)\right) \\
&= \beta_1\left(\overline{G^2} - E(G^2)\right) + \left(\overline{Ge} - E(Ge)\right) \\
&-\beta_1\bar{G}\left(\bar{G} - E(G)\right) - \bar{G}\left(\bar{e} - E(e)\right) \\
&+\beta_1 E(G^2) + E(Ge) - \beta_1\bar{G}E(G) - \bar{G}E(e) \qquad \textbf{Eq. A-3}
\end{aligned}$$

It follows from **Eq. 1-2** and **Eq. A-3** that

$$T_1 - C_1 = \frac{(\beta_1, 1, \beta_1\bar{G}, -\bar{G})}{D_1}\begin{pmatrix}\sqrt{n}\left(\overline{G^2} - E(G^2)\right)\\ \sqrt{n}\left(\overline{Ge} - E(Ge)\right)\\ \sqrt{n}\left(\bar{G} - E(G)\right)\\ \sqrt{n}\left(\bar{e} - E(e)\right)\end{pmatrix}, \qquad \textbf{Eq. A-4}$$

where

$$C_1 = \frac{\beta_1 E(G^2) + E(Ge) - \beta_1\bar{G}E(G) - \bar{G}E(e)}{D_1} = \frac{\beta_1[E(G^2) - \bar{G}E(G)]}{D_1}$$

And

$$D_1 = \sqrt{\hat{\sigma}_Y^2\hat{\sigma}_G^2 - \hat{\sigma}_{Y,G}^2}.$$

Similarly, substituting equations **Eq. A-1** and **Eq. A-2** into the definition of $\hat{\sigma}_{Y^2,G^2}$, we derive

$$\begin{aligned}\hat{\sigma}_{Y^2,G^2} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - \overline{Y^2}\right)\left(G_i^2 - \overline{G^2}\right)\\ &= \beta_1^2\left(\overline{G^4} - E(G^4)\right) + \beta_1^2 E(G^4)\\ &-\beta_1^2\overline{G^2}\left(\overline{G^2} - E(G^2)\right) - \beta_1^2\overline{G^2}E(G^2)\\ &+2\beta_1\left(\overline{G^3e} - E(G^3e)\right) + 2\beta_1 E(G^3e)\\ &-2\beta_1\overline{G^2}\left(\overline{Ge} - E(Ge)\right) - 2\beta_1\overline{G^2}E(Ge)\\ &+\left(\overline{G^2e^2} - E(G^2e^2)\right) + E(G^2e^2)\\ &-\overline{G^2}\left(\overline{e^2} - E(e^2)\right) - \overline{G^2}E(e^2)\end{aligned} \qquad \textbf{Eq. A-5}$$

Then we can obtain

$$T_2 - C_2$$

$$= \frac{\left(\beta_1^2, -\beta_1^2\overline{G^2}, 2\beta_1, -2\beta_1\overline{G^2}, 1, -\overline{G^2}\right)}{D_2} \begin{pmatrix} \sqrt{n}\left(\overline{G^4} - E(G^4)\right) \\ \sqrt{n}\left(\overline{G^2} - E(G^2)\right) \\ \sqrt{n}\left(\overline{G^3 e} - E(G^3 e)\right) \\ \sqrt{n}\left(\overline{Ge} - E(Ge)\right) \\ \sqrt{n}\left(\overline{G^2 e^2} - E(G^2 e^2)\right) \\ \sqrt{n}\left(\overline{e^2} - E(e^2)\right) \end{pmatrix}, \quad \textbf{Eq. A-6}$$

where

$$C_2 = \frac{\beta_1^2 E(G^4) - \beta_1^2\overline{G^2}E(G^2) + 2\beta_1 E(G^3 e) - 2\beta_1\overline{G^2}E(Ge) + E(G^2 e^2) - \overline{G^2}E(e^2)}{D_2}$$

$$= \frac{\beta_1^2 E(G^4) - \beta_1^2\overline{G^2}E(G^2) + E(G^2 e^2) - \overline{G^2}E(e^2)}{D_2}$$

And

$$D_2 = \sqrt{\hat{\sigma}_{Y^2}^2 \hat{\sigma}_{G^2}^2 - \hat{\sigma}_{Y^2,G^2}^2}$$

Using **Eq. A-4** and **Eq. A-6**, we write

$$\begin{pmatrix} T_1 - C_1 \\ T_2 - C_2 \end{pmatrix} = \boldsymbol{A_n Z}_n, \quad \textbf{Eq. A-7}$$

where

$$\boldsymbol{A_n} = \begin{pmatrix} 0 & \frac{\beta_1}{D_1} & 0 & \frac{1}{D_1} & 0 & \frac{\beta_1 \bar{G}}{D_1} & 0 & -\frac{\bar{G}}{D_1} \\ \frac{\beta_1^2}{D_2} & -\frac{\beta_1^2\overline{G^2}}{D_2} & \frac{2\beta_1}{D_2} & -\frac{2\beta_1\overline{G^2}}{D_2} & \frac{1}{D_2} & 0 & -\frac{\overline{G^2}}{D_2} & 0 \end{pmatrix}$$

and

$$\boldsymbol{Z}_n = \begin{pmatrix} \sqrt{n}\left(\overline{G^4} - E(G^4)\right) \\ \sqrt{n}\left(\overline{G^2} - E(G^2)\right) \\ \sqrt{n}\left(\overline{G^3 e} - E(G^3 e)\right) \\ \sqrt{n}\left(\overline{Ge} - E(Ge)\right) \\ \sqrt{n}\left(\overline{G^2 e^2} - E(G^2 e^2)\right) \\ \sqrt{n}\left(\bar{G} - E(G)\right) \\ \sqrt{n}\left(\overline{e^2} - E(e^2)\right) \\ \sqrt{n}\left(\bar{e} - E(e)\right) \end{pmatrix}$$

According to standard asymptotic normality theorem,

$$\boldsymbol{Z}_n \xrightarrow{a.d.} \mathcal{N}_8(\mathbf{0}, \boldsymbol{\Lambda}). \qquad \textbf{Eq. A-8}$$

Let $\mu_k \overset{\text{def}}{=} E(G^k)$ and $\varsigma_k \overset{\text{def}}{=} E(e^k)$ for integer k. $\boldsymbol{\Lambda} = \left(\lambda_{ij}\right)$ is the variance-covariance matrix of random vector $(G^4, G^2, G^3 e, Ge, G^2 e^2, G, e^2, e)'$. We derive explicit formulae of λ_{ij}'s as below. Specifically,

$$\lambda_{11} = Var(G^4) = \mu_8 - \mu_4^2,$$

$$\lambda_{12} = Cov(G^4, G^2) = \mu_6 - \mu_4\mu_2,$$

$$\lambda_{13} = Cov(G^4, G^3 e) = \varsigma_1(\mu_7 - \mu_3\mu_4) = 0,$$

$$\lambda_{14} = Cov(G^4, Ge) = \varsigma_1(\mu_5 - \mu_4\mu_1) = 0,$$

$$\lambda_{15} = Cov(G^4, G^2 e^2) = \varsigma_2(\mu_6 - \mu_4\mu_2),$$

$$\lambda_{16} = Cov(G^4, G) = \mu_5 - \mu_4\mu_1,$$

$$\lambda_{17} = Cov(G^3, e^2) = 0,$$

$$\lambda_{18} = Cov(G^3, e) = 0,$$

$$\lambda_{22} = Var(G^2) = \mu_4 - \mu_2^2,$$

$$\lambda_{23} = Cov(G^2, G^3 e) = \varsigma_1(\mu_5 - \mu_2\mu_3) = 0,$$

$$\lambda_{24} = Cov(G^2, Ge) = \varsigma_1(\mu_3 - \mu_2\mu_1) = 0,$$

$$\lambda_{25} = Cov(G^2, G^2e^2) = \varsigma_2(\mu_4 - \mu_2^2),$$

$$\lambda_{26} = Cov(G^2, G) = \mu_3 - \mu_2\mu_1,$$

$$\lambda_{27} = Cov(G^2, e^2) = 0,$$

$$\lambda_{28} = Cov(G^2, e) = 0,$$

$$\lambda_{33} = Var(G^3e) = \mu_6\varsigma_2 - \mu_3^2\varsigma_1^2 = \mu_6\sigma_e^2,$$

$$\lambda_{34} = Cov(G^3e, Ge) = \varsigma_2\mu_4 - \varsigma_1^2\mu_1\mu_3 = \mu_4\sigma_e^2,$$

$$\lambda_{35} = Cov(G^3e, G^2e^2) = \mu_5\varsigma_3 - \varsigma_1\varsigma_2\mu_2\mu_3 = 0,$$

$$\lambda_{36} = Cov(G^3e, G) = \varsigma_1(\mu_4 - \mu_1\mu_3) = 0,$$

$$\lambda_{37} = Cov(G^3e, e^2) = \mu_3(\varsigma_3 - \varsigma_1\varsigma_2) = 0,$$

$$\lambda_{38} = Cov(G^3e, e) = \mu_3(\varsigma_2 - \varsigma_1^2) = \mu_3\sigma_e^2,$$

$$\lambda_{44} = Var(Ge) = \mu_2\varsigma_2 - \mu_1^2\varsigma_1^2 = \mu_2\sigma_e^2,$$

$$\lambda_{45} = Cov(Ge, G^2e^2) = \mu_3\varsigma_3 - \mu_1\mu_2\varsigma_1\varsigma_2 = 0,$$

$$\lambda_{46} = Cov(Ge, G) = \varsigma_1(\mu_2 - \mu_1^2) = 0,$$

$$\lambda_{47} = Cov(Ge, e^2) = \mu_1(\varsigma_3 - \varsigma_1\varsigma_2) = 0,$$

$$\lambda_{48} = Cov(Ge, e) = \mu_1(\varsigma_2 - \varsigma_1^2) = \mu_1\sigma_e^2,$$

$$\lambda_{55} = Var(G^2e^2) = \mu_4\varsigma_4 - \mu_2^2\varsigma_2^2,$$

$$\lambda_{56} = Cov(G^2e^2, G) = \varsigma_2(\mu_3 - \mu_1\mu_2),$$

$$\lambda_{57} = Cov(G^2e^2, e^2) = \mu_2(\varsigma_4 - \varsigma_2^2),$$

$$\lambda_{58} = Cov(G^2e^2, e) = \mu_2(\varsigma_3 - \varsigma_1\varsigma_2) = 0,$$

$$\lambda_{66} = Var(G) = \mu_2 - \mu_1^2,$$

$$\lambda_{67} = Cov(G, e^2) = 0,$$

$$\lambda_{68} = Cov(G, e) = 0,$$

$$\lambda_{77} = Var(e^2) = \varsigma_4 - \varsigma_2^2,$$

$$\lambda_{78} = Cov(e^2, e) = \varsigma_3 - \varsigma_2\varsigma_1 = 0,$$

and

$$\lambda_{88} = Var(e) = \varsigma_2 - \varsigma_1^2 = \sigma_e^2.$$

By large-number theory, when $n \to \infty$ we have

$$D_1 \xrightarrow{Pr.} \delta_1 = \sqrt{\sigma_Y^2\sigma_G^2 - \sigma_{Y,G}^2},$$

$$D_2 \xrightarrow{Pr.} \delta_2 = \sqrt{\sigma_{Y^2}^2\sigma_{G^2}^2 - \sigma_{Y^2,G^2}^2},$$

$$\bar{G} \xrightarrow{Pr.} \mu_1 = E(G),$$

and

$$\overline{G^2} \xrightarrow{Pr.} \mu_2 = E(G^2).$$

It follows that

$$\boldsymbol{A_n} \xrightarrow{Pr.} \boldsymbol{A}$$

$$= \begin{pmatrix} 0 & \frac{\beta_1}{\delta_1} & 0 & \frac{1}{\delta_1} & 0 & \frac{\beta_1\mu_1}{\delta_1} & 0 & -\frac{\mu_1}{\delta_1} \\ \frac{\beta_1^2}{\delta_2} & -\frac{\beta_1^2\mu_2}{\delta_2} & \frac{2\beta_1}{\delta_2} & -\frac{2\beta_1\mu_2}{\delta_2} & \frac{1}{\delta_2} & 0 & -\frac{\mu_2}{\delta_2} & 0 \end{pmatrix}. \qquad \textbf{Eq. A-9}$$

According to Slutsky's theorem, we obtain from **Eq. A-8** and **Eq. A-9** that

$$\begin{pmatrix} T_1 - C_1 \\ T_2 - C_2 \end{pmatrix} = \boldsymbol{A}_n\boldsymbol{Z}_n \xrightarrow{a.d.} \mathcal{N}_2(\mathbf{0}, \boldsymbol{A\Lambda A}'), \qquad \textbf{Eq. A-10}$$

when $n \to \infty$. Algebraically, we verify that

$$\boldsymbol{A\Lambda A'} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \textbf{Eq. A-11}$$

and

$$\rho = \frac{\beta_1}{\delta_1 \delta_2}[\sigma_e^2(3\mu_4 - 3\mu_2^2 + \mu_1^2\mu_2 - \mu_1\mu_3) + \beta_1^2(\mu_6 - 2\mu_2\mu_4 + \mu_1\mu_5 - \mu_1^2\mu_4 + \mu_2^3 - \mu_1\mu_2\mu_3 + \mu_1^2\mu_2^2)]. \qquad \textbf{Eq. A-12}$$

By **Eq. A-12**, T_1 and T_2 are asymptotically dependent if $\rho \neq 0$. But if $\beta_1 = 0$, we have $\rho = 0$ together with $C_1 = 0$ and $C_2 \xrightarrow{Pr.} 0$, and therefore

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \xrightarrow{a.d.} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \qquad \textbf{Eq. A-13}$$

when $n \to \infty$. By **Eq. A-13**, T_1 and T_2 are asymptotically independent if $\beta_1 = 0$.

A.2 Supplemental Figures

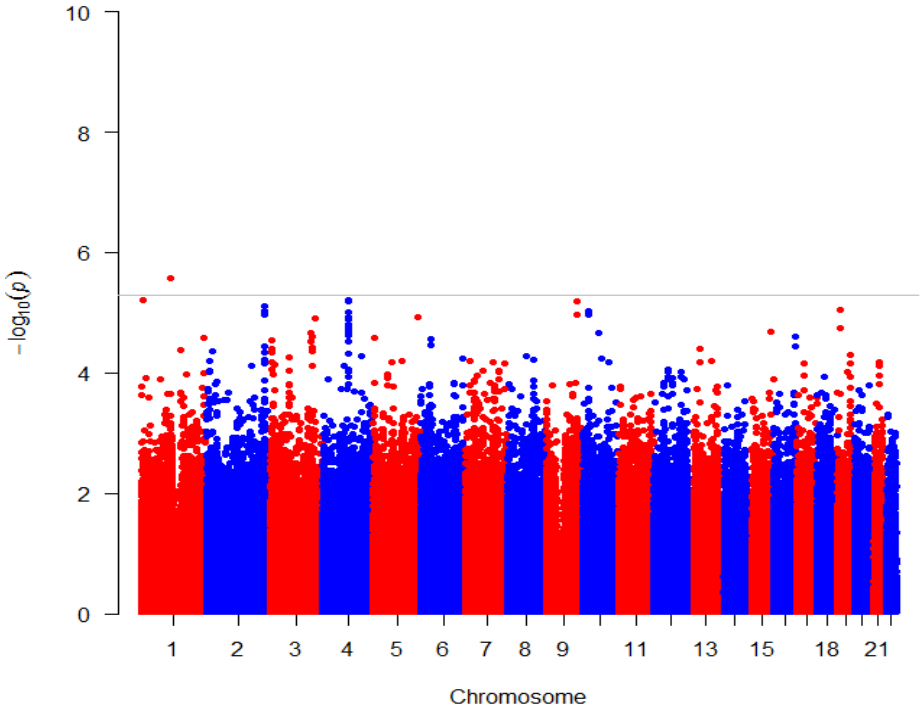

Figure A-1: The Manhattan plot of MT.

A.3 Candidate SNPs selected by HSAT

Table A-1: Top-ranked Significant SNPs by the HSAT (5×10^{-5}) .

Chr	rs	Pos	MAF	Gene	HSAT	MT	JLS	LRT
7	rs849436	106367588	0.050454	NA	5.72E-10	6.77E-05	1.04E-07	1.59E-05
6	rs2842519	38042247	0.007064	ZFAND3	2.70E-08	2.72E-05	0.00024	0.000132
6	rs1335535	79999203	0.040363	HMGN3	7.86E-08	0.027669	1.72E-06	0.003534
6	rs9350803	79999595	0.040445	HMGN3	7.86E-08	0.027669	1.72E-06	0.003534
7	rs849370	106307179	0.050556	PIK3CG	7.89E-08	0.000268	1.23E-06	6.21E-05
6	rs11963886	37970017	0.007576	ZFAND3	1.16E-07	3.41E-05	0.000298	0.00017
6	rs1537740	80035233	0.040868	NA	1.41E-07	0.031622	2.08E-06	0.004591
19	rs10402645	13058752	0.008138	NFIX	1.51E-07	9.14E-06	8.22E-05	4.91E-05

6	rs7738508	80048256	0.04002	NA	1.58E-07	0.020429	1.71E-06	0.004319
19	rs306045	2992700	0.089808	NA	1.62E-07	0.000533	1.46E-05	0.000385
19	rs11881808	13054782	0.007576	NFIX	2.25E-07	1.77E-05	0.000198	9.66E-05
6	rs7763232	80030157	0.04154	NA	2.49E-07	0.034876	2.52E-06	0.005855
7	rs849406	106320153	0.0444	PIK3CG	2.75E-07	0.000423	7.04E-06	0.000145
6	rs4706754	79969588	0.036327	HMGN3	2.84E-07	0.008009	1.75E-06	0.004203
6	rs7772967	80051380	0.043592	NA	4.05E-07	0.038085	4.22E-06	0.006727
6	rs10806163	79951373	0.039354	NA	4.35E-07	0.007589	0.000446	0.001569
6	rs9343886	79983800	0.036831	HMGN3	5.13E-07	0.00996	2.14E-06	0.00559
6	rs16890450	79949239	0.039434	NA	5.86E-07	0.006776	0.000432	0.00159
19	rs1688114	2988787	0.092012	NA	7.56E-07	0.004381	5.38E-05	0.001079
2	rs17020307	37294768	0.007107	CEBPZ	8.01E-07	0.032838	0.093755	0.11547
2	rs28548299	37340278	0.007064	PRKD3	8.01E-07	0.032838	0.093755	0.11547
7	rs2453840	45920337	0.183519	IGFBP3	1.08E-06	0.000291	4.29E-05	0.000177
19	rs1654678	2985077	0.093023	NA	1.11E-06	0.002849	5.32E-05	0.001221
6	rs2322219	80038110	0.041877	NA	1.32E-06	0.032227	0.000868	0.004562
2	rs2421738	62877847	0.00555	EHBP1	1.86E-06	0.680389	0.000271	0.005476
2	rs17027558	63065438	0.005567	EHBP1	1.86E-06	0.680389	0.000271	0.005476
7	rs849408	106329620	0.049495	PIK3CG	1.88E-06	0.000919	6.15E-06	1.25E-05
7	rs849390	106296223	0.047427	PIK3CG	1.91E-06	0.001406	2.51E-05	0.000515
2	rs2871608	57499324	0.072149	NA	2.73E-06	0.306335	5.74E-05	0.009102
6	rs1414283	80036646	0.038384	NA	2.81E-06	0.025533	4.63E-06	0.014626
9	rs10982123	116050914	0.1	COL27A1	3.95E-06	1.10E-05	8.18E-05	0.000304
2	rs16829835	151831949	0.009082	NA	4.22E-06	0.093899	0.188207	0.151819
5	rs159981	6042136	0.155051	NA	4.28E-06	0.161182	0.000116	0.002683
7	rs4236534	96311548	0.192929	NA	4.75E-06	0.344937	0.000159	0.004283
8	rs6988232	121699353	0.10101	SNTB1	5.23E-06	0.01313	0.001271	0.000508
7	rs940823	17102971	0.006067	NA	5.38E-06	6.36E-05	0.000466	0.000257
2	rs11687001	76515912	0.029828	NA	5.55E-06	0.000287	0.002581	0.003707
5	rs2434738	6047196	0.153455	NA	5.98E-06	0.192654	0.00016	0.003367
7	rs6463939	1581188	0.005139	NA	6.02E-06	0.157577	0.011945	0.009269
1	rs6687647	111326016	0.129424	NA	6.07E-06	2.65E-06	2.10E-05	8.47E-05
20	rs1535253	19630909	0.311806	SLC24A3	6.11E-06	0.008563	5.97E-05	4.52E-05
4	rs16875269	23796971	0.014632	NA	6.31E-06	0.00779	4.67E-05	0.001791
4	rs11734262	61206591	0.290274	NA	6.55E-06	0.005641	0.000892	0.001085
2	rs2264692	57585448	0.077195	NA	6.70E-06	0.818274	0.001112	0.015968
3	rs6775197	38729651	0.018668	SCN10A	6.87E-06	0.081582	0.014534	0.03464
4	rs11931196	122194095	0.024242	C4orf31	7.21E-06	0.066794	0.055587	0.076014
2	rs1347861	139765677	0.077778	NA	7.49E-06	0.081001	2.92E-05	0.001484
4	rs2349960	161459516	0.054711	NA	8.10E-06	0.546463	0.000187	0.023878

20	rs6112527	19625176	0.427851	SLC24A3	8.15E-06	0.000197	4.05E-05	0.000286
3	rs13060227	144022179	0.035318	PCOLCE2	8.95E-06	0.068773	2.53E-07	0.000158
8	rs6986444	121695275	0.10202	SNTB1	9.60E-06	0.015676	0.001301	0.000505
6	rs426133	149839810	0.04596	ZC3H12D	1.03E-05	0.228343	0.029236	0.028184
21	rs2826498	21054936	0.119072	NA	1.06E-05	0.009687	0.000332	0.00184
5	rs16901192	31491067	0.0111	RNASEN	1.07E-05	0.005038	0.004793	0.003423
5	rs288837	73499743	0.279011	NA	1.07E-05	0.002054	0.001096	0.00124
7	rs2453839	45920098	0.196465	IGFBP3	1.11E-05	0.000111	0.000282	0.001004
3	rs6786387	2797150	0.06559	CNTN4	1.12E-05	4.57E-05	9.35E-05	0.000978
5	rs12173038	31520652	0.012626	RNASEN	1.13E-05	0.022093	0.003145	0.003405
4	rs10516521	106772696	0.052977	FLJ20184	1.22E-05	0.000735	0.000385	0.003244
21	rs7283239	21062388	0.120585	NA	1.23E-05	0.01708	0.000348	0.001853
6	rs7738385	127819209	0.324924	KIAA0408	1.27E-05	0.000781	1.65E-05	0.000616
22	rs7289613	46862049	0.097376	NA	1.28E-05	0.00115	0.001205	0.00096
2	rs7607803	151922716	0.006098	TNFAIP6	1.30E-05	0.001628	0.011932	0.003989
19	rs3745180	58311203	0.314329	ZNF415	1.39E-05	0.007603	0.000589	0.001928
2	rs1514748	219901259	0.149849	NA	1.42E-05	7.68E-06	4.04E-05	0.000305
2	rs6725931	219913390	0.149849	NA	1.42E-05	7.68E-06	4.04E-05	0.000305
17	rs4646364	17408693	0.009586	PEMT	1.45E-05	0.006452	0.000992	0.003869
3	rs17609118	10104118	0.008089	FANCD2	1.45E-05	0.315449	0.00269	0.006777
10	rs2607830	87957082	0.211111	GRID1	1.50E-05	0.355212	0.000118	0.001939
3	rs704597	100359964	0.033367	NA	1.51E-05	0.213852	0.038338	0.023054
11	rs17486172	83253375	0.058527	DLG2	1.52E-05	0.00025	0.000425	0.003559
7	rs1403179	96307040	0.195455	NA	1.58E-05	0.402389	0.000475	0.006887
5	rs2770952	180566460	0.011134	NA	1.60E-05	0.05791	0.221341	0.251756
9	rs12347248	2896287	0.008586	NA	1.63E-05	0.061392	0.003891	0.007652
2	rs7600417	219877736	0.146317	PTPRN	1.65E-05	9.81E-06	9.82E-05	0.000486
11	rs7102041	12127193	0.064646	MICAL2	1.70E-05	0.01322	0.001021	0.007743
7	rs1636804	106379027	0.150505	NA	1.84E-05	0.034885	0.001877	0.002478
9	rs3847255	3065469	0.012109	NA	1.93E-05	0.000926	0.001702	0.001844
2	rs908194	219906491	0.150353	NA	1.93E-05	1.08E-05	5.67E-05	0.000418
7	rs2232106	43883498	0.016194	URG4	1.94E-05	0.522197	0.002729	0.003116
7	rs1724278	106376466	0.150657	NA	1.95E-05	0.033474	0.000869	0.003638
2	rs2271593	219876851	0.149849	PTPRN	1.99E-05	9.25E-06	6.39E-05	0.000401
9	rs10113990	23221298	0.019173	NA	2.01E-05	0.042497	0.169218	0.018756
1	rs947633	111330302	0.179474	NA	2.12E-05	0.001044	0.000292	0.001988
6	rs9387278	97878404	0.053481	NA	2.16E-05	0.000769	0.00012	0.001019
16	rs1540610	79019538	0.112969	LOC729847	2.16E-05	0.014814	5.16E-05	0.005243
6	rs2842518	38033650	0.015167	ZFAND3	2.17E-05	0.006342	0.002993	0.00266
4	rs12500426	95733632	0.435419	PDLIM5	2.22E-05	6.30E-06	3.56E-05	5.86E-05

3	rs9878578	8620666	0.416246	NA	2.26E-05	0.000563	0.000331	0.000928
4	rs17263714	106909495	0.042132	GSTCD	2.27E-05	0.008277	0.000378	0.002156
4	rs17264527	106960873	0.041919	GSTCD	2.27E-05	0.008277	0.000378	0.002156
10	rs6584778	108757592	0.008073	SORCS1	2.27E-05	0.004444	0.006837	0.0049
1	rs4838884	111330093	0.180851	NA	2.33E-05	0.0012	0.00033	0.00217
19	rs2112464	13393476	0.284561	CACNA1A	2.37E-05	0.423801	0.000755	0.005022
8	rs6469297	111217804	0.020182	NA	2.38E-05	0.000291	2.42E-05	0.001361
15	rs7169262	38462785	0.116162	C15orf23	2.43E-05	0.016235	0.004095	0.002904
6	rs1408913	164612796	0.085267	LOC728275	2.48E-05	0.1033	0.000472	0.007985
21	rs11909439	41475912	0.008595	BACE2	2.49E-05	0.02077	0.001507	0.00651
8	rs4464955	125691713	0.064077	MTSS1	2.56E-05	0.144627	0.000981	0.020451
6	rs7755769	103737195	0.00555	NA	2.59E-05	0.019803	0.001123	0.010545
16	rs7198517	81712286	0.266162	CDH13	2.63E-05	0.360489	0.000387	0.002633
3	rs11915300	111904344	0.019697	NA	2.66E-05	0.000862	0.00207	0.011318
3	rs9825259	111923855	0.019697	NA	2.66E-05	0.000862	0.00207	0.011318
3	rs11925026	111935960	0.019677	NA	2.66E-05	0.000862	0.00207	0.011318
1	rs12725071	105022488	0.082569	NA	2.70E-05	0.401595	0.001311	0.012736
4	rs6826001	170183269	0.114995	NA	2.77E-05	0.125259	0.006784	0.006888
9	rs3789255	115171500	0.142929	BSPRY	2.80E-05	0.000615	0.001053	0.003096
20	rs3790286	19603938	0.43441	SLC24A3	2.85E-05	0.000369	0.000136	0.000848
1	rs17838268	209807801	0.119576	NA	2.93E-05	0.492022	8.96E-05	0.001804
9	rs10819692	101561702	0.207871	NA	2.94E-05	0.021543	0.000846	0.00604
19	rs11881337	13041863	0.01665	NFIX	2.99E-05	0.000224	0.00077	0.000814
22	rs4823340	43339825	0.095046	NA	3.00E-05	0.147564	0.000836	0.013243
17	rs4924892	17344692	0.149239	NA	3.06E-05	0.053651	0.001418	0.009824
4	rs6852740	84284062	0.013636	NA	3.08E-05	0.00224	0.000844	0.002602
9	rs3827661	115171694	0.136082	BSPRY	3.17E-05	0.000989	0.001531	0.00372
4	rs17378658	95609180	0.211616	PDLIM5	3.18E-05	0.006596	0.000568	0.00183
1	rs11811690	6239708	0.012121	GPR153	3.23E-05	0.071074	0.001109	0.005539
5	rs288864	73463799	0.482846	NA	3.23E-05	6.53E-05	3.89E-05	0.000254
21	rs2255892	41960841	0.165319	NA	3.24E-05	0.055255	0.002255	0.003264
21	rs2826511	21068427	0.119697	NA	3.37E-05	0.029514	0.000877	0.003346
3	rs11926949	172577788	0.132323	TNIK	3.43E-05	1.24E-05	1.40E-05	3.31E-05
8	rs10505136	111210397	0.015167	NA	3.45E-05	0.000602	0.001801	0.007424
10	rs7093513	91970434	0.014213	NA	3.51E-05	0.780072	0.002939	0.004354
1	rs656843	111538195	0.158746	DENND2D	3.65E-05	0.001819	0.000258	0.001369
4	rs11938297	175039558	0.230303	NA	3.65E-05	0.48093	0.036322	0.007811
3	rs16851691	142912773	0.025227	NA	3.76E-05	0.002931	0.000616	0.001636
8	rs1481800	72293980	0.264646	EYA1	3.78E-05	5.27E-05	0.000168	0.000597

10	rs11016332	130253792	0.005045	NA	3.87E-05	0.89803	0.000973	0.014244
13	rs7331710	84469501	0.072811	NA	3.90E-05	6.47E-05	0.000415	0.001826
8	rs4448295	6909100	0.039899	NA	3.90E-05	0.005693	0.002043	0.017091
8	rs1369453	143845212	0.317125	LYNX1	3.93E-05	0.000911	0.00108	0.001465
8	rs16917049	53069017	0.021695	NA	4.01E-05	0.021298	0.047268	0.064028
8	rs3758081	143821375	0.324045	NA	4.09E-05	0.001378	0.002874	0.000779
2	rs4664931	151655431	0.329798	NA	4.10E-05	0.092537	0.000246	0.005188
1	rs12743480	36479573	0.006579	THRAP3	4.10E-05	0.27519	0.01572	0.012261
2	rs4386359	69107531	0.062121	ANTXR1	4.13E-05	0.238141	0.00547	0.012061
8	rs4736323	143827361	0.332659	LYPD2	4.19E-05	0.001726	0.002154	0.00101
15	rs12592245	22470643	0.071719	C15orf2	4.21E-05	0.003804	0.005316	0.006048
8	rs2738100	6780991	0.374369	DEFA4	4.23E-05	0.000153	0.000161	0.000627
15	rs8041151	92427459	0.300202	NA	4.24E-05	2.04E-05	7.16E-05	0.000365
4	rs28661179	69913518	0.006054	NA	4.25E-05	0.034357	0.001216	0.012085
22	rs10427772	43344028	0.095046	RP3-474I12.5	4.48E-05	0.191036	0.001267	0.016731
1	rs7530862	209803093	0.119072	NA	4.51E-05	0.566176	0.000151	0.002535
1	rs6682769	36098783	0.051515	NA	4.52E-05	0.065073	0.033751	0.011428
5	rs1379855	162337912	0.243814	NA	4.55E-05	0.001147	0.001442	0.003039
5	rs152439	141904579	0.077778	NA	4.62E-05	0.003998	0.005161	0.011858
2	rs12467436	166205057	0.371342	FAM130A2	4.64E-05	0.001099	1.79E-05	0.000163
17	rs12453224	10995988	0.091919	NA	4.64E-05	0.020513	0.007116	0.015387
17	rs12450301	11002172	0.092331	NA	4.64E-05	0.020513	0.007116	0.015387
13	rs9584155	86460262	0.005051	NA	4.69E-05	0.762073	0.000263	0.0156
9	rs10760693	101426527	0.20376	NA	4.73E-05	0.040601	0.001782	0.008499
12	rs17094348	42931797	0.006054	TMEM117	4.79E-05	0.125246	0.00641	0.013973
8	rs35292150	7141446	0.030287	FAM90A20	4.85E-05	0.317925	0.000694	0.002449
13	rs17348373	84641135	0.108476	NA	4.87E-05	0.000538	0.000173	0.002494
8	rs16880474	111237809	0.018668	NA	4.87E-05	0.0003	0.000613	0.003227
1	rs1366990	234948326	0.12109	ACTN2	4.95E-05	9.90E-05	0.000567	0.001519

APPENDIX B

SUPPLEMENTARY OF INTEGRATING MEAN AND HIGH-ORDER HETEROGENEITIES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES

B.1 Proof of Proposition about the null independence between the mean and variance heterogeneity tests under normality setting

At gene i, let the two samples of sizes (n_1, n_2) follow an identical normal distribution, $N(\mu, \sigma^2)$. Namely, H_{03} ($\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma^2$ and $\mu_{i1} = \mu_{i2} = \mu$) is true. Then,

$$Q_1 \stackrel{\text{def}}{=} \frac{(n_1 - 1)\hat{\sigma}_{i1}^2}{\sigma^2} \sim \chi_{n_1-1}^2, \qquad \textbf{Eq. B-1}$$

$$Q_2 \stackrel{\text{def}}{=} \frac{(n_1 - 1)\hat{\sigma}_{i2}^2}{\sigma^2} \sim \chi_{n_2-1}^2, \qquad \textbf{Eq. B-2}$$

$$Z \stackrel{\text{def}}{=} \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1), \qquad \textbf{Eq. B-3}$$

and Q_1, Q_2 and Z are independently distributed. The classical two-sample *F* statistic can be rewritten as

$$\hat{F} = \frac{\hat{\sigma}_{i1}^2}{\hat{\sigma}_{i2}^2} = \frac{Q_1/(n_1 - 1)}{Q_2/(n_2 - 1)} \sim F_{n_1-1, n_2-1}. \qquad \textbf{Eq. B-4}$$

The classical two-sample Student *t* statistic can be rewritten as

$$\hat{t} = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-\frac{1}{2}} (\hat{\mu}_{i1} - \hat{\mu}_{i2})}{\sqrt{\frac{n_1 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i1}^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i2}^2}}$$

$$= \frac{Z}{\sqrt{(Q_1 + Q_2)/(n_1 + n_2 - 2)}} \qquad \textbf{Eq. B-5}$$

and the two-sample Welch *t* statistic can be rewritten as

$$\hat{t}_w = \frac{(\hat{\mu}_{i1} - \hat{\mu}_{i2})}{\sqrt{\frac{1}{n_1}\hat{\sigma}_{i1}^2 + \frac{1}{n_2}\hat{\sigma}_{i2}^2}}$$

$$= \frac{Z}{\sqrt{n_2 Q_1/(n_1 + n_2)(n_1 - 1) + n_1 Q_2/(n_1 + n_2)(n_2 - 1)}}, \qquad \textbf{Eq. B-6}$$

let $\Gamma(\cdot)$ and $\mathrm{Beta}(\cdot,\cdot)$ denote Γ and Beta functions, respectively. By their mutual independence, Q_1, Q_2 and Z have joint probability density function

$$p_{(Q_1,Q_2,Z)}(q_1, q_2, z)$$

$$= \frac{q_1^{\frac{n_1-1}{2}-1} q_2^{\frac{n_2-1}{2}-1}}{\mathrm{Beta}\left(\frac{n_1 - 1}{2}, \frac{n_2 - 1}{2}\right) \mathrm{Beta}\left(\frac{1}{2}, \frac{n_1 + n_2}{2} - 1\right)}$$

$$\times \frac{\exp\left(-\frac{q_1 + q_2}{2}\right) \exp\left(-\frac{z^2}{2}\right)}{2^{\frac{n_1+n_2-1}{2}} \Gamma\left(\frac{n_1 + n_2 - 1}{2}\right)}. \qquad \textbf{Eq. B-7}$$

By the density formula of a multivariate transformation, the joint probability density function of $\left(\hat{t}, \hat{F}, Z\right)$ is given by

$$p_{(\hat{t},\hat{F},Z)}(t,f,z) = p_{(Q_1,Q_2,Z)}(q_1,q_2,z) \times |J|, \quad \textbf{Eq. B-8}$$

where

$$\begin{cases} q_1 = \dfrac{(n_1+n_2-2)z^2}{t^2}\dfrac{(n_1-1)f}{(n_1-1)f+n_2-1}, \\ q_2 = \dfrac{(n_1+n_2-2)z^2}{t^2}\dfrac{n_2-1}{(n_1-1)f+n_2-1}, \\ z = z, \end{cases} \quad \textbf{Eq. B-9}$$

and

$$|J| = \left\| \frac{\partial(q_1,q_2,z)'}{\partial(t,f,z)} \right\|$$

$$= \frac{2(n_1-1)(n_2-1)}{((n_1-1)f+(n_2-1))^2}\frac{(n_1+n_2-2)^2 z^4}{|t|^5} \quad \textbf{Eq. B-10}$$

is the absolute Jacobian determinant of the multivariate transformation (**Eq. B-9**).

The support of the joint density of $\hat{t}$, $\hat{F}$ and Z are defined into two sets $\{(t,f,z)|t>0,f>0,z>0\}$ and $\{(t,f,z)|t<0,f>0,z<0\}$. Substituting **Eq. B-9** into **Eq. B-10**, we obtain the joint density of $(\hat{t},\hat{F})$ by integrating variable Z

$$p_{(\hat{t},\hat{F})}(t,f) = \int_{-\infty}^{\infty} p_{(\hat{t},\hat{F},Z)}(t,f,z)\,\mathrm{dz}$$

$$= \int_{-\infty}^{\infty} p_{(Q_1,Q_2,Z)}(q_1,q_2,z) \times |J|\,\mathrm{dz}$$

$$= \frac{\frac{1}{f}\left(\frac{\left((n_1-1)f\right)^{(n_1-1)}(n_2-1)^{n_2-1}}{\left((n_1-1)f+n_2-1\right)^2}\right)^{\frac{1}{2}}}{\mathrm{Beta}\left(\frac{1}{2},\frac{n_1+n_2}{2}-1\right)\mathrm{Beta}\left(\frac{n_1-1}{2},\frac{n_2-1}{2}\right)}$$

$$\times \int_0^{\infty} \frac{\left(\frac{(n_1+n_2-2)z^2}{t^2}\right)^{\frac{n_1+n_2}{2}-1}\exp\left\{-\frac{z^2}{2}\left(1+\frac{n_1+n_2-2}{t^2}\right)\right\}}{2^{\frac{n_1+n_2-3}{2}}\Gamma\left(\frac{n_1+n_2-1}{2}\right)|t|}\mathrm{dz}$$

$$= p_{\hat{F}}(f) \times p_{\hat{t}}(t), \quad \textbf{Eq. B-11}$$

where

$$p_{\hat{F}}(f)=\frac{\frac{1}{f}\left(\frac{((n_1-1)f)^{(n_1-1)}(n_2-1)^{(n_2-1)}}{((n_1-1)f+n_2-1)^2}\right)^{\frac{1}{2}}}{\mathrm{Beta}\left(\frac{n_1-1}{2},\frac{n_2-1}{2}\right)}, \qquad \textbf{Eq. B-12}$$

is the probability density function of the F statistic, and

$$p_{\hat{t}}(t)=\frac{\left(1+\frac{t^2}{n_1+n_2-2}\right)^{-\frac{n_1+n_2-1}{2}}}{\sqrt{n_1+n_2-2}\mathrm{Beta}\left(\frac{1}{2},\frac{n_1+n_2}{2}-1\right)}, \qquad \textbf{Eq. B-13}$$

is the probability density function of the Student t statistic. In summary, if H_{03} holds, then $\hat{F}$ and $\hat{t}$ are independently distributed. Under the normality setting, the null independence of Welch t statistic to F statistics can be similarly proved. Specifically, we only need to consider transformation system

$$\begin{cases} q_1=\frac{(n_1+n_2)z^2}{t_w^2}\frac{(n_1-1)f}{n_2f+n_1} \\ q_2=\frac{(n_1+n_2)z^2}{t_w^2}\frac{n_2-1}{n_2f+n_1} \\ z=z. \end{cases}. \qquad \textbf{Eq. B-14}$$

Substituting **Eq. B-14** into **Eq. B-8** and repeating the other steps can prove the null independence between F statistic and Welch t statistic.

B.2 Proof of Proposition about the null independence between mean and variance heterogeneity tests under generic spherically symmetric setting

Mean heterogeneity tests in two-sample comparisons can be equivalent to a simple linear regression model:

$$G_{ij} = \beta_{0i} + \beta_i K_j + e_{ij}, \qquad \textbf{Eq. B-15}$$

where G_{ij} is the expression level of the i^{th} gene of the j^{th} subject, $K_j = 1$ if the j^{th} subject belongs to Group 1, and $K_j = 0$ if otherwise, β_{0i} is intercept and β_i is the effect of group on gene expression levels, and e_{ij} is random error. According to ordinary least squares (OLS) method, we obtain $\hat{\beta}_i = \hat{\mu}_{i1} - \hat{\mu}_{i2}$ and $\hat{\beta}_{0i} = \hat{\mu}_{i2}$. The standard error of $\hat{\beta}_i$ is

$$\mathrm{SE}_{\widehat{\beta}_i} = \sqrt{\left(\frac{1}{n_1 + n_2 - 2}\right)\left(\frac{\sum_{j=1}^{n_1+n_2}\left(G_{ij} - \hat{G}_{ij}\right)^2}{\sum_{j=1}^{n_1+n_2}\left(K_j - \overline{K}_j\right)^2}\right)}$$

$$= \sqrt{\left(\frac{1}{n_1 + n_2 - 2}\right)\left(\frac{\sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}_{i1}\right)^2 + \sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}_{i1}\right)^2}{\sum_{j=1}^{n_1+n_2}\left(K_j - \overline{K}_j\right)^2}\right)}$$

$$= \sqrt{\left(\frac{1}{n_1 + n_2 - 2}\right)\left(\frac{(n_1 - 1)\hat{\sigma}_{i1}^2 + (n_2 - 1)\hat{\sigma}_{i2}^2}{\frac{n_1 n_2}{n_1 + n_2}}\right)}, \qquad \textbf{Eq. B-16}$$

where $\hat{G}_{ij} = \hat{\beta}_{0i} + \hat{\beta}_i K_j$. The statistic to test $\beta_i = 0$ in **Eq. B-15** can be written as

$$t_{regression} = \frac{\hat{\beta}_i}{\mathrm{SE}_{\widehat{\beta}_i}}$$

$$= \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-\frac{1}{2}}(\hat{\mu}_{i1} - \hat{\mu}_{i2})}{\sqrt{\frac{n_1 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i1}^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}\hat{\sigma}_{i2}^2}} \qquad \textbf{Eq. B-17}$$

Thus, it is mathematically the Student *t* statistic in two-sample group comparisons. Under spherically symmetric distribution conditions, the density of gene expression levels is

$$L_G = \prod_{j=1}^{(n_1+n_2)} \frac{1}{\sigma} g\left(\frac{\left(G_{ij} - E\left(G_{ij}\right)\right)^2}{\sigma^2}\right)$$

$$= \frac{1}{\sigma^{n_1+n_2}} g\left(\sum_{j=1}^{(n_1+n_2)} \frac{\left(G_{ij} - E\left(G_{ij}\right)\right)^2}{\sigma^2}\right), \qquad \textbf{Eq. B-18}$$

where $g(.)$ is a given monotone function called the generating function with respect to the Lebesgue measure in $\mathbb{R}$, $E\left(G_{ij}\right) = \beta_{0i} + \beta_i K_j$ is the conditional expectation given K_j. Similar to the theorem for exponential family in Lehmann's book [44], the complete sufficient statistic for gene expression distribution is $\boldsymbol{T} = \left(\sum_{j=1}^{n_1} G_{ij1}^2 + \sum_{j=1}^{n_2} G_{ij2}^2, \sum_{j=1}^{n_1} G_{ij1} + \sum_{j=1}^{n_2} G_{ij2}, \sum_{j=1}^{n_1} G_{ij1}\right)$. Note that the *t* statistic of mean heterogeneity test is a function of $\boldsymbol{T}$. In addition, the *LF* statistic of Levene's test approximately follows *F* distribution with 1 and $(\mathrm{n}_1 + n_2 - 2)$ degree of freedoms. And this F distribution does not depend on parameters $\beta_{0i}, \beta_i, \sigma^2$ in (*B*1). Therefore, according to Basu's theorem [122], the *LF* and the Student t statistics are independently distributed ($\widehat{LF} \perp\!\!\!\perp \hat{t}$). Within the family of spherically symmetric distributions, mean and mode is the same and thus the BF statistic is also independent of Student t statistic ($\widehat{BF} \perp\!\!\!\perp \hat{t}$). $\widehat{LF} \perp\!\!\!\perp \hat{t}_w$ and $\widehat{BF} \perp\!\!\!\perp \hat{t}_w$ can be similarly proved. Since spherically symmetric distribution family is a very broad distribution family that include spherical exponential family, Student distribution, Laplace distribution, exponential power distribution and many other distributions, the Student and Welch t-statistics are independent of the Levene and Brown-Forsythe statistics under normality settings by letting random error *e* follow normal distribution in **Eq. B-15**.

B.3 Two-sample likelihood ratio test

Herein, we derive the formula of the two-sample likelihood ratio test under the joint null hypothesis. For the i^{th} gene, let $\boldsymbol{G_{i1}} = \left(G_{i11}, G_{i21}, \dots, G_{in_1 1}\right)'$ and $\boldsymbol{G_{i2}} = \left(G_{i12}, G_{i22}, \dots, G_{in_2 2}\right)'$ be expression levels of two independent random samples from normal populations $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$ and $\mathcal{N}(\mu_{i2}, \sigma_{i2}^2)$, respectively. The full likelihood function is given by

$$L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2) = \left(\frac{1}{2\pi\sigma_{i1}^2}\right)^{\frac{n_1}{2}} \left(\frac{1}{2\pi\sigma_{i2}^2}\right)^{\frac{n_2}{2}} \exp\left(-\frac{1}{2}\left[\sum_{j=1}^{n_1}\left(\frac{G_{ij1}-\mu_{i1}}{\sigma_{i1}}\right)^2 + \sum_{j=1}^{n_2}\left(\frac{G_{ij2}-\mu_{i2}}{\sigma_{i2}}\right)^2\right]\right). \quad \textbf{Eq. B-19}$$

Under the joint null hypothesis, $\mu_{i1} = \mu_{i2} = \mu_i$ and $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_i^2$, the reduced likelihood (joint function) can be rewritten as

$$L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2) = L(\mu_i, \mu_i, \sigma_i^2, \sigma_i^2) = \left(\frac{1}{2\pi\sigma_i^2}\right)^{\frac{n_1+n_2}{2}} \exp\left(-\frac{1}{2\sigma_i^2}\left[\sum_{j=1}^{n_1}\left(G_{ij1}-\mu_i\right)^2 + \sum_{j=1}^{n_2}\left(G_{ij2}-\mu_i\right)^2\right]\right). \quad \textbf{Eq. B-20}$$

Solving the system of equations of $\frac{\partial \ln L(\mu_i, \mu_i, \sigma_i^2, \sigma_i^2)}{\partial \mu_i} = 0$ and $\frac{\partial \ln L(\mu_i, \mu_i, \sigma_i^2, \sigma_i^2)}{\partial \sigma_i^2} = 0$, we derive the maximum likelihood estimators

$$\hat{\mu}_i = \frac{1}{n_1 + n_2}\left(\sum_{j=1}^{n_1} G_{ij1} + \sum_{j=1}^{n_2} G_{ij2}\right)$$

and

$$\hat{\sigma}_i^2 = \frac{1}{n_1 + n_2}\left(\sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}_i\right)^2 + \sum_{j=1}^{n_2}\left(G_{ij2} - \hat{\mu}_i\right)^2\right).$$

The maximum of the reduced likelihood $L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)$ under the joint null hypothesis is

$$\max_{\mu_i, \sigma_i^2} L(\mu_i, \mu_i, \sigma_i^2, \sigma_i^2) = L(\hat{\mu}_i, \hat{\mu}_i, \hat{\sigma}_i^2, \hat{\sigma}_i^2)$$

$$= \frac{1}{\sigma^{n_1+n_2}} g\left(\sum_{j=1}^{(n_1+n_2)} \frac{\left(G_{ij} - E\left(G_{ij}\right)\right)^2}{\sigma^2}\right)$$

$$= \left(\frac{1}{2\pi\hat{\sigma}_i^2}\right)^{\frac{n_1+n_2}{2}} \exp\left(-\frac{n_1 + n_2}{2}\right), \qquad \textbf{Eq. B-21}$$

using the full likelihood of $(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)$, we obtain the following system of equations:

$$\begin{cases} \dfrac{\partial \ln L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)}{\partial \mu_{i1}} = 0, \\ \dfrac{\partial \ln L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)}{\partial \mu_{i2}} = 0, \\ \dfrac{\partial \ln L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)}{\partial \sigma_{i1}^2} = 0, \\ \dfrac{\partial \ln L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)}{\partial \sigma_{i2}^2} = 0. \end{cases}$$

Solving the system, we derive

$$\hat{\mu}_{i1} = \frac{1}{n_1}\sum_{j=1}^{n_1} G_{ij1},$$

$$\hat{\mu}_{i2} = \frac{1}{n_2}\sum_{j=1}^{n_2} G_{ij2},$$

$$\hat{\sigma}_{i1}^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}_{i1}\right)^2,$$

and

$$\hat{\sigma}_{i2}^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\left(G_{ij2} - \hat{\mu}_{i2}\right)^2.$$

Then maximum of $L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)$ over the full parameter space can be derived as below:

$$\begin{aligned}
&\max_{\mu_{i1},\mu_{i2},\sigma_{i1}^2,\sigma_{i2}^2} L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2) \\
&= L(\hat{\mu}_{i1}, \hat{\mu}_{i2}, \hat{\sigma}_{i1}^2, \hat{\sigma}_{i2}^2) \\
&= \left(\frac{1}{2\pi\hat{\sigma}_{i1}^2}\right)^{\frac{n_1}{2}} \left(\frac{1}{2\pi\hat{\sigma}_{i2}^2}\right)^{\frac{n_2}{2}} \exp\left(-\frac{1}{2}\left[\sum_{j=1}^{n_1}\left(\frac{G_{ij1} - \hat{\mu}_{i1}}{\hat{\sigma}_{i1}}\right)^2 \right.\right. \\
&\qquad \left.\left. + \sum_{j=1}^{n_2}\left(\frac{G_{ij2} - \hat{\mu}_{i2}}{\hat{\sigma}_{i2}}\right)^2\right]\right) \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n_1+n_2}{2}} \exp\left(-\frac{n_1+n_2}{2}\right)\left(\frac{1}{\hat{\sigma}_{i1}^2}\right)^{\frac{n_1}{2}} \left(\frac{1}{\hat{\sigma}_{i2}^2}\right)^{\frac{n_2}{2}}. \qquad \textbf{Eq. B-22}
\end{aligned}$$

From **Eq. B-21** and **Eq. B-22**, we derive the likelihood ratio

$$\begin{aligned}
LR &= \frac{\max_{\mu_i,\sigma_i^2} L(\mu_i, \mu_i, \sigma_i^2, \sigma_i^2)}{\max_{\mu_{i1},\mu_{i2},\sigma_{i1}^2,\sigma_{i2}^2} L(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2)} = \frac{(\hat{\sigma}_{i1}^2)^{\frac{n_1}{2}}(\hat{\sigma}_{i2}^2)^{\frac{n_2}{2}}}{(\hat{\sigma}^2)^{\frac{n_1+n_2}{2}}} \\
&= \frac{\left(\frac{1}{\mathrm{n}_1}\sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}_{i1}\right)^2\right)^{\frac{n_1}{2}} \left(\frac{1}{\mathrm{n}_2}\sum_{j=1}^{n_2}\left(G_{ij2} - \hat{\mu}_{i2}\right)^2\right)^{\frac{n_2}{2}}}{\left(\frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1}\left(G_{ij1} - \hat{\mu}\right)^2 + \sum_{j=1}^{n_2}\left(G_{ij2} - \hat{\mu}\right)^2\right)\right)^{\frac{n_1+n_2}{2}}}.
\end{aligned}$$

For large samples, the statistic $-2\ln(LR)$ of likelihood ratio test follows asymptotically chi-square distribution with $\mathrm{df} = 2$ under H_{03}. The finite-sample performance of the LRT depends on the sample size, and χ_2^2 distribution may not well approximate the exact

distribution of $-2\ln(LR)$ for a small sample, which is intractable even under normality setting.

B.4 Proof about the asymptotical null independence between $\mathbf{T_{w_1}}$ and $\mathbf{T_{w_2}}$

Let the dataset contain expression levels of M gene probes of n_1 unrelated subjects from control groups and n_2 unrelated subjects for treatment group, respectively. For a specific gene, let $X_1 = \left(X_{11}, X_{12}, \dots, X_{1n_1}\right)$ be the expression level of gene probes under control group and $X_2 = (X_{21}, X_{22}, \dots, X_{2n_2})$ be the expression level of gene probes under treatment group. The total sample size is $n = n_1 + n_2$ be. Let $\mu_1(X_c)$ and $\sigma^2_{X_c}$ be the gene-specific mean and variance of the expression levels of gene probe under condition c (i.e., $c = 1$ for control group, and $c = 2$ for treatment group). And let $\mu_2(X_c) = E(X_c^2)$ be the second-order moment of X_c. According to the definition of second order moment, $\mu_2(X_c) = \left(\mu_1(X_c)\right)^2 + \sigma^2_{X_c}$.

Without loss of generality, we assume $\overline{X_1} < \overline{X_2}$. Define $Y_{1i} = X_{1i} - \overline{X_1}$ and $Y_{2j} = X_{2j} - \overline{X_1}$, where $i = 1,2, \dots, n_1, j = 1,2, \dots, n_2, \overline{X_1} = \sum_{i=1}^{n_1} X_{1i}$ and $\overline{X_2} = \sum_{j=1}^{n_2} X_{2j}$. Firstly, we constructed the first welch *t* test statistic to capture the mean heterogeneity between two groups.

Now we have

$$T_{w_1} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S^2_{Y_1}}{n_1} + \frac{S^2_{Y_2}}{n_2}}}, \qquad \textbf{Eq. B-23}$$

and

$$T_{w_2} = \frac{\overline{Y_1^2} - \overline{Y_2^2}}{\sqrt{\frac{S_{Y_1}^2}{n_1} + \frac{S_{Y_2}^2}{n_2}}}. \qquad \textbf{Eq. B-24}$$

,

The null hypothesis is $\mu_1(X_1) = \mu_1(X_2)$ and $\mu_2(X_1) = \mu_2(X_2)$. Next we prove the following conclusion: When $n_1 = n_2$ and $n_1, n_2 \longrightarrow \infty$, T_{w_1} and T_{w_2} would converge in distribution to a bivariate normal distribution with unit variance and zero correlation coefficient below

$$\begin{pmatrix} T_{w1} \\ T_{w2} \end{pmatrix} \xrightarrow{a.d.} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \qquad \textbf{Eq. B-25}$$

Proof:

Under $n_1, n_2 \longrightarrow \infty$, to simplify the proof, we redefine

$$Y_{1i} = \sqrt{\frac{n_1 + 1}{n_1 - 1}}(X_{1i} - \overline{X_1}) \approx (X_{1i} - \overline{X_1}), \qquad \textbf{Eq. B-26}$$

Eq. B-23 can be written as

$$T_{w_1} - \frac{\sqrt{n_1}(\mu_1(X_1) - \mu_1(X_2))}{\sqrt{S_{X_1}^2 + \frac{n_1}{n_2}S_{X_2}^2}} = \frac{\begin{bmatrix} 1 & 0 & -\sqrt{\frac{n_1}{n_2}} & 0 \end{bmatrix}}{\sqrt{S_{X_1}^2 + \frac{n_1}{n_2}S_{X_2}^2}}$$

$$\times \begin{pmatrix} \sqrt{n_1}(\overline{X_1} - \mu_1(X_1)) \\ \sqrt{n_1}(\overline{X_1^2} - \mu_2(X_1)) \\ \sqrt{n_2}(\overline{X_2} - \mu_1(X_2)) \\ \sqrt{n_2}(\overline{X_2^2} - \mu_2(X_2)) \end{pmatrix}, \qquad \textbf{Eq. B-27}$$

Similar as T_{w_1}, T_{w_1} can be written as

$$T_{w_2} - \frac{\sqrt{n_1}\left[\sigma_{X_1}^2 - \sigma_{X_2}^2 - (\mu_1 - \mu_2)^2\right]}{\sqrt{S_{Y_1^2}^2 + \frac{n_1}{n_2} S_{Y_2^2}^2}}$$

$$= \frac{\sqrt{n_1}\left(\overline{Y_1^2} - \overline{Y_2^2}\right) - \left(\sigma_{X_1}^2 - \sigma_{X_2}^2 - (\mu_1 - \mu_2)^2\right)}{\sqrt{S_{Y_1^2}^2 + \frac{n_1}{n_2} S_{Y_2^2}^2}}$$

$$= \frac{1}{\sqrt{S_{Y_1^2}^2 + \frac{n_1}{n_2} S_{Y_2^2}^2}} \left[-2 \left(\mu_1(X_2) \right.\right.$$

$$\left.\left. + \frac{\mu_1(X_1) + \overline{X_1}}{n-1} \right), \frac{n_1 + 1}{n_1 - 1}, -2\sqrt{\frac{n_1}{n_2}}\mu_1(X_1), \sqrt{\frac{n_1}{n_2}} \right]$$

$$\times \begin{pmatrix} \sqrt{n_1}\left(\overline{X_1} - \mu_1(X_1)\right) \\ \sqrt{n_1}\left(\overline{X_1^2} - \mu_2(X_1)\right) \\ \sqrt{n_2}\left(\overline{X_2} - \mu_1(X_2)\right) \\ \sqrt{n_2}\left(\overline{X_2^2} - \mu_2(X_2)\right) \end{pmatrix}$$

$$+ \frac{1}{\sqrt{S_{Y_1^2}^2 + \frac{n_1}{n_2} S_{Y_2^2}^2}} \left(\frac{2}{n_1 - 1} \frac{\sigma_{X_1}^2}{\sqrt{n_1}} \right.$$

$$\left. - 2\sqrt{n_1}\left(\overline{X_1} - \mu_1(X_1)\right)\left(\overline{X_2} - \mu_1(X_2)\right) \right). \qquad \textbf{Eq. B-28}$$

Letting $\frac{n_1}{n_2} = r$

$$\begin{pmatrix} T_{w_1} - C_{w_1} \\ T_{w_2} - C_{w_2} \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{[1 \quad 0 \quad -\sqrt{r} \quad 0]}{\sqrt{S_{X_1}^2 + rS_{X_2}^2}} \\ \dfrac{\left[-2\left(\mu_1(X_2) + \dfrac{\mu_1(X_1) + \overline{X_1}}{n-1}\right) \quad \dfrac{n_1+1}{n_1-1} \quad -2\sqrt{r}\mu_1(X_1) \quad \sqrt{r}\right]}{\sqrt{S_{Y_1^2}^2 + rS_{Y_2^2}^2}} \end{pmatrix}$$

$$\begin{pmatrix} \sqrt{n_1}\left(\overline{X_1} - \mu_1(X_1)\right) \\ \sqrt{n_1}\left(\overline{X_1^2} - \mu_2(X_1)\right) \\ \sqrt{n_2}\left(\overline{X_2} - \mu_1(X_2)\right) \\ \sqrt{n_2}\left(\overline{X_2^2} - \mu_2(X_2)\right) \end{pmatrix}$$

$$+ \begin{pmatrix} 0 \\ \dfrac{1}{\sqrt{S_{Y_1^2}^2 + \dfrac{n_1}{n_2}S_{Y_2^2}^2}}\left(\dfrac{2}{n_1 - 1}\dfrac{\sigma_{X_1}^2}{\sqrt{n_1}} - 2\sqrt{n_1}\left(\overline{X_1} - \mu_1(X_1)\right)\left(\overline{X_2} - \mu_1(X_2)\right)\right) \end{pmatrix} \quad \textbf{Eq. B-29}$$

where

$$C_{w_1} = \frac{\sqrt{n_1}(\mu_1(X_1) - \mu_1(X_2))}{\sqrt{S_{X_1}^2 + \dfrac{n_1}{n_2}S_{X_2}^2}}, C_{w_2} = \frac{\sqrt{n_1}\left[\sigma_{X_1}^2 - \sigma_{X_2}^2 - (\mu_1(X_1) - \mu_1(X_2))^2\right]}{\sqrt{S_{Y_1^2}^2 + \dfrac{n_1}{n_2}S_{Y_2^2}^2}}$$

As $n_1, n_2 \longrightarrow \infty$

$$\begin{pmatrix} T_{w_1} - C_{w_1} \\ T_{w_2} - C_{w_2} \end{pmatrix}$$

$$\xrightarrow{a.d.} \begin{pmatrix} \frac{[1 \quad 0 \quad -\sqrt{r} \quad 0]}{\sqrt{\sigma_{X_1}^2 + r\sigma_{X_2}^2}} \\ \frac{\left[-2\left(\mu_1(X_2)\right) \quad 1 \quad -2\sqrt{r}\mu_1(X_1) \quad \sqrt{r}\right]}{\sqrt{\sigma_{Y_1^2}^2 + r\sigma_{Y_2^2}^2}} \end{pmatrix} \times N_4(\mathbf{0}, \Omega)$$

$$= N_4(\mathbf{0}, A\Omega\mathrm{A}'), \qquad \textbf{Eq. B-30}$$

where

$$\sigma_{Y_1^2}^2 = \lim_{n_1, n_2 \to \infty} S_{Y_1^2}^2,$$

$$\sigma_{Y_2^2}^2 = \lim_{n_1, n_2 \to \infty} S_{Y_2^2}^2$$

$$A = \begin{pmatrix} \frac{[1 \quad 0 \quad -\sqrt{r} \quad 0]}{\sqrt{\sigma_{X_1}^2 + r\sigma_{X_2}^2}} \\ \frac{\left[-2\left(\mu_1(X_2)\right) \quad 1 \quad -2\sqrt{r}\mu_1(X_1) \quad \sqrt{r}\right]}{\sqrt{\sigma_{Y_1^2}^2 + r\sigma_{Y_2^2}^2}} \end{pmatrix}$$

$$\Omega = \begin{pmatrix} \sigma_{X_1}^2 & Cov(X_1, X_1^2) & 0 & 0 \\ Cov(X_1, X_1^2) & \sigma_{X_1^2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{X_2}^2 & Cov(X_2, X_2^2) \\ 0 & 0 & Cov(X_2, X_2^2) & \sigma_{X_2^2}^2 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 \\ \frac{1}{\sqrt{S_{Y_1^2}^2 + \frac{n_1}{n_2} S_{Y_2^2}^2}} \left(\frac{2}{n_1 - 1} \frac{\sigma_{X_1}^2}{\sqrt{n_1}} - 2\sqrt{n_1}\left(\overline{X_1} - \mu_1(X_1)\right)\left(\overline{X_2} - \mu_1(X_2)\right) \right) \end{pmatrix} \xrightarrow{a.d.} o(1)$$

Then we obtain

$$\begin{pmatrix} T_{w_1} - C_{w_1} \\ T_{w_2} - C_{w_2} \end{pmatrix} \xrightarrow{a.d.} N_4\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where

ρ

$$= \frac{\left(\mu_3(X_1) - \mu_1(X_1)\mu_2(X_1)\right) - 2\mu_1(X_2)\sigma^2_{X_1^2} - r\left(\mu_3(X_2) - \mu_1(X_2)\mu_2(X_2) - 2\mu_1(X_1)\sigma^2_{X_2^2}\right)}{\sqrt{\sigma^2_{X_1} + r\sigma^2_{X_2}}\sqrt{\sigma^2_{Y_1^2} + r\sigma^2_{Y_2^2}}}$$

Under the null hypothesis,

$$\rho = \frac{(1-r)(\mu_3(X_1) - \mu_1(X_1)\mu_2(X_1) - 2\mu_1(X_1)\sigma^2_{X_1})}{\sqrt{\sigma^2_{X_1} + r\sigma^2_{X_2}}\sqrt{\sigma^2_{Y_1^2} + r\sigma^2_{Y_2^2}}}$$

When $n_1 = n_2$, $r = 1$, then $\rho = 0$

B.5 Supplemental Figures

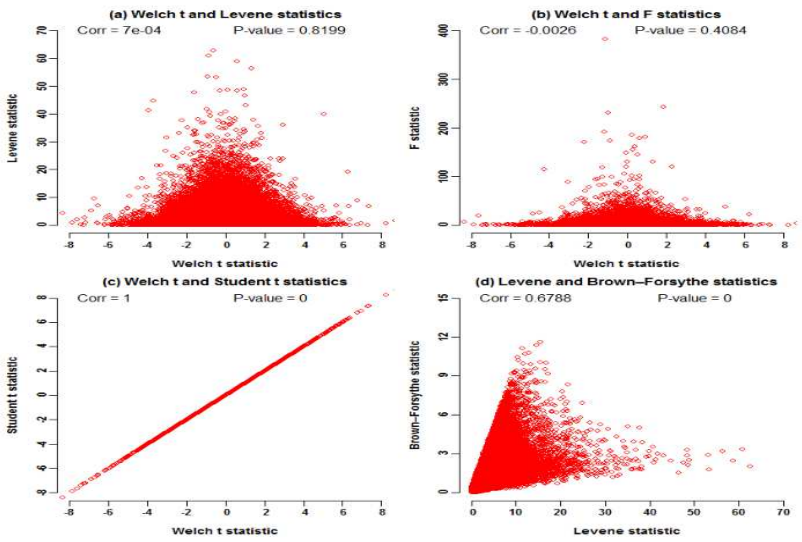

Figure B-1: Null joint distributions of mean and variance test statistics under 5 vs. 5 normality setting.

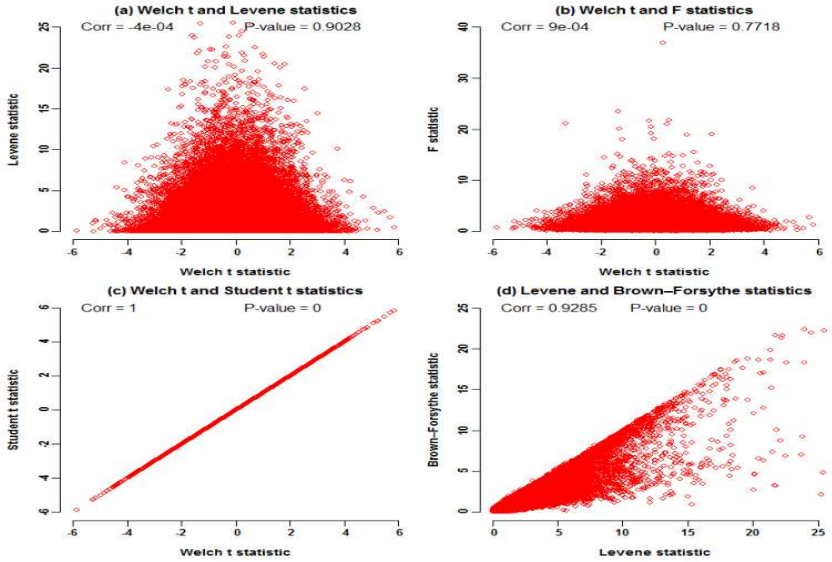

Figure B-2: Null joint distributions of mean and variance test statistics under 10 vs.10 normality setting.

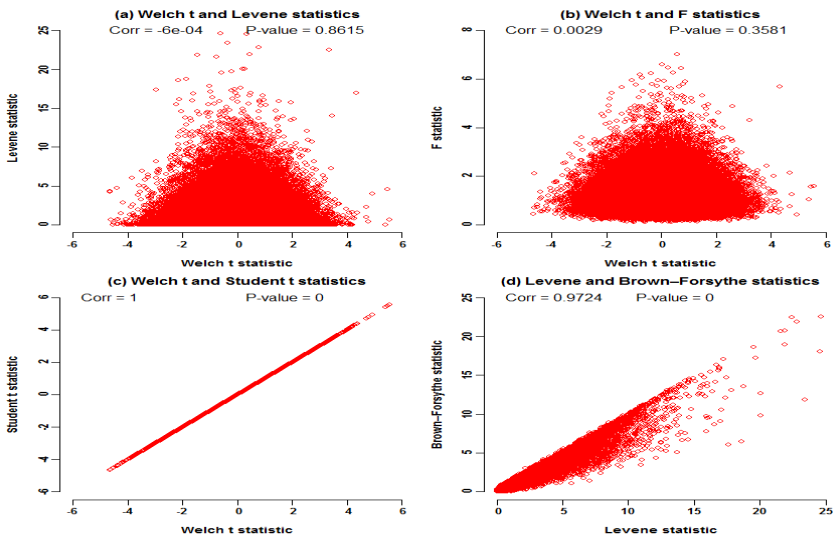

Figure B-3: Null joint distributions of mean and variance test statistics under 20 vs. 20 normality setting.

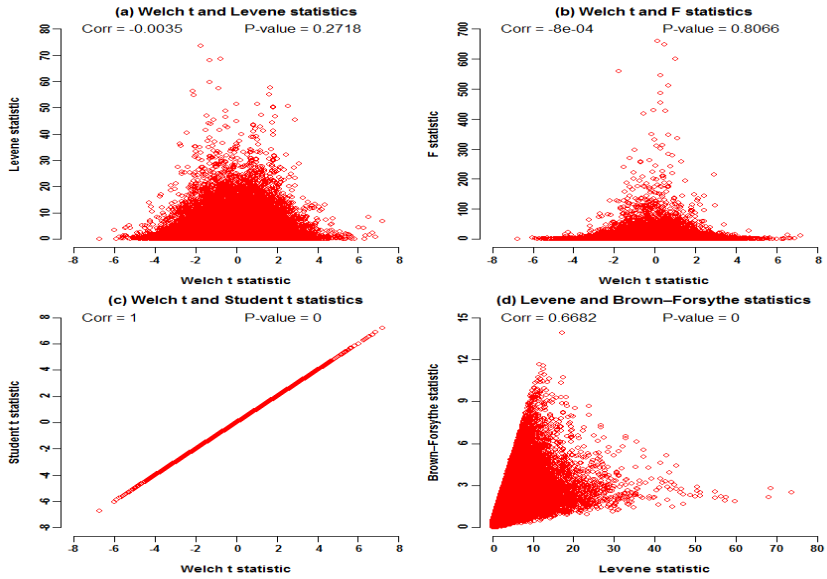

Figure B-4: Null joint distributions of mean and variance test statistics under 5 vs. 5 Laplace setting

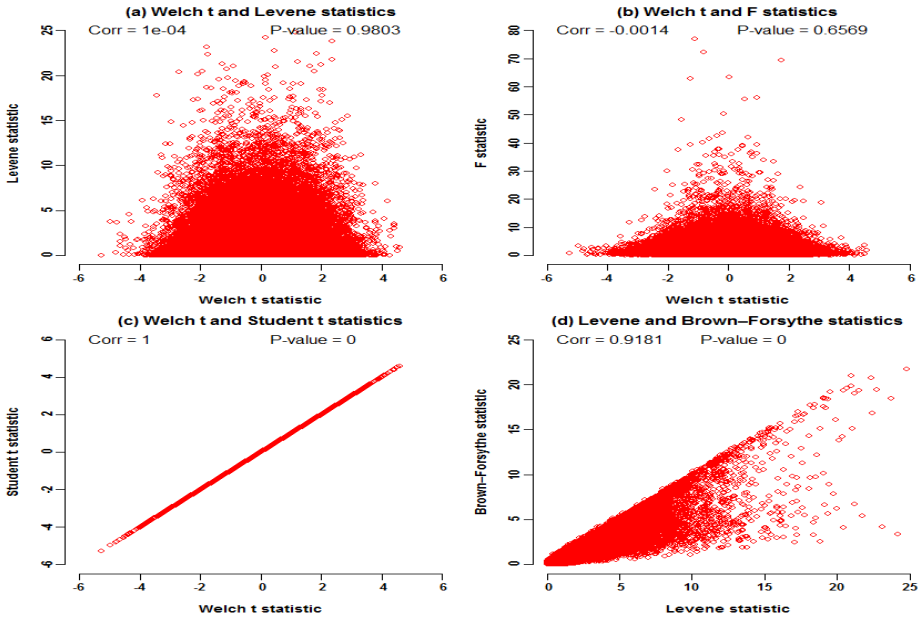

Figure B-5: Null joint distributions of mean and variance test statistics under 10 vs. 10 Laplace setting.

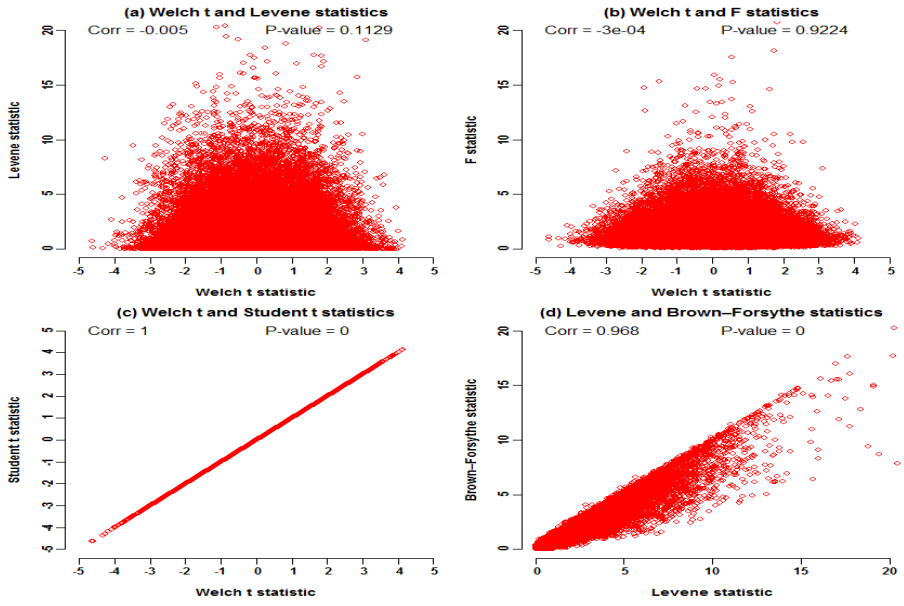

Figure B-6: Null joint distributions of mean and variance test statistics under 20 vs. 20 Laplace setting.

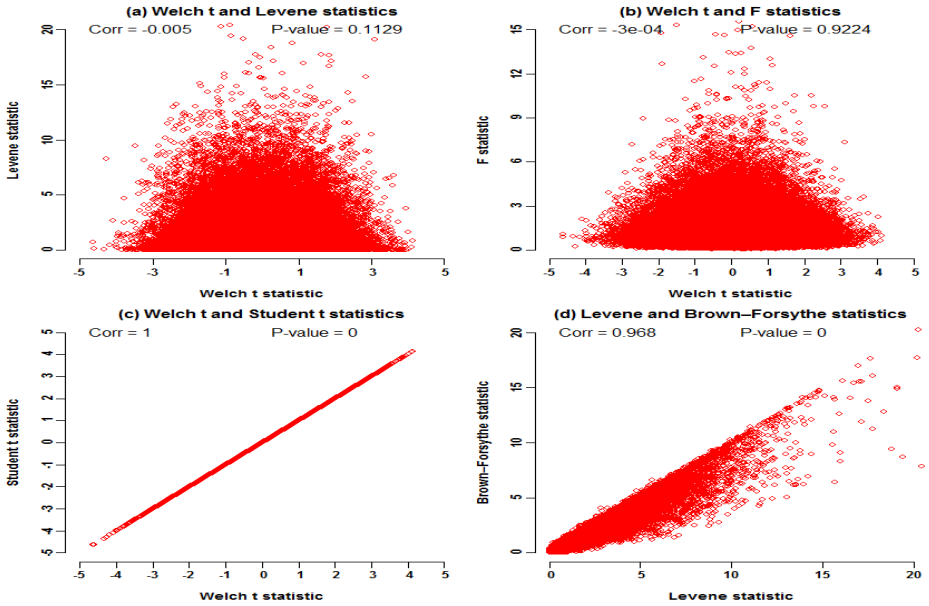

Figure B-7: Null joint distributions of mean and variance test statistics under 40 vs. 40 Laplace setting.

B.6 Supplemental Tables

Table B-1: The first 2 and significant PCs of all the experiment-wide gene probes

PC Index	Eigen Value	Variation Proportion	*P* values of			
			ST	WT	MWT	Levene
1	1.48E+13	**0.9824**	0.5343	0.5355	0.8459	0.2359
2	4.89E+10	**0.0032**	0.5786	0.5782	0.5752	0.9243
4	2.01E+10	0.0013	**1.84E-15**	**1.91E-15**	**4.49E-15**	0.0935
6	1.08E+10	0.0007	0.1148	0.1134	0.1102	**0.0003**
9	7.15E+09	0.0005	0.9725	0.9723	0.972	**0.0015**
14	4.32E+09	0.0003	0.5616	0.5598	0.557	**0.0114**
28	1.88E+09	0.0001	0.5533	0.5519	0.549	**0.0185**
38	1.33E+09	0.0001	0.9091	0.9095	0.9084	**0.0419**
49	9.81E+08	0.0001	0.3012	0.2993	0.2958	**0.0142**
78	4.47E+08	2.97E-05	0.8199	0.8217	0.8189	**0.0129**

Table B-2: The first 2 and significant PCs of 13415 experiment-wide robust gene probes

PC Index	Eigen Value	variation Proportion	*P* values of			
			ST	WT	MWT	Levene
1	9.06E+12	**0.9835**	0.6702	0.6712	0.8922	0.1865
2	3.40E+10	**0.0037**	0.6448	0.6442	0.6419	0.6613
12	2.89E+09	0.0003	**0.0064**	**0.0063**	**0.006**	0.0557
14	2.36E+09	0.0003	**0.0035**	**0.0036**	**0.0033**	0.538
16	2.03E+09	0.0002	**0.0135**	**0.0134**	**0.0129**	0.3799
18	1.67E+09	0.0002	**0.0151**	**0.0149**	**0.0144**	0.1453
25	1.17E+09	0.0001	**0.0059**	**0.0058**	**0.0056**	0.9441
28	1.01E+09	0.0001	0.2272	0.2253	0.222	**0.0069**
29	9.91E+08	0.0001	0.6090	0.6074	0.605	**0.0119**
30	9.78E+08	0.0001	0.9674	0.9675	0.9672	**0.0208**

Table B-3: Discoveries of the IMVT by controlling FDR below 0.1

AffyID	Gene	Local FDR			
		IMVT	**STSD**	**MWT**	**WT**
203558_at	*CUL7*	0.0025	0.0344	0.4676	0.4701
208307_at	*RBMY1J*	0.0043	0.3490	0.6452	0.6345
204384_at	*GOLGA2*	0.0116	0.2158	0.4607	0.4701
206359_at	*SOCS3*	0.0116	0.3158	0.4766	0.4915
208294_x_at	*CSHL1*	0.0116	0.2155	0.4445	0.4701
210106_at	*RDH5*	0.0116	0.3568	0.6046	0.6039
214436_at	*FBXL2*	0.0116	0.7192	0.8635	0.8527
218922_s_at	*CERS4*	0.0116	0.3158	0.4761	0.4915
214886_s_at	*N4BP2L1*	0.0137	0.3231	0.7155	0.6722
210492_at	*MFAP3L*	0.0160	0.3306	0.6466	0.6267
206162_x_at	*SYT5*	0.0162	0.4133	0.6414	0.6317
215840_at	*DNAH2*	0.0176	0.2158	0.5222	0.5227
214257_s_at	*SEC22B*	0.0240	0.2158	0.4445	0.4701
219829_at	*ITGB1BP2*	0.0249	0.3953	0.6717	0.6557
211789_s_at	*MLXIP*	0.0305	0.3158	0.6595	0.6430
209461_x_at	*WDR18*	0.0359	0.6251	0.7799	0.7725
210974_s_at	*AP3D1*	0.0373	0.2457	0.4445	0.4701
214145_s_at	*SPTB*	0.0373	0.3231	0.4761	0.4915
210922_at	*BC000772*	0.0397	0.5364	0.7180	0.7128
220625_s_at	*ELF5*	0.0404	0.3134	0.4676	0.4788
222260_at	*AK026947*	0.0404	0.3306	0.5112	0.5224
203532_x_at	*CUL5*	0.0441	0.3810	0.6342	0.6281
214138_at	*ZNF79*	0.0460	0.5952	0.7812	0.7772
221208_s_at	*MSANTD2*	0.0460	0.1610	0.2808	0.3123
203609_s_at	*ALDH5A1*	0.0514	0.3271	0.4766	0.4915
222256_s_at	*JMJD7*	0.0517	0.3402	0.5112	0.5179
204947_at	*E2F1*	0.0544	0.2158	0.4766	0.4915
214803_at	*CDH6*	0.0643	0.3450	0.6213	0.6184
221528_s_at	*ELMO2*	0.0643	0.4408	0.7337	0.7364
218659_at	*ASXL2*	0.0710	0.3134	0.4761	0.4915
209666_s_at	*CHUK*	0.0783	0.3158	0.4676	0.4788
203918_at	*PCDH1*	0.0808	0.3262	0.5112	0.5224
208524_at	*GPR15*	0.0808	0.1610	0.2748	0.3123
209850_s_at	*CDC42EP2*	0.0816	0.3297	0.5033	0.5175
204854_at	*LEPREL2*	0.0848	0.3946	0.6046	0.6031

206604_at	*OVOL1*	0.0848	0.3490	0.5143	0.5227
207961_x_at	*MYH11*	0.0848	0.4337	0.6289	0.6317
216975_x_at	*NPAS1*	0.0848	0.3231	0.5947	0.5859
222015_at	*CSNK1E*	0.0848	0.2005	0.4445	0.4566
200080_s_at	*H3F3AP4*	0.0886	0.1912	0.2808	0.3123
205391_x_at	*ANK1*	0.0886	0.3953	0.5947	0.5921
209156_s_at	*COL6A2*	0.0886	0.3810	0.6897	0.6713
210565_at	*GCGR*	0.0886	0.3231	0.5161	0.5227
216006_at	*AF070620*	0.0886	0.3231	0.4761	0.4915
216584_at	*216584_at*	0.0886	0.3134	0.4873	0.4915
219733_s_at	*SLC27A5*	0.0886	0.3158	0.4761	0.4915
205387_s_at	*CGB7*	0.0905	0.3490	0.5267	0.5314
222084_s_at	*SBF1*	0.0924	0.3158	0.4676	0.4788
206298_at	*ARHGAP22*	0.0955	0.3564	0.5847	0.5854
207150_at	*SLC18A3*	0.0969	0.3262	0.4761	0.4915
215786_at	*AK022170*	0.0969	0.2158	0.4094	0.4535
219729_at	*PRRX2*	0.0969	0.3490	0.5847	0.5793
220735_s_at	*SENP7*	0.0969	0.3231	0.4761	0.4915
216313_at	*PCDHB17*	0.0972	0.2242	0.4980	0.4933
212514_x_at	*DDX3X*	0.0990	0.1610	0.2595	0.2962

APPENDIX C DETAILED DISCRIPTIONS OF FIGURES

Figure 1-2: Comparison of false positive rates of eight methods under null hypothesis. The result was computed from 100000 replicates with the specified samples size 1000. At each significance level, the false positive rate of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$. The gray belt is the 95% concentration band of the false positive rates of a typical test that can properly control false positive rates at given nominal significance levels.

Figure 1-3: Power comparison of MT, JLS, LRT and HSAT under Scenario I at nominal level $\mathbf{5 \times 10^{-8}}$. Setting MAF to be 0.01, 0.025 and 0.05, powers of the four methods were computed from 100000 replicates with samples size 1000. X-axis is the heritability of genotype (h^2) that ranges from 0% to 2% for single locus and Y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at significance level 5×10^{-8}

Figure 1-4: Power comparison of MT, JLS, LRT and HSAT under Scenario II at nominal level $\mathbf{5 \times 10^{-8}}$. Setting the main genetic effect β to be 0.01, 0.05 and 0.1, powers of the four methods were computed from 100000 replicates with samples size 1000. The x-axis is the effect size of $G \times E$ interaction term that ranges from 0 to 1 by grid of 0.1 and y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at significance level 5×10^{-8}.

Figure 1-5: Power comparison of MT, JLS, LRT and HSAT under Scenario III at nominal level $\mathbf{5 \times 10^{-8}}$. (a) MAF of the common causal variant ranges from 0.05 to 0.5. Setting the main genetic effect β to be 0.25, the x-axis is the effect of genotype (γ) on variance that ranges from 0 to 0.5 by grid of 0.05;(b) MAF of the common causal variant ranges from 0.005 to 0.05. Setting the main genetic effect size β to be 0.5, the x-axis is the effect size of genotype (γ) on variance that ranges from 0 to 0.5 by grid of 0.05.The y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$. Powers of the four methods were computed from 100000 replicates with samples size 1000.

Figure 1-6: Q-Q plots of MT, JLS, LRT and HSAT. The inflation factors of MT, JLS and HSAT appeared reasonable and indicate no obvious inflation. While the curve of LRT clearly appeared under the gray band (95% concentration band), which indicates the conservativeness of LRT method.

Figure 1-7: The Manhattan plot of HSAT. 856149 SNPs with MAF>0.005 was plotted. Obvious association signal peaks were observed on chromosomes 2, 6, 7 and 19. The gray line is the suggestive nominal level 5×10^{-6}.

Figure 2-2: Comparison of false positive rates of eight methods under different nominal levels. The result was computed from 100000 replicates with the specified samples size 1000. At each significance level, the false positive rate of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$. The gray belt is the 95% concentration band of the false positive rates of a typical test that can properly control false positive rates at given nominal significance levels.

Figure 2-3: Comparison of empirical powers of eight methods at different nominal levels under HP model. Setting $m = 50$ and $l = 10$, the empirical powers of the eight methods were computed from 10000 replicates with samples size 1000. The percentage of positive causal variants (+) are set to be 100%, 80%, 50% and 20% respectively. The X-axis is the nominal level α that ranges from 0 to 0.05 and Y-axis is the empirical

powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at the nominal level α.

Figure 2-4: Comparison of false positive rates of eight methods at different nominal levels under Fisher's model framework. Setting $m = 50$ and $l = 10$, the empirical powers of the eight methods were computed from 10000 replicates with samples size 1000. The percentage of positive causal variants (+) are set to be 100%, 80%, 50% and 20% respectively. The X-axis is the nominal level α that ranges from 0 to 0.05 and Y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at the nominal level α.

Figure 2-5: Comparison of empirical power of eight methods levels when latent G×E interaction exists at nominal level 0.005(a) and 0.0005(b). Setting rs811589 in OPA3 as the causal loci and the main genetic effect β is 0.25, powers of the eight methods were computed from 10000 replicates with samples size 991. There are 23 test SNPs with MAF>0.005 in OPA3. The x-axis is the effect size δ of $G \times E$ interaction term that ranges from 0 to 0.5 by grid of 0.05 and y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at significance level 0.005 and 0.0005 respectively.

Figure 2-6: Comparison of empirical power of eight methods levels when latent G×G interaction exists at nominal level 0.005(a) and 0.0005(b). Setting rs811589 in

OPA3 as the causal loci and the main genetic effect β is 0.25, powers of the eight methods were computed from 10000 replicates with samples size 991. There are 23 test SNPs with MAF>0.005 in OPA3. Setting the main genetic effect β to be 0.25, the x-axis is the effect of genotype (γ) on variance that ranges from 0 to 0.25 by grid of 0.05. The y-axis is the empirical powers estimated by the empirical proportion that the method rejected the dual null hypothesis $H_{0.mh}$ at significance level 0.005 and 0.0005 respectively.

Figure 2-7: Q-Q plots of eight gene-based methods. The inflation factors of HGAT and wHGAT appeared reasonable and indicate no obvious inflation. The curve of other methods appeared within the gray band (95% concentration band), which indicates no obvious inflations.

Figure 3-1: Null joint distributions of the test statistics on mean and variance heterogeneities under normality setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 40$) from the standard normal distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure 3-2: Comparison of false positive rates of eight methods under standard normality setting. Each panel was computed from 100000 replicates of two-group samples with the specified samples sizes simulated from $\mathcal{N}(0,1)$. At each significance level, the false positive rate of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis H_{03}. The gray belt is the 95% concentration band of the false positive rates of a typical test that can properly control false positive rates at given nominal significance levels.

Figure 3-3: Comparison of false positive rates of eight methods under standard Laplace setting. Each panel was computed from 100000 replicates of two-group samples with the specified samples sizes simulated from $\mathcal{L}aplace(0,1)$. At each significance level, the false positive rate of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis H_{03}. The gray belt is the 95% concentration band of the false positive rates of a typical test that can properly control false positive rates at given nominal significance levels.

Figure 3-4: Power comparison of six methods under two-condition normality setting. In each panel, for each specific (r, s) pair, powers of the six methods were computed from 100000 replicates of two-group samples with samples sizes (40 vs. 40) simulated from $\mathcal{N}(0,1)$ and $\mathcal{N}(r, (1+s)^2)$, respectively. At each (r, s) pair, the power of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis H_{03} at significance level 0.05. For the SMVT, both the significance

level of Welch test and that of the Levene test were set to be $1-\sqrt{1-0.05}$ to control overall type I error rate at 0.05.

Figure 3-5: Power comparison of six methods under two-condition Laplace setting. In each panel, for each specific (r, s) pair, powers of the six methods were computed from 100000 replicates of two-group samples with samples sizes (40 vs. 40) simulated from $\mathcal{L}aplace(0,1)$ and $\mathcal{L}aplace(r,(1+s)^2)$, respectively. At each (r, s) pair, the power of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis H_{03} at significance level 0.05. For the SMVT, both the significance level of Welch test and that of the Levene test were set to be $1-\sqrt{1-0.05}$ to control overall type I error rate at 0.05.

Figure 3-6: Q-Q plots of the five competitors without adjusting for latent data structure and covariates. Using the MAS5, we normalized the raw expression data of the 22283 gene probes on the 39 smokers and 40 nonsmokers. We then compute gene probe specific statistics and *p* values of the tests statistics based on the MAS5 normalized data. The inflation factors of all the tests appeared unreasonably huge, especially that of the STSD. All the curves clearly appeared above the gray band (95% concentration band). The striking inflations implied that some latent factors severely confounded the competitors.

Figure 3-7: Global data structure of all the experiment-wide gene expression levels. Using MAS5, we normalized the raw expression levels of the 22283 experiment-wide

gene probes and computed the PCs of all the normalized expression levels. PC1 alone accounted for 98.24% of the total variation and was the unique major PC. PC2 merely accounted for 0.32% of total variation. Neither PC1 nor PC2 displayed mean heterogeneity or variance heterogeneity. PC4 displayed strikingly significant mean heterogeneity ($p_{WT} = 1.91 \times 10^{-15}$), even if it only accounted for 0.13% of the total variation. PC6 displayed very significant variance heterogeneity ($p_{LF} = 3.18 \times 10^{-4}$) even if it accounted for 0.07% of the total variation only. PC4 and PC6 clearly distinguished the smokers and the nonsmokers.

Figure 3-8: Deflations due to the over adjustment of the experiment-wide data structure. Among all the 79 global PCs, only PC4 displayed significant mean heterogeneity (p_{WT}=4.49E-15). PC6, 9, 14, 28, 38, 49 and 78 displayed variance heterogeneity (p_{LF} ranged from 3.18E-4 to 0.0419). After adjusting for the significant global PCs, age and menopausal status, the *Q*-*Q* plots of all the five competitors displayed severe deflations. All the genomic inflation factors turned out to be much smaller than 1. The *Q*-*Q* plots of the four mean heterogeneity tests fell below the diagonal, where those of the WT and the MWT fell below the lower limit of the 95% concentration band. Global PCs did not distinguish informative heterogeneities and impediment heterogeneities. The significant global PCs would account for big portions of informative mean and variance heterogeneities due to DE genes. Therefore, adjusting for the significant PCs of all the experiment-wide gene probes would reduce statistical powers

Figure 3-9: Background data structure of the expression levels of robust gene probes. From the MAS5 normalized data, we selected 13415 robust gene probes and conducted background PCA. PC1 alone accounted for 98.35% of the total variation and was the unique major PC. PC2 merely accounted for 0.37% of total variation. Neither PC1 nor PC2 displayed mean heterogeneity or variance heterogeneity. PC14 displayed significant mean heterogeneity ($p_{WT} = 0.0036$), even if it only accounted for 0.03% of the total variation. PC28 displayed significant variance heterogeneity ($p_{LF} = 0.0069$) even if it accounted for 0.01% of the total variation only. PC14 and PC28 displayed clear stratification of the smokers and the nonsmokers.

Figure 3-10: Q-Q plots of the five competitors after adjusting for background data structure and covariates. Among all the 79 background PCs, PC14, PC25, PC12, PC16, and PC18 displayed significant mean heterogeneity (p_{WT} ranged from 0.0036 to 0.0149). PC28, PC29 and PC30 displayed variance heterogeneity (p_{LF} ranged from 0.0069 to 0.0208). After adjusting for these significant background PCs, age and menopausal status, the *Q-Q* plots of all the five tests climbed above the diagonal. Especially, the Q-Q plot of the IMVT climbed above the upper limit of the 95% concentration band. All the tests displayed reasonable inflation factors. The mild inflation could be due to weak differentials or residual correlations between DE genes. Adjusting for significant background PCs was necessary to prevent false positives and false negatives.

Figure 3-11: Boxplots of four experiment-wide significant gene probes. After calibrating the background data structure, no gene probes appeared experiment-wide significant mean heterogeneity. All of these four genes displayed certain significance of mean heterogeneity and displayed nearly experiment-wide significant variance heterogeneity. Integrating variance heterogeneity and mean heterogeneity led us to identify these four gene probes to be experiment-wide significant.

Figure 3-12: Comparison of false positive rates of six methods under standard normality setting. WT, MWT, STSD, IMVT, SMVT and DWT are the competitors here. Each panel was computed from 100000 replicates of two-group samples with the specified samples sizes simulated from $\mathcal{N}(0,1)$. At each significance level, the false positive rate of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis. The gray belt is the 95% concentration band of the false positive rates of a typical test that can properly control false positive rates at given nominal significance levels.

Figure 3-13: Power comparison of six methods with different mean heterogeneities levels at nominal level 0.05. WT, MWT, STSD, IMVT, SMVT and DWT are the competitors here. We consider $r = 0,0.25,0.5$ in (a)-(c). In each panel, for each specific s, powers of the six methods were computed from 100000 replicates of two-group samples with samples sizes (40 vs. 40) simulated from $\mathcal{N}(0,1)$ and $\mathcal{N}(r,(1+s)^2)$, respectively. The power of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis at significance level 0.05. For the SMVT,

both the significance level of Welch test and that of the Levene test were set to be $1 - \sqrt{1 - 0.05}$ to control overall type I error rate at 0.05.

Figure 3-14: Power comparison of DWT and IMVT at nominal level 0.05 and 0.005, respectively. IMVT and DWT are the competitors here. When no variance heterogeneity exist ($s = 0$), for each specific r, powers of the six methods were computed from 100000 replicates of two-group samples with samples sizes (40 vs. 40) simulated from $\mathcal{N}(0,1)$ and $\mathcal{N}(r, (1+s)^2)$, respectively. The power of each method was estimated by the empirical proportion that the method rejected the dual null hypothesis at s at nominal level 0.05 and 0.005, respectively

Figure A-1: The Manhattan plot of MT. 856149 SNPs with MAF>0.005 was plotted. No obvious association signal peaks were observed using MT. The gray line is the suggestive nominal level 5×10^{-6}.

Figure B-1: Null joint distributions of mean and variance test statistics under 5 vs. 5 normality setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 5$) from the standard normal distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-2: Null joint distributions of mean and variance test statistics under 10 vs.10 normality setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 10$) from the standard normal distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-3: Null joint distributions of mean and variance test statistics under 20 vs. 20 normality setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 20$) from the standard normal distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-4: Null joint distributions of mean and variance test statistics under 5 vs. 5 Laplace setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 5$) from the standard Laplace distribution. Panel (**a**) shows the null independence between

Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-5: Null joint distributions of mean and variance test statistics under 10 vs. 10 Laplace setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 10$) from the standard Laplace distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-6: Null joint distributions of mean and variance test statistics under 20 vs. 20 Laplace setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 20$) from the standard Laplace distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

Figure B-7: Null joint distributions of mean and variance test statistics under 40 vs. 40 Laplace setting. Each panels displays 100000 pairs of the specified test statistics, which were computed from 100000 replicates of two-group samples of sizes ($n_1 = n_2 = 40$) from the standard Laplace distribution. Panel (**a**) shows the null independence between Welch *t* statistic and Levene statistic. Panel (**b**) shows the null independence between Welch *t*-statistic and *F*-statistic. Panel (**c**) shows the equivalence between Welch *t* statistic and Student t statistic. Panel (**d**) shows the high correlation between Levene test statistic and Brown-Forsythe statistic.

APPENDIX D R CODES

All the codes of Methods and Simulation, figures are written in R. Please click the link of my github account to access them https://github.com/oyww710.

APPENDIX E PUBLICATIONS

Ouyang, W., An, Q., Zhao, J., & Qin, H. (2016). Integrating mean and variance heterogeneities to identify differentially expressed genes. BMC Bioinformatics, 17(1), 497.

Qin, H. and **Ouyang, W.** (2016). Asymmetric risk of the stein variance estimator under a mispecified linear regression model. *Statistics & Probability Letters*, *116*, 94-100.

Wenan, C., Ren, C., Qin, H., Archer, K., **Ouyang, W.**, Nianjun Liu, and Xiangning Chen. (2016). A generalized sequential bonferroni procedure for GWAS in admixed

populations incorporating admixture mapping information into association tests. *Human heredity, 79(*2), 80-92.

Qin, H. and **Ouyang, W.** (2015). Statistical properties of gene-gene correlations in omics experiments. *Statistics & Probability Letters, 97*, 206-211.

BIBLIOGRAPHY

1. Nelder JA, Baker RJ: **Generalized linear models**. *Encyclopedia of statistical sciences* 1972.
2. Breslow NE, Clayton DG: **Approximate inference in generalized linear mixed models**. *Journal of the American statistical Association* 1993, **88**(421):9-25.
3. McCulloch CE, Neuhaus JM: **Generalized linear mixed models**: Wiley Online Library; 2001.
4. Schork NJ, Nath SK, Fallin D, Chakravarti A: **Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects**. *The American Journal of Human Genetics* 2000, **67**(5):1208-1218.
5. Q Deng W, Paré G: **A fast algorithm to optimize SNP prioritization for gene‐gene and gene‐environment interactions**. *Genetic epidemiology* 2011, **35**(7):729-738.
6. Perry GM, Nehrke KW, Bushinsky DA, Reid R, Lewandowski KL, Hueber P, Scheinman SJ: **Sex modifies genetic effects on residual variance in urinary calcium excretion in rat (Rattus norvegicus)**. *Genetics* 2012, **191**(3):1003-1013.
7. Fraser HB, Schadt EE: **The quantitative genetics of phenotypic robustness**. *PloS one* 2010, **5**(1):e8635.

8. Rönnegård L, Valdar W: **Detecting major genetic loci controlling phenotypic variability in experimental crosses**. *Genetics* 2011, **188**(2):435-447.
9. Struchalin MV, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS: **Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations**. *BMC genetics* 2010, **11**(1):92.
10. Shen X, Pettersson M, Rönnegård L, Carlborg Ö: **Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana**. *PLoS Genet* 2012, **8**(8):e1002839.
11. Cao Y, Wei P, Bailey M, Kauwe JS, Maxwell TJ: **A versatile omnibus test for detecting mean and variance heterogeneity**. *Genetic epidemiology* 2014, **38**(1):51-59.
12. Hulse AM, Cai JJ: **Genetic variants contribute to gene expression variability in humans**. *Genetics* 2013, **193**(1):95-108.
13. Soave D, Corvol H, Panjwani N, Gong J, Li W, Boëlle P-Y, Durie PR, Paterson AD, Rommens JM, Strug LJ: **A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways**. *The American Journal of Human Genetics* 2015, **97**(1):125-138.
14. Levene H: **Robust tests for equality of variances**. *Contributions to probability and statistics* 1960, **1**:278-292.
15. Rosenthal R: **Combining results of independent studies**. *Psychological bulletin* 1978, **85**(1):185.
16. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S: **A genome-wide association study of alcohol dependence**. *Proceedings of the National Academy of Sciences* 2010, **107**(11):5082-5087.
17. Smyth GK, Verbyla AP: **Adjusted likelihood methods for modelling dispersion in generalized linear models**. *Environmetrics* 1999, **10**(6):695-709.
18. Beech RD, Qu J, Leffert JJ, Lin A, Hong KA, Hansen J, Umlauf S, Mane S, Zhao H, Sinha R: **Altered expression of cytokine signaling pathway genes in peripheral blood cells of alcohol dependent subjects: preliminary findings**. *Alcoholism: Clinical and Experimental Research* 2012, **36**(9):1487-1496.
19. Charlesworth JC, Curran JE, Johnson MP, Göring HH, Dyer TD, Diego VP, Kent JW, Mahaney MC, Almasy L, MacCluer JW: **Transcriptomic epidemiology of smoking: the effect of smoking on**

gene expression in lymphocytes. *BMC medical genomics* 2010, **3**(1):1.

20. Dick DM, Meyers J, Aliev F, Nurnberger J, Kramer J, Kuperman S, Porjesz B, Tischfield J, Edenberg HJ, Foroud T: **Evidence for genes on chromosome 2 contributing to alcohol dependence with conduct disorder and suicide attempts**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2010, **153**(6):1179-1188.
21. Johnson C, Drgon T, Liu QR, Walther D, Edenberg H, Rice J, Foroud T, Uhl GR: **Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2006, **141**(8):844-853.
22. Ehringer MA, Thompson J, Conroy O, Xu Y, Yang F, Canniff J, Beeson M, Gordon L, Bennett B, Johnson TE: **High-throughput sequence identification of gene coding variants within alcohol-related QTLs**. *Mammalian genome* 2001, **12**(8):657-663.
23. Edwards AC, Aliev F, Bierut LJ, Bucholz KK, Edenberg H, Hesselbrock V, Kramer J, Kuperman S, Nurnberger Jr JI, Schuckit MA: **Genome-wide association study of comorbid depressive syndrome and alcohol dependence**. *Psychiatric genetics* 2012, **22**(1):31.
24. Riley B, Kalsi G, Kuo P, Vladimirov V, Thiselton D, Vittum J, Wormley B, Grotewiel M, Patterson D, Sullivan PF: **Alcohol dependence is associated with the ZNF699 gene, a human locus related to Drosophila hangover, in the Irish Affected Sib Pair Study of Alcohol Dependence (IASPSAD) sample**. *Molecular psychiatry* 2006, **11**(11):1025-1031.
25. Clark SL, Aberg KA, Nerella S, Kumar G, McClay JL, Chen W, Xie LY, Harada A, Shabalin AA, Gao G: **Combined whole methylome and genomewide association study implicates CNTN4 in alcohol use**. *Alcoholism: Clinical and Experimental Research* 2015, **39**(8):1396-1405.
26. Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, De Moor MH, Smit AB, Hottenga J-J, Richter MM, Heath AC: **A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations**. *Twin research and human genetics: the*

official journal of the International Society for Twin Studies 2010, **13**(1):10.

27. Gelernter J, Kranzler H, Sherva R, Almasy L, Koesterer R, Smith A, Anton R, Preuss U, Ridinger M, Rujescu D: **Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci**. *Molecular psychiatry* 2014, **19**(1):41-49.
28. Spanagel R, Bartsch D, Brors B, Dahmen N, Deussing J, Eils R, Ende G, Gallinat J, Gebicke‐Haerter P, Heinz A: **An integrated genome research network for studying the genetics of alcohol addiction**. *Addiction biology* 2010, **15**(4):369-379.
29. Johnson C, Drgon T, Walther D, Uhl GR: **Genomic regions identified by overlapping clusters of nominally-positive SNPs from genome-wide studies of alcohol and illegal substance dependence**. *PloS one* 2011, **6**(7):e19210.
30. Kendler KS, Kalsi G, Holmans PA, Sanders AR, Aggen SH, Dick DM, Aliev F, Shi J, Levinson DF, Gejman PV: **Genomewide association analysis of symptoms of alcohol dependence in the molecular genetics of schizophrenia (MGS2) control sample**. *Alcoholism: Clinical and Experimental Research* 2011, **35**(5):963-975.
31. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.
32. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene**. *science* 2006, **314**(5804):1461-1463.
33. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR: **A common allele on chromosome 9 associated with coronary heart disease**. *Science* 2007, **316**(5830):1488-1491.
34. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.

35. Morris AP, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies**. *Genetic epidemiology* 2010, **34**(2):188-193.
36. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N: **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease**. *Nature genetics* 2011, **43**(11):1066-1073.
37. Asselbergs FW, Guo Y, Van Iperen EP, Sivapalaratnam S, Tragante V, Lanktree MB, Lange LA, Almoguera B, Appelman YE, Barnard J: **Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci**. *The American Journal of Human Genetics* 2012, **91**(5):823-838.
38. Diogo D, Kurreeman F, Stahl EA, Liao KP, Gupta N, Greenberg JD, Rivas MA, Hickey B, Flannick J, Thomson B: **Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis**. *The American Journal of Human Genetics* 2013, **92**(1):15-27.
39. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data**. *The American Journal of Human Genetics* 2008, **83**(3):311-321.
40. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare-variant association testing for sequencing data with the sequence kernel association test**. *The American Journal of Human Genetics* 2011, **89**(1):82-93.
41. Lee S, Wu MC, Lin X: **Optimal tests for rare variant effects in sequencing association studies**. *Biostatistics* 2012, **13**(4):762-775.
42. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team ELP, Christiani DC, Wurfel MM, Lin X: **Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies**. *The American Journal of Human Genetics* 2012, **91**(2):224-237.
43. Sun J, Zheng Y, Hsu L: **A Unified Mixed‐Effects Model for Rare‐Variant Association in Sequencing Studies**. *Genetic epidemiology* 2013, **37**(4):334-344.
44. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies**. *Genetics* 2003, **164**(4):1567-1587.

45. Moschopoulos P, Canada W: **The distribution function of a linear combination of chi-squares**. *Computers & mathematics with applications* 1984, **10**(4-5):383-386.
46. Sankararaman S, Sridhar S, Kimmel G, Halperin E: **Estimating local ancestry in admixed populations**. *The American Journal of Human Genetics* 2008, **82**(2):290-303.
47. Zhu X, Tang H, Risch N: **Admixture mapping and the role of population structure for localizing disease genes**. *Advances in genetics* 2008, **60**:547-569.
48. Consortium GP: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
49. Zhang J, Stram DO: **The role of local ancestry adjustment in association studies using admixed populations**. *Genetic epidemiology* 2014, **38**(6):502-515.
50. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS: **Truncated product method for combining P‐values**. *Genetic epidemiology* 2002, **22**(2):170-185.
51. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. *Genetics* 1989, **121**(1):185-199.
52. Taillon-Miller P, Bauer-Sardiña I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok P-Y: **Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28**. *Nature genetics* 2000, **25**(3):324-328.
53. Collins A, Lonjou C, Morton N: **Genetic epidemiology of single-nucleotide polymorphisms**. *Proceedings of the National Academy of Sciences* 1999, **96**(26):15173-15177.
54. Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A: **Extent and distribution of linkage disequilibrium in three genomic regions**. *The American Journal of Human Genetics* 2001, **68**(1):191-197.
55. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes**. *Nature genetics* 1999, **22**(2):139-144.
56. Kempthorne O: **The correlation between relatives on the supposition of Mendelian inheritance**. *American journal of human genetics* 1968, **20**(4):402.

57. Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L, Bierut LJ, Bucholz KK, Goate A, Aliev F: **Genome‐wide association study of alcohol dependence implicates a region on chromosome 11**. *Alcoholism: Clinical and Experimental Research* 2010, **34**(5):840-852.
58. Zuo L, Lu L, Tan Y, Pan X, Cai Y, Wang X, Hong J, Zhong C, Wang F, Zhang XY: **Genome‐wide association discoveries of alcohol dependence**. *The American journal on addictions* 2014, **23**(6):526-539.
59. Sikela JM, MacLaren EJ, Kim Y, Karimpour‐Fard A, Cai WW, Pollack J, Hitzemann R, Belknap J, McWeeney S, Kerns RT: **DNA microarray and proteomic strategies for understanding alcohol action**. *Alcoholism: Clinical and Experimental Research* 2006, **30**(4):700-708.
60. Moller S, Becker U, Juul A, Skakkebaek NE, Christensen E: **Prognostic value of insulinlike growth factor I and its binding protein in patients with alcohol‐induced liver disease**. *Hepatology* 1996, **23**(5):1073-1078.
61. Pochareddy S, Edenberg HJ: **Chronic alcohol exposure alters gene expression in HepG2 cells**. *Alcoholism: Clinical and Experimental Research* 2012, **36**(6):1021-1033.
62. Tessema M, Yingling CM, Liu Y, Tellez CS, Van Neste L, Baylin SS, Belinsky SA: **Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers**. *Carcinogenesis* 2014, **35**(6):1248-1257.
63. Na H-K, Kim M, Chang S-S, Kim S-Y, Park JY, Chung MW, Yang M: **Tobacco smoking-response genes in blood and buccal cells**. *Toxicology letters* 2015, **232**(2):429-437.
64. Zuo L, Gelernter J, Zhang CK, Zhao H, Lu L, Kranzler HR, Malison RT, Li C-SR, Wang F, Zhang X-Y: **Genome-wide association study of alcohol dependence implicates KIAA0040 on chromosome 1q**. *Neuropsychopharmacology* 2012, **37**(2):557-566.
65. Zuo L, Zhang CK, Wang F, Li C-SR, Zhao H, Lu L, Zhang X-Y, Lu L, Zhang H, Zhang F: **A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study**. *PloS one* 2011, **6**(11):e26726.
66. Wang K-S, Liu X, Zhang Q, Zeng M: **ANAPC1 and SLCO3A1 are associated with nicotine dependence: meta-analysis of genome-**

wide association studies. *Drug and alcohol dependence* 2012, **124**(3):325-332.

67. Hill SY, Jones BL, Zezza N, Stiffler S: **Family-based association analysis of alcohol dependence implicates KIAA0040 on Chromosome 1q in multiplex alcohol dependence families**. *Open journal of genetics* 2013, **3**(4):243.
68. Wang K-S, Liu X, Zhang Q, Pan Y, Aragam N, Zeng M: **A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence**. *Journal of psychiatric research* 2011, **45**(11):1419-1425.
69. Forero DA, López-León S, Shin HD, Park BL, Kim D-J: **Meta-analysis of six genes (BDNF, DRD1, DRD3, DRD4, GRIN2B and MAOA) involved in neuroplasticity and the risk for alcohol dependence**. *Drug and alcohol dependence* 2015, **149**:259-263.
70. Wang L, Liu X, Luo X, Zeng M, Zuo L, Wang K-S: **Genetic variants in the fat mass-and obesity-associated (FTO) gene are associated with alcohol dependence**. *Journal of Molecular Neuroscience* 2013, **51**(2):416-424.
71. Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, De Moor MH, Smit AB, Hottenga J-J, Richter MM, Heath AC: **A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations**. *Twin Research and Human Genetics* 2010, **13**(01):11-29.
72. Nikpay M: **Genome wide search for genetic determinants of habitual alcohol, tobacco and coffee use, obesity-related traits, response to mental and physical stress and hemodynamic traits**. 2011.
73. Latella MC, Di Castelnuovo A, De Lorgeril M, Arnout J, Cappuccio FP, Krogh V, Siani A, Van Dongen M, Donati MB, De Gaetano G: **Genetic variation of alcohol dehydrogenase type 1C (ADH1C), alcohol consumption, and metabolic cardiovascular risk factors: results from the IMMIDIET study**. *Atherosclerosis* 2009, **207**(1):284-290.
74. Mulligan CJ, Robin RW, Osier MV, Sambuughin N, Goldfarb LG, Kittles RA, Hesselbrock D, Goldman D, Long JC: **Allelic variation at alcohol metabolism genes (ADH1B, ADH1C, ALDH2) and alcohol dependence in an American Indian population**. *Human genetics* 2003, **113**(4):325-336.
75. Peters ES, McClean MD, Liu M, Eisen EA, Mueller N, Kelsey KT: **The ADH1C polymorphism modifies the risk of squamous cell**

carcinoma of the head and neck associated with alcohol and tobacco use. *Cancer Epidemiology and Prevention Biomarkers* 2005, **14**(2):476-482.

76. Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, Moessner R, Gaebel W, Dahmen N: **Genome-wide association study of alcohol dependence**. *Archives of general psychiatry* 2009, **66**(7):773-784.
77. Treutlein J, Rietschel M: **Genome-wide association studies of alcohol dependence and substance use disorders**. *Current psychiatry reports* 2011, **13**(2):147-155.
78. Xu W, Cohen-Woods S, Chen Q, Noor A, Knight J, Hosang G, Parikh SV, De Luca V, Tozzi F, Muglia P: **Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1**. *BMC medical genetics* 2014, **15**(1):2.
79. Wang K-S, Liu X, Zhang Q, Wu L-Y, Zeng M: **Genome-wide association study identifies 5q21 and 9p24. 1 (KDM4C) loci associated with alcohol withdrawal symptoms**. *Journal of Neural Transmission* 2012, **119**(4):425-433.
80. Yin X-y, Cheng H, Wang W-j, Wang W-j, Fu H-y, Liu L-h, Zhang F-y, Yang S, Zhang X-j: **TNIP1/ANXA6 and CSMD1 variants interacting with cigarette smoking, alcohol intake affect risk of psoriasis**. *Journal of dermatological science* 2013, **70**(2):94-98.
81. Han S, Yang B-Z, Kranzler HR, Liu X, Zhao H, Farrer LA, Boerwinkle E, Potash JB, Gelernter J: **Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence**. *The American Journal of Human Genetics* 2013, **93**(6):1027-1034.
82. Yang BZ, Kranzler HR, Zhao H, Gruen JR, Luo X, Gelernter J: **Haplotypic variants in DRD2, ANKK1, TTC12, and NCAM1 are associated with comorbid alcohol and drug dependence**. *Alcoholism: Clinical and Experimental Research* 2008, **32**(12):2117-2127.
83. Yang B-Z, Kranzler HR, Zhao H, Gruen JR, Luo X, Gelernter J: **Association of haplotypic variants in DRD2, ANKK1, TTC12 and NCAM1 to alcohol dependence in independent case–control and family samples**. *Human molecular genetics* 2007, **16**(23):2844-2853.
84. Mulligan MK, Ponomarev I, Hitzemann RJ, Belknap JK, Tabakoff B, Harris RA, Crabbe JC, Blednov YA, Grahame NJ, Phillips TJ: **Toward understanding the genetics of alcohol drinking through**

transcriptome meta-analysis. *Proceedings of the National Academy of Sciences* 2006, **103**(16):6368-6373.

85. Biernacka JM, Geske J, Jenkins GD, Colby C, Rider DN, Karpyak VM, Choi D-S, Fridley BL: **Genome-wide gene-set analysis for identification of pathways associated with alcohol dependence**. *International Journal of Neuropsychopharmacology* 2013, **16**(2):271-278.

86. Duell EJ, Sala N, Travier N, Muñoz X, Boutron-Ruault MC, Clavel-Chapelon F, Barricarte A, Arriola L, Navarro C, Sánchez-Cantalejo E: **Genetic variation in alcohol dehydrogenase (ADH1A, ADH1B, ADH1C, ADH7) and aldehyde dehydrogenase (ALDH2), alcohol consumption and gastric cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort**. *Carcinogenesis* 2011:bgr285.

87. Crabb DW, Edenberg HJ, Bosron WF, Li T-K: **Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2 (2) allele is dominant**. *Journal of Clinical Investigation* 1989, **83**(1):314.

88. Morozova TV, Goldman D, Mackay TF, Anholt RR: **The genetic basis of alcoholism: multiple phenotypes, many genes, complex networks**. *Genome biology* 2012, **13**(2):239.

89. Zuo L, Zhang F, Zhang H, Zhang XY, Wang F, Li CSR, Lu L, Hong J, Lu L, Krystal J: **Genome‐wide search for replicable risk gene regions in alcohol and nicotine co‐dependence**. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2012, **159**(4):437-444.

90. Guan Y: **Detecting structure of haplotypes and local ancestry**. *Genetics* 2014, **196**(3):625-642.

91. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.

92. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron J, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proceedings of the National Academy of Sciences* 2003, **100**(14):8418-8423.

93. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT: **Gene**

expression profiling predicts clinical outcome of breast cancer. *nature* 2002, **415**(6871):530-536.

94. Jeanmougin M, De Reynies A, Marisa L, Paccard C, Nuel G, Guedj M: **Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies**. *PloS one* 2010, **5**(9):e12336.

95. Glass GV, Peckham PD, Sanders JR: **Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance**. *Review of educational research* 1972, **42**(3):237-288.

96. Welch BL: **The generalization ofstudent's' problem when several different population variances are involved**. *Biometrika* 1947, **34**(1/2):28-35.

97. Demissie M, Mascialino B, Calza S, Pawitan Y: **Unequal group variances in microarray data analyses**. *Bioinformatics* 2008, **24**(9):1168-1174.

98. Qin H, Feng T, Harding SA, Tsai C-J, Zhang S: **An efficient method to identify differentially expressed genes in microarray experiments**. *Bioinformatics* 2008, **24**(14):1583-1589.

99. Qin H, Ouyang W: **Statistical properties of gene–gene correlations in omics experiments**. *Statistics & Probability Letters* 2015, **97**:206-211.

100. Markowski CA, Markowski EP: **Conditions for the effectiveness of a preliminary test of variance**. *The American Statistician* 1990, **44**(4):322-326.

101. Levene H: **Robust tests for equality of variances1**. *Contributions to probability and statistics: Essays in honor of Harold Hotelling* 1960, **2**:278-292.

102. Brown MB, Forsythe AB: **Robust tests for the equality of variances**. *Journal of the American Statistical Association* 1974, **69**(346):364-367.

103. Pan F, Yang T-L, Chen X-D, Chen Y, Gao G, Liu Y-Z, Pei Y-F, Sha B-Y, Jiang Y, Xu C: **Impact of female cigarette smoking on circulating B cells in vivo: the suppressed ICOSLG, TCF3, and VCAM1 gene functional network may inhibit normal cell function**. *Immunogenetics* 2010, **62**(4):237-251.

104. Games PA, Keselman HJ, Clinch JJ: **Tests for homogeneity of variance in factorial designs**. *Psychological Bulletin* 1979, **86**(5):978.

105. O'Brien RG: **Robust techniques for testing heterogeneity of variance effects in factorial designs**. *Psychometrika* 1978, **43**(3):327-342.
106. Devlin B, Roeder K: **Genomic control for association studies**. *Biometrics* 1999, **55**(4):997-1004.
107. Gagnon-Bartsch JA, Speed TP: **Using control genes to correct for unwanted variation in microarray data**. *Biostatistics* 2012, **13**(3):539-552.
108. Geraghty P, Wyman AE, Garcia-Arcos I, Dabo AJ, Gadhvi S, Foronjy R: **STAT3 modulates cigarette smoke-induced inflammation and protease expression**. *Second hand smoke and COPD: lessons from animal studies* 2015:23.
109. Halappanavar S, Russell M, Stampfli MR, Williams A, Yauk CL: **Induction of the interleukin 6/signal transducer and activator of transcription pathway in the lungs of mice sub-chronically exposed to mainstream tobacco smoke**. *BMC medical genomics* 2009, **2**(1):1.
110. Nasreen N, Gonzalves L, Peruvemba S, Mohammed KA: **Fluticasone furoate is more effective than mometasone furoate in restoring tobacco smoke inhibited SOCS-3 expression in airway epithelial cells**. *International immunopharmacology* 2014, **19**(1):153-160.
111. Rager JE, Bauer RN, Müller LL, Smeester L, Carson JL, Brighton LE, Fry RC, Jaspers I: **DNA methylation in nasal epithelial cells from smokers: identification of ULBP3-related effects**. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 2013, **305**(6):L432-L438.
112. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas Y-M, Calner P, Sebastiani P: **Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer**. *Nature medicine* 2007, **13**(3):361-366.
113. Boelens MC, van den Berg A, Fehrmann RS, Geerlings M, de Jong WK, te Meerman GJ, Sietsma H, Timens W, Postma DS, Groen HJ: **Current smoking - specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer**. *The Journal of pathology* 2009, **218**(2):182-191.
114. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW: **Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival**. *PloS one* 2008, **3**(2):e1651.

115. Wang X, Chorley BN, Pittman GS, Kleeberger SR, Brothers II J, Liu G, Spira A, Bell DA: **Genetic variation and antioxidant response gene expression in the bronchial airway epithelium of smokers at risk for lung cancer**. *PLoS One* 2010, **5**(8):e11934.
116. Gümüş ZH, Du B, Kacker A, Boyle JO, Bocker JM, Mukherjee P, Subbaramaiah K, Dannenberg AJ, Weinstein H: **Effects of tobacco smoke on gene expression and cellular pathways in a cellular model of oral leukoplakia**. *Cancer Prevention Research* 2008, **1**(2):100-111.
117. Boyle JO, Gümüş ZH, Kacker A, Choksi VL, Bocker JM, Zhou XK, Yantiss RK, Hughes DB, Du B, Judson BL: **Effects of cigarette smoke on the human oral mucosal transcriptome**. *Cancer Prevention Research* 2010, **3**(3):266-278.
118. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research**. *Behavioural brain research* 2001, **125**(1):279-284.
119. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995:289-300.
120. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency**. *Annals of statistics* 2001:1165-1188.
121. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures**. *Bioinformatics* 2003, **19**(3):368-375.
122. Casella G, Berger RL: **Statistical inference**, vol. 2: Duxbury Pacific Grove, CA; 2002.