EXPLORE THE RELATIONSHIP BETWEEN GENETIC VARIATIONS AND PHENOTYPES WITH BAYESIAN APPROACHES

by

Jianan Zhan

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2018

© Jianan Zhan 2018

All rights reserved

Abstract

Genome-wide association studies (GWAS) have had great success in identifying human genetic variants associated with human traits. With recent developments in high throughput biology, immense amount of data have been generated, thus calling for novel statistical and computational approaches to be developed and draw biological meaningful conclusions. A current direction for GWAS method development has been to use Bayesian approaches, where prior beliefs of variant effects are incorporated into test statistics, to boost the power to detect real associations. With previous success in developing Bayesian-based GWAS method for single phenotype, in this work the Bayesian idea is extended to multiple phenotypes, aiming at developing a method that detects pleiotropic genome-wide associations. Alongside with the method development, analytical simulations were also performed to investigate into the possible power gain by using such Bayesian approaches, as well as to understand how different factors influence the behavior of Bayesian-based GWAS methods.

Many variants are pleiotropic, and discovery of these variants could help reveal disease mechanisms, suggest new therapeutic options. Therefore, we developed a pleiotropic GWAS method based on Bayesian framework, SNP And Pleiotropic PHenotype Organization (SAPPHO), which learns pleiotropy using identified associations to discover additional associations with shared patterns. SAPPHO was applied on two sets of real data: 1. Atherosclerosis Risk in Communities (ARIC) study of 8,000 individuals, whose gold-standard associations were provided by meta-analysis of 40,000 to 100,000 individuals from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium; 2. Cancer phenotypes from UK Biobank project, consisting several hundred to 15,000 individuals, with gold-standard obtained from GWAS catalog. For both data sets, SAPPHO was able to detect additional associations that were not detected with the conventional univariate test, and boost power when different variants follow the same association patterns.

Bayesian approaches boost power for GWAS through alleviating burdens from multiple hypothesis testings, which is usually on the scale of thousands to millions. Intuitively, by making use of prior probabilities that bias favored sets thought to be enriched for significant findings, power for detecting true associations could be increased. Therefore, an analytical study was conducted here to see theoretically to what extent power could gain by using such approaches, and how does this gain depend on different factors. By calculating test power assuming perfect knowledge of a prior distribution, the population size increase required to provided the same boost without a prior was obtained, and it is shown that population size is exponentially more important than prior, providing a rigorous proof for the lack of use for prior-

ABSTRACT

based GWAS methods.

Thesis Advisor:

Joel S. Bader, Ph.D.

Thesis Committee:

Dan E. Arking, Ph.D. (Chair)

Alexis Battle, Ph.D.

Joel S. Bader, Ph.D. (Advisor)

Acknowledgments

First I would like to thank my advisor, Dr. Joel Bader, who has been a great advisor and mentor. He has been a constant source of advice and encouragement. His nice personality and endless flow of ideas have made it a great pleasure to work with him. I have learned from him how to be a good biomedical engineering researcher, and more importantly, how to be a great person. Without his constant guidance and support, I would have never made it through my PhD study.

Thanks to my thesis committee for their advice over the years: Dr. Alexis Battle and Dr. Dan Arking. Thank you for introducing me to the field of Human Genetics, and overseeing me closely on my thesis projects. Your expertise in the field and feedback on my research were of great value to me.

Thanks to all my labmates in the Bader Lab, including Dr. Hailiang Huang, Dr. Pritam Chanda, Dr. Nirmalya Bandyopadhyay, Dr. Giovanni Stracquadanio, Dr. Elisa Pappalardo, Dr. Kun Yang, Dr. Shlomit Edinger (post-doc fellows), Dr. Yongjin Park, Dr. Yasir Suhail, Mr. Will Matern (graduate students). It was such a pleasure working with all of you. Your thoughtful and unselfish suggestions on my research as well as life were always more than helpful to me. I am really blessed to have the opportunity to work with all of you.

Finally, I would like to say thanks to my family. Thanks to my parents who have been supporting me all the time on every decision I made. They do not speak English and might never know what I am writing here, but I am sure they can feel it on the other side of the planet. Thanks to my wife Ying, who has been standing on my side through all these years, and companied me through all of my happiness and sorrows. Thanks to my son Francis and my daughter Fang, who taught me to enjoy and appreciate the real value of life.

My research is supported by the NIH grant R01HL116747 and NSF grant DMS-1228248.

Contents

Abstract ii

Acknowledgments v

List of Tables x

List of Figures xii

1 Introduction 1

2 Discovering patterns of pleiotropy in genome-wide association studies 7

2.1 Introduction 11

2.2 Methods 16

2.2.1 Genetic model 16

2.2.2 Model prior 20

2.2.3 Model score 23

2.2.4 Model search and variant ranking 25

2.2.5 Test statistic and significance thresholds 27

2.2.6 Other methods 29

2.3 Assessment with simulated data 32

2.3.1 Methods 32

2.3.1.1 Genes-only simulations 34

2.3.1.2 Genes-and-environment simulations 34

2.3.1.3 Mixture of genetic and non-genetic phenotypes . . . 36

2.3.2 Results 37

2.3.2.1 Genes-only simulations 37

2.3.2.2 Genes-and-environment simulations 38

2.3.2.3 Mixture of genetic and non-genetic phenotypes . . . 40

2.4 Assessment with real data and pleiotropy on ECG phenotypes 42

2.4.1 Electrocardiogram phenotypes 42

2.4.2 Methods 43

2.4.3 Results 49

2.4.3.1 Assessment for different methods on ARIC data . . . 49

2.4.3.2 Pleiotropy of ECG traits in ARIC and CHARGE . . 52

2.5 Application to UK Biobank cancer phenotypes 56

2.5.1 UK Biobank project and cancer phenotypes 56

2.5.2 Methods 58

2.5.3 Results . . . 62
2.6 Discussion . . . 64
3 The paltry power of priors versus populations 84
3.1 Introduction . . . 87
3.2 Methods . . . 89
3.2.1 Hypothesis testing . . . 89
3.2.2 Hard prior . . . 91
3.2.3 Soft prior . . . 95
3.3 Results . . . 99
3.3.1 Hard prior . . . 99
3.3.2 Soft prior . . . 100
3.3.2.1 Power for the two classes . . . 101
3.3.2.2 Average power for combining both classes . . . 102
3.4 Discussion . . . 103
4 Discussion 110
Bibliography 116
Curriculum Vitae 123

List of Tables

2.1 Running time for different methods for a simulation with 10,000 individuals, 6 phenotypes, and 638 SNPs of which 24 had associations, with Core i5 2.9 GHz CPU, 8 GB RAM. 12 SNPs are associated the first three phenotypes, while the other 12 SNPs are associated with the second three phenotypes. 75

2.2 Assessed methods provide qualitatively different types of predictions. ‘Parsimonious’ indicates that a single SNP is selected from an LD block, whereas ‘Non-Parsimonious’ indicates that all SNPs in an LD block are reported. ‘Phenotype selection’ indicates that the subset of associated phenotypes is reported, whereas ‘No phenotype selection’ indicates that the results do not specify which phenotypes are associated with a SNP and which are not. 76

2.3 Additional novel associations detected by SAPPHO-I in known gold-standard loci. **Locus**: Gene symbol or symbols spanning an associated region. **Univariate GWS**: phenotypes detected at 5×10^{-8} genome-wide significance threshold. **SAPPHO 0.05 FDR**: phenotypes detected by SAPPHO at 0.05 FDR. **SAPPHO, SNPs**: SNPs detected by SAPPHO, with rows corresponding to the previous *SAPPHO 0.05 FDR* column. $\boldsymbol{P_{PR}}$, $\boldsymbol{P_{QRS}}$, $\boldsymbol{P_{QT}}$: univariate p-values from CHARGE meta-analysis. New associations added to a locus are in two categories: (1) a variant already associated with at least one trait is associated with a new trait or traits (the first 7 loci in the table); (2) a new variant is introduced and is associated with a trait not previously associated with the locus (the last 8 loci in the table). 77

2.4 Consistency of direction of effects and phenotypic correlations for each pair of ECG phenotypes. ‘Correlation’ denotes the correlation of each phenotype-phenotype pair; ‘+/+’ denotes the count of SNPs whose effects are in the same direction; ‘+/−’ denotes the count of SNPs whose effects are in different directions. Of the 23 pairs of association effects, 16 were consistent with the phenotypic correlation, corresponding to a probability of 0.70 (binomial parameter 95% confidence interval $[0.49, 0.84]$) that the direction of effect agreed with the phenotypic correlation. These results demonstrate that SAPPHO is able to detect variants whether or not the direction of genetic effect matches the overall direction of phenotypic correlation. 78
2.5 Information for cancer phenotypes on UKB data. Columns are cancer phenotypes, cancer identification codes according to ICD9 coding, cancer identification codes according to ICD10 coding, and number of cases for each type of cancer. 79
2.6 Number of associations for each cancer phenotype for univariate test at $p = 5 \times 10^{-8}$ and SAPPHO at 0.05 FDR on locus level. No association was detected for either method for 4 out of 13 cancer phenotypes. Univariate test detected in total 54 associations with 5 cancer phenotypes, and SAPPHO detected 66 associations with 9 phenotypes. More than ten associations were detected by both methods for skin cancer, breast cancer, and cancer for male genital organs, with a few associations detected for other cancer phenotypes. All loci detected by univariate tests were also detected by SAPPHO at 0.05 FDR. 80
2.7 Pleiotropic loci detected by univariate test at $p = 5 \times 10^{-8}$ and SAPPHO at 0.05 FDR. The same four loci were detected to be pleiotropic by both methods. For two out of all four pleiotropic loci, SAPPHO was able to detect one extra association. All associations were verified by comparing the detected associations with known associations from GWAS catalog. 81

List of Figures

2.1 Histogram of pleiotropic variants. Counts are for variants associated with at least one trait at genome-wide significance ($p < 5 \times 10^{-8}$) (Welter et al., 2014); no attempt was made to correct for correlation between variants (linkage disequilibrium) or between traits. The number of total variants is 17607, and total number of phenotypes is 785. 68

2.2 Power for associations of 6 phenotypes simulated to be correlated through genetic effects only. Scenarios are Independent (6 phenotypes, no pleiotropy), Block (2 blocks of 3 phenotypes each), and Correlated (all 6 phenotypes correlated). Error bars indicate 95% confidence intervals estimated from 5 repeated runs and a binomial distribution. For the independent scenario lacking pleiotropy, SAPPHO methods performed the best. For the block correlation scenario, MultiPhen leading performed best, followed by SAPPHO. For a single correlated block, all pleiotropy methods perform well. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI); univariate test at loose empirical threshold $p < 5\times10^{-8}$ (UNILE); univariate test at stringent empirical threshold $p < 5\times10^{-8}/6$ for 6 phenotypes (UNISE). 69

2.3 Power to detect associations for scenarios involving strong environmental correlations. Two sets of SNPs each containing 12 variants were simulated to contribute to 4 phenotypes. For the same-direction scenario, all 12 SNPs have positive effect with all phenotypes; for the different-direction scenario, the 12 variants had positive effects for the first 2 phenotype and negative effects for the second 2 phenotypes. SAPPHO-C and SHet were unable to detect same-effect SNPs; SHom could not detect different-direction SNPs. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI). . . 70

2.4 Power to detect associations for scenarios involving weak environmental correlations. Four sets of SNPs each containing 6 variants were simulated to contribute to 4 phenotypes. Scenarios as described in the main text were DBSE, different-block-same-effect; DBDE, different-block-different-effect; SBSE, same-block-same-effect; and SBDE, same-block-different-effect. SAPPHO-I and MultiPhen performed the best over all scenarios. SAPPHO-C experiences dramatic loss of power under the sbse scenario, similar to its loss of power for similar scenarios involving strong environmental correlations. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI). . . 71

2.5 Simulations of a mixture of genetic and non-genetic phenotypes, with 1, 2, or 3 genetic phenotypes embedded as part of 13 total phenotypes. The remaining non-genetic phenotypes generated as standard normal random variables. The SAPPHO parameter was set to include associations with p-value $< 5 \times 10^{-4}$. Methods were MultiPhen in selection mode (MS), SAPPHO-I with a greedy forward (SAPPHO-I), and the SAPPHO-I score for the true model (True Model). The greedy forward search limits the power of SAPPHO-I; a more sophisticated strategy could improve its power. 72

2.6 Pleiotropy methods were used to detect associations with the PR, QRS, and QT phenotypes in the ARIC cohort. The three measured phenotypes were then augmented with 3, 6, and 10 noise phenotypes. SAPPHO had the greatest power to detect associations regardless of noise phenotypes were present. The pleiotropy methods SHet, MS, MNS had power similar to standard univariate tests. PCH and SHom had lower power than univariate tests, and the performance of SHom degraded further as noise phenotypes were added. All methods were controlled for type I error at FWER = 0.05. Methods were MultiPhen-Selection (MS), Mutiphen-NonSelection (MNS), Principal Components of Heritability (PCH), Homogeneous Test Statistics (SHom), Heterogeneous Test Statistics (SHet), and univariate tests corrected using permutations (UNI). 73

2.7 Number of associations recovered versus number previously known ($p < 5 \times 10^{-8}$ in CHARGE for univariate single phenotype test). First row: CHARGE meta-analysis data as analyzed by univariate tests for each phenotype ($p < 5 \times 10^{-8}$) and by SAPPHO-I (0.05 FDR). The first row denotes univariate test (with $p < 5 \times 10^{-8}$ cut-off) and SAPPHO-I run on CHARGE meta-analysis results; Second row: ARIC data as analyzed by SAPPHO-I and SAPPHO-C using 0.05 FWER threshold. Third row: ARIC data as analyzed by MultiPhen and univariates test using 0.05 FWER threshold, equivalent to $p < 5 \times 10^{-8}/3$ for univariate tests. More discoveries are made with CHARGE (top row) because it is a larger cohort and because 0.05 FDR is a less stringent threshold than 0.05 FWER. SAPPHO has greater power for the ARIC cohort than MultiPhen or univariate tests. Pleiotropic associations discovered by MultiPhen in the ARIC cohort (bottom left panel) may be overpredictions as these associations were not genome-wide significant in the much larger CHARGE cohort. 74

3.1 **Contour plot for critical R^2 for $p = 5 \times 10^{-8}$ at power = 0.8 as a function of prior strength and population size.** The color corresponding to critical value R^2 changes rapidly as population size changes, and doesn't change much as a function of prior strength, indicating that power is much more sensitive to population size compared to restricting tests to a subset of variants. Relationship between prior strength and population size appears to follow a linear relationship on the log-log scale, and the slope of the linear relationship given a fixed value of R^2 indicates the relative importance of the two factors. . . . 105

3.2 **Population size and prior strength to achieve same power.** X-axis denotes the prior strength S; y-axis of the left panel denotes population fraction increase; y-axis of the right panel denotes population size exponent for a fixed effect value. The left panel shows the population size and prior strength to achieve the same power; a prior strength at 1000 will have equivalent power if the cohort size increases by 70%. The right panel shows the relation between population size exponent and prior strength given a fixed effect size. Both figures shows numerically the linear relationship between population size and logarithmic of prior strength. 106

3.3 **Contour plot of power for the favored class and non-favored class at $p = 5 \times 10^{-8}$ threshold for 50% power.** The left panel shows power for the favored class, and the right panel shows power for the non-favored class. X-axis denotes prior strength S, which is equal to inverse of fraction of SNPs in the favored class; larger S value denotes smaller group of favored class and a more focused subset of variants with higher prior. Y-axis denotes fold enrichment of the favored class $\Pr(M_F)/\Pr(M_{NF})$. For power for the favored class, it could be observed that when prior strength is small, the power is insensitive to prior fold enrichment; this is because large fraction for the favored class is essentially equivalent to a less well-defined subset of variants, thus effect of prior enrichment becomes less obvious. When the favored class is more well-defined, corresponding to larges S values and smaller fractions of favored class, the effect of incorporating prior becomes more obvious; this is reflected by the power gain at large S values, and the power increases with higher fold-enrichment. Power of the non-favored class remains at around 50% when fraction of favored class is small or fold-enrichment is small, because small fold enrichment or small fraction of favored subset is unlikely to make big impact to the non-favored class; when the favored class is given a large prior and consists of large proportion of the total variants, the non-favored class begins to loose power. 107

3.4 **Average power for both classes combined at $p = 5 \times 10^{-8}$ threshold for 50% power.** Axes are defined in the same way as for Fig.3.3. Average power is smaller than power for the favored class and larger than power for the non-favored class given the same prior strength and fold enrichment, as a balance between gain of power for the favored class and loss of power for the non-favored class. Shape of the contour is determined by weights of the two different classes: for small prior strength and large fold enrichment, shape of contour is more similar to the favored class in Fig. 3.3; for large prior strength and small fold enrichment, shape is similar to the non-favored class. 108

3.5 **Population size fraction increase to achieve the same average power for at 5×10^{-8}.** Axis are defined in the same way as for Fig. 3.3. Values are N_1/N_2 where N_1 is the population size without prior, and N_2 is population size to achieve the same power using prior. The maximum population size fraction increase is 1.3 fold to obtain the maximum power gain fulfilled through incorporating a prior, further prove the point that population size is of a more crucial factor for association testing. 109

Chapter 1

Introduction

Genome-wide association studies (GWAS) have had great success in identifying human genetic variants associated with human traits, including human diseases, disease risk factors, and other biomedical phenotypes. According to GWAS catalog, more than 17607 genetic variants, primarily single nucleotide polymorphisms (SNPs), have been proved to be associated with at least 785 distinct phenotypes, at genome-wide significance of 5×10^{-8} (Welter et al., 2014). With recent developments in high throughput biology, immense amount of data have been generated; this strongly calls for novel statistical and computational approaches to be developed and draw biological meaningful conclusions.

Among the multiple approaches taken for GWAS method development (Zhu et al., 2015; Liu et al., 2010; Wu et al., 2011; Klei et al., 2008; O'Reilly et al., 2012; Wu et al., 2009; Kim and Xing, 2009; Huang et al., 2011; Chanda et al., 2013), a reasonable and widely-adopted direction for GWAS method development has been to use Bayesian approaches, where prior beliefs of variant effects are incorporated into test statistics, in order to boost the power to detect real associations (Hindorff et al., 2009; Schork et al., 2013; Petersen et al., 2013; Sveinbjornsson et al., 2016; Meyer et al., 2018). For example, a method developed by Pickrell et al. incorporated 450 different annotations into GWAS analysis of 18 human traits, and showed the number of loci with high confidence associations was increased by around 5% (Pickrell, 2014); our group has also previously developed and implemented a Bayesian method for gene-based association tests called Gene-Wide Significance (GWiS), which aggregates statistical associations of multiple independent variants within a gene for a single phenotype (Huang et al., 2011; Chanda et al., 2013); the method was proved to have greater power than univariate tests as well as other gene-based methods including L1/LASSO and summing the effects over all variants(Wu et al., 2009; Liu et al., 2010). Therefore in this work, we first follow the same idea which GWiS was developed based on, and tried to expand the Bayesian framework to test on multiple phenotypes on the whole genome scale, in order to detect pleiotropic associations genome-wide. Then, simulations were performed to analytically investigate into the dependency of power gain by using such Bayesian approaches on different factors, including prior strength and population size.

Even after correction for trait-trait correlations, many genetic variants are associated with multiple phenotypes, which motivates systematic approaches to identify these variants. While there are no pleiotropic association methods in general use, there have been three general directions for methods development: 1. for small collection of highly correlated phenotypes, a reasonable approach is to aggregate associations over all phenotypes and get a pooled z-score (Zhu et al., 2015); 2. when phenotypes are highly correlated and possibly redundant, a second direction is to use orthogonalization methods, like principal component analysis or singular value decomposition (Klei et al., 2008; O'Reilly et al., 2012); 3. regularization methods like L1/LASSO that favor selection of sparser sets of features have also been tried, despite the difficulty to scale on the genome level (Kim and Xing, 2009). While few Bayesian methods have been applied to the pleiotropic GWAS problem, such approaches have the potential to use observed patterns of pleiotropy to identify additional variants that follow the same pattern; in other words, the observed patterns would provide prior probabilities that could boost the confidence that other variants with the same pattern are true associations, even if there univariate p-values do not reach conventional genome-wide significance thresholds.

Therefore, we introduce this pleiotropic GWAS method based on Bayesian framework, SNP And Pleiotropic PHenotype Organization (SAPPHO), which learns pleiotropy using identified associations to discover additional associations with shared patterns. The detailed approaches for method development is described in Chapter 2. SAPPHO, along with other state of art approaches for pleiotropic association tests were tested using simulations over a range of scenarios with different association architectures and phenotypic correlations; scenarios where SAPPHO performed the best and worst were explored, and possible explanations were given. Then, SAPPHO was further assessed with two sets of real data: 1. Atherosclerosis Risk in Communities (ARIC) study of 8,000 individuals, whose gold-standard associations were provided by meta-analysis of 40,000 to 100,000 individuals from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium; 2. Cancer phenotypes from UK Biobank project, consisting several hundred to 15,000 individuals, with gold-standard obtained from GWAS catalog. It was shown that for both data sets, SAPPHO was able to detect additional associations that were not detected with the conventional univariate test, and boost power when different variants follow the same association patterns.

Despite the great efforts put into the method development using Bayesian approaches, performance of these methods wasn't as great as they were expected to be. For application to ARIC ECG phenotypes, GWiS only detected one additional hit compared to univariate test; while the pleiotropic method SAPPHO detected 5 more additional loci associated with the three ECG traits, additional pleiotropic effect was detected for only one loci compared with univariate test (Huang et al., 2011; Zhan et al., 2018). For Bayesian methods developed by other groups, it was also observed the average gain of power is on the range of 5% to 10%(Sveinbjornsson et al., 2016; Pickrell, 2014). Therefore, an analytical study was conducted to demonstrate why Bayesian approaches wasn't very successful in beating unbiased univariate tests. Given that Bayesian approaches usually incorporates prior probabilities that bias significance for favored sets thought to be enriched for significant findings, the simulation was done by calculating test power assuming perfect knowledge of a prior distribution, and then derive the population size increase required to provided the same boost without a prior. Two types of priors were considered: 1. hard prior, where only a fraction of total hypotheses are tested; 2. soft prior, for which all variants are divided into two classes, and a higher prior value is given to the favored class which is believed to be enriched with true associations. It is proved that for the case of hard prior, population size is exponentially more important than prior; and for soft prior, large power gain could only be obtained under very specific circumstances, and on average the power gain is only around 5-10%. These findings provide a rigorous explanation for the observed avoidance of prior-based methods, and support for the continued use of simple, robust univariate test rather than more sophisticated approaches.

Chapter 2

Discovering patterns of pleiotropy in genome-wide association studies

CHAPTER 2. DISCOVERING PATTERNS OF PLEIOTROPY IN GENOME-WIDE ASSOCIATION STUDIES

Abstract

Genome-wide association studies have had great success in identifying human genetic variants associated with disease, disease risk factors, and other biomedical phenotypes. Many variants are associated with multiple traits, even after correction for trait-trait correlation. Discovering subsets of variants associated with a shared subset of phenotypes could help reveal disease mechanisms, suggest new therapeutic options, and increase the power to detect additional variants with similar pattern of associations. Here we introduce two methods based on a Bayesian framework, SNP And Pleiotropic PHenotype Organization (SAPPHO), one modeling independent phenotypes (SAPPHO-I) and the other incorporating a full phenotype covariance structure (SAPPHO-C). These two methods learn patterns of pleiotropy from genotype and phenotype data, using identified associations to discover additional associations with shared patterns. The SAPPHO methods, along with other recent approaches for pleiotropic association tests, were first assessed by performing simulations over a range of scenarios and different genetic architectures. It is observed that SAPPHO methods perform the best when the true association patterns are simple, namely when association is between one single phenotype and each active variant, or when

all phenotypes are associated with each active variants. For scenarios where several but not all phenotypes are associated with the active variants, by comparing results obtained from greedy forward heuristic search and feeding SAPPHO with true association model, it was suggested that more sophisticated searching approach might have to be adopted for SAPPHO to obtain better performance.

SAPPHO was further assessed with data from the Atherosclerotic Risk in Communities (ARIC) study of 8,000 individuals on electrocardiogram(ECG) phenotypes, whose gold-standard associations were provided by meta-analysis of 40,000 to 100,000 individuals from the CHARGE consortium. Using power to detect gold-standard associations at genome-wide significance (0.05 family-wise error rate) as a metric, SAPPHO performed best. The SAPPHO methods were also uniquely able to select the most significant variants in a parsimonious model, excluding other less likely variants within a linkage disequilibrium block. For meta-analysis, the SAPPHO methods implement summary modes that use sufficient statistics rather than full phenotype and genotype data. Meta-analysis applied to CHARGE detected 16 additional associations to the gold-standard loci, as well as 124 novel loci, at 0.05 false discovery rate. With SAPPHO we were able to learn genetic structures that were hidden using the traditional univariate tests.

With SAPPHO's outstanding performance in both assessments using simulated and ECG data, we further applied it on cancer phenotypes from UK Biobank (UKB) project. The UKB project consists of 500,000 individuals, with 99% of the participants genotyped. 13 cancer phenotypes were tested with SAPPHO in summary mode, and the results were compared with univariate test and genome-wide associates from GWAS catalog. With number of cases for each cancer phenotype ranging from 300 to 14,647, univariate test detected 51 hits on the loci level, at 5×10^{-8} genome-wide significance; at 0.05 FDR, SAPPHO run in summary mode detected 59 loci, which includes all 51 loci detected by univariate test and 8 novel loci. 4 loci were pleiotropic, which were detected by both univariate and SAPPHO; out of the 4 loci, SAPPHO detected additional associations for 2 of them. All detected loci were compared with associations from GWAS catalog to gain further insights.

2.1 Introduction

Genome-wide association studies (GWAS) have had remarkable success in identifying genetic variants responsible for human disease, disease risk factors, and other biomedical phenotypes. To date, more than 17607 variants, primarily single nucleotide polymorphisms (SNPs), have been associated at genome-wide significance with at least 785 distinct traits, according to the GWAS catalog (Welter et al., 2014). Many variants are pleiotropic, with significant associations with multiple traits (Fig. 2.1). Observations of pleiotropy motivate systematic approaches to identify pleiotropic variants. Such approaches could use observed patterns of pleiotropy to identify additional variants that follow the same pattern. In a Bayesian setting, the observed patterns would provide prior probabilities that could boost the confidence that other variants with the same pattern are true associations, even if their univariate p-values do not reach conventional genome-wide significance thresholds. A second valuable application could be to use pleiotropic associations to infer mechanisms shared by multiple diseases, which could lead to new therapeutic approaches including drug repurposing.

A challenge is that we do not know in general which traits share causal genetic factors. While these pleiotropic patterns may be discovered from genome-wide association study data, re-use of the same data for pattern discovery and association discovery requires new methods to control false discovery rates. A second important challenge is to develop methods that provide as direct a route as possible to the most significant variant within an association locus. Methods that produce parsimonious models, selecting just the most significant variant and excluding the neighboring linked variants, have great value. A third challenge is to incorporate the phenotype-phenotype covariance structure in the analysis to discriminate between a model in which a variant affects two phenotypes directly and an alternative model in which a variant directly affects one phenotype, which in turn affects a second correlated phenotype.

While there are no pleiotropic association methods in general use, there have been three general directions for methods development. First, for small collections of highly correlated phenotypes, a reasonable approach is to aggregate associations over all the phenotypes. This approach has greatest power when the true model is that a variant affects each phenotype. A recent report demonstrated good power for phenotypes related to hypertension (Zhu et al., 2015). This approach is similar in motivation to gene-based methods for GWAS signal aggregation, including VEGAS for common variants (Liu et al., 2010) and SKAT for rare variants (Wu et al., 2011).

When phenotypes are highly correlated and possibly redundant, a second direction has been to use orthogonalization methods, usually principal component analysis or singular value decomposition, to identify a rank-reduced set of linear combinations. A representative method is principal components of heritability (PCH), which generates linear combination of phenotypes with highest heritability for each genetic variant (Klei et al., 2008). A drawback of this approach is that validating an association with a linear combination of phenotypes is more difficult than validating an association with a single phenotype, particularly when different studies assess different sets of phenotypes. An alternative approach is to use a linear combination of phenotypes as a feature to predict the variant genotype, reversing the typical direction of regression. The MultiPhen method uses ordinal regression to perform this type of test, with increased power for variants affecting multiple phenotypes (O'Reilly et al., 2012). While methods such as canonical correlation analysis (CCA) and MANOVA that assume that genotypes follow a normal distribution have inflated type-I error, MultiPhen produces no such inflation when parametric p-values are used.

A third approach has been to adapt methods like L1/LASSO (Wu et al., 2009) that favor selection of sparser sets of features (Kim and Xing, 2009). Despite the promise of this approach, it has not been widely used, possibly because the L1 regularization is still too weak to reject variants introduced through correlation only and not causation and because of computational costs for genome-wide applications.

The approach pursued here is to exploit observed association patterns to identify additional variants following the same pattern. These patterns are biclusters, with subsets of variants associated with subsets of phenotypes. Biclustering has been a productive approach for identifying block structure in gene expression data (Pontes et al., 2015). Biclustering is not directly applicable to GWAS data, however, because blocks of SNPs identified by naïve application of standard biclustering algorithms would be dominated by non-causal variants in linkage disequilibrium or haplotype blocks with a single causal variant.

We report results for a new Bayesian framework for genome-wide association studies of multiple phenotypes with shared genetics: SNP And Pleiotropic PHenotype Organization, SAPPHO. The SAPPHO method is motivated by our previous work developing a Bayesian method for gene-based association tests, Gene-Wide Significance (GWiS) (Huang et al., 2011; Chanda et al., 2013), which aggregates statistical associations of multiple independent variants within a gene for a single phenotype. Each identified variant within a gene effectively updates the prior probability that additional variants within the same gene are also associated, permitting successive identification of variants with smaller effects that could be missed by conventional univariate tests of individual SNPs. Using an assessment with real data and gold standards from meta-analysis, GWiS was found to have greater power than univariate tests and also greater power than other gene-based methods, including methods based on summing the effects over all variants (Liu et al., 2010) and methods using L1/LASSO (Wu et al., 2009). GWiS had robust performance across different genetic architectures, including the number of true effects per gene and the minor allele frequencies. The robust performance was in part due to a lack of tuning parameters. Instead, most parameters in the GWiS model were treated as nuisance parameters and removed by integration.

The SAPPHO method uses a similar approach to associate individual variants with multiple phenotypes in a single genome-wide model for testing T total SNPs for association with P total phenotypes. Model priors interpolate between two probability distributions for genetic architecture, one which each of the $T \times P$ possible SNP-phenotype associations is independent, and a second in which each of the 2^P possible patterns of association has its own prior probability. All of the remaining association structure parameters are integrated out, yielding a method with only a single adjustable parameter, the mixing fraction of the two priors. This parameter is essentially the threshold for the weakest possible variant-phenotype association that can be entered into a regression model. Identifying the most likely model is NP-hard, and heuristics are needed for an acceptable runtime. SAPPHO uses a greedy forward approach to identify a local optimum with an algorithm that scales linearly with the number of SNP-phenotype associations identified in the data.

We evaluate our proposed method through simulation and analysis of cardiovascular electrocardiogram (ECG) phenotypes. For ECG phenotypes, meta-analyses of studies with 40,000 to 100,000 individuals have been conducted with overlapping sets of variants associated with PR, QRS, and QT intervals. Notwithstanding concerns about missing heritability (Silva et al., 2015), the fraction of heritability explained by genome-wide significant associations for these traits ranges from 4% to 17% (Pfeufer et al., 2010; Sotoodehnia et al., 2010; Pfeufer et al., 2009; Arking et al., 2014). The genome-wide significant findings provided by meta-analysis provide gold-standard true positive associations for assessing the power of different methods. The SAPPHO methods have the further potential to provide new biomedical knowledge by revealing classes of variants that contribute to distinct subsets of ECG parameters. Given the outstanding performance of SAPPHO on both simulations and ECG phenotypes, SAPPHO was further applied to UK Biobank project to try to explore pleiotropic patterns on cancer phenotypes.

The Methods section provides a mathematical description of SAPPHO and the algorithm used to identify an optimum in the space of all possible models. Briefer summaries of other approaches are provided, together with summaries of real and simulated data used for assessment. Then follows section reports on the assessment results and the pleiotropic patterns observed for simulations representing a range of scenarios of phenotypes that share genetic and environmental factors and also for ECG traits. The Discussion concludes with an interpretation of the benefits and drawbacks of different pleiotropy methods and a vision for possible future directions to discover and exploit pleiotropy in human genetic association studies.

2.2 Methods

2.2.1 Genetic model

SAPPHO has been developed for quantitative phenotypes. Case/control or other dichotomous phenotypes can be represented as 0/1 encodings, which general retain high power when causal variants have small effects (Chanda et al., 2013). Similarly, rank-ordered categories can be represented as corresponding integers. Unranked categories can in principle be represented as 0/1 indicators for each category; in practice, these are less common than quantitative, graded, or dichotomous phenotypes. Extensions to dichotomous phenotypes or general linear models in the exponential family are possible but more computationally intensive without a corresponding gain in power (Chanda et al., 2013).

The SAPPHO statistical model considers a population of N unrelated individuals and P distinct phenotypes, with each individual assessed for each phenotype. The phenotype data is represented as a real-valued phenotype matrix $\mathbf{Y}$ with N rows and P columns. Individuals are also genotyped at distinct loci corresponding to T total independent tests, represented as a real-valued genotype matrix $\mathbf{X}$ with N rows and T columns. In most applications, the genotype values will correspond to allele frequencies or dosages for bi-allelic single nucleotide polymorphisms (SNPs), measured directly or imputed. All data elements of $\mathbf{Y}$ and $\mathbf{X}$ are assumed present. In practice, some individuals will lack data for some genotypes and phenotypes. In this work, for simplicity only individuals with complete data are retained. Exclusion could be done at the level of individual means, variances, and covariances of phenotypes and genotypes, which in theory leads to non-positive-definite covariance matrices but in practice usually does not cause numerical instabilities (Chanda et al., 2013).

An association model, denoted M, specifies the direct effects of variants on phenotypes. For this work, we restrict attention to linear models. Thus, a model specifies which elements of a regression coefficient matrix $\boldsymbol{\beta}$ with T rows and P columns may be non-zero; the number of non-zero elements is denoted $|M|$. The model does not specify the corresponding values; these are treated as nuisance parameters that are integrated out. We consider two different models representing alternative assumptions about the phenotype covariance matrix: SAPPHO-I models each phenotype as independent given the genetic effects; SAPPHO-C models the complete phenotype-phenotype covariance structure. Given the model M and the genotypes $\mathbf{X}$ for an individual, the probability distribution for phenotypes $\mathbf{Y}$ is multivariate normal with covariance matrix $\mathbf{\Omega}$, diagonal for SAPPHO-I and including off-diagonal elements for SAPPHO-C,

$$\begin{aligned}\Pr(\mathbf{Y}|\mathbf{X}, M) = &\int_{\boldsymbol{\beta}} \Pr(\boldsymbol{\beta}) d\boldsymbol{\beta} \int_{\mathbf{\Omega}} \Pr(\mathbf{\Omega}) d\mathbf{\Omega} \times \\ &\prod_{i=1}^{N} (2\pi)^{-P/2} |\mathbf{\Omega}|^{-1/2} \times \\ &\exp[-(1/2)(\mathbf{y_i} - \mathbf{x_i}\boldsymbol{\beta})^{+} \mathbf{\Omega}^{-1} (\mathbf{y_i} - \mathbf{x_i}\boldsymbol{\beta})] \end{aligned} \tag{2.2.1}$$

where $\mathbf{x_i}$ and $\mathbf{y_i}$ are genotype and phenotype vectors for individual i, and the superscript $^+$ denotes transpose. The integral over $\boldsymbol{\beta}$ is over the $|M|$ non-zero elements, and the $\boldsymbol{\beta}$ and $\mathbf{\Omega}$ integrals include formal normalization factors $\Pr(\boldsymbol{\beta})$ and $\Pr(\mathbf{\Omega})$. The notation $|\mathbf{\Omega}|$ denotes the determinant of the phenotype covariance matrix $\mathbf{\Omega}$. This covariance matrix does not include the genetic contributions; the observed covariance matrix $\mathbf{V}$ is

$$\mathbf{V} = \mathbf{\Omega} + (N-1)^{-1} \boldsymbol{\beta}^{+} \mathbf{X}^{+} \cdot [I - N^{-1} \mathbf{1}\mathbf{1}^{+}] \cdot \mathbf{X}\boldsymbol{\beta}. \tag{2.2.2}$$

The normalization factors $\Pr(\boldsymbol{\beta})$ and $\Pr(\boldsymbol{\Omega})$ formally depend on meta-parameters for regularization. In practice, we use the asymptotic limit that excludes the contribution of the meta-parameters, as we did with GWiS (Huang et al., 2011), keeping terms of order $\ln N$ and greater. Performing steepest descents around the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Omega}}$, equivalent to the Bayesian Information Criterion or BIC (Schwarz, 1978), the asymptotic limit for the log-probability of the observed phenotypes given genotypes and model is

$$\begin{aligned} \ln \Pr(\mathbf{Y}|\mathbf{X}, M) &\sim \ln \Pr(\mathbf{Y}|\mathbf{X}, \widehat{\beta}, \widehat{\boldsymbol{\Omega}}) - (|M|/2) \ln N \\ &= -(1/2) \ln[(2\pi)^P |\widehat{\boldsymbol{\Omega}}|] - (N/2) \\ &\quad -(|M|/2) \ln N. \end{aligned} \tag{2.2.3}$$

For SAPPHO-I, we assume that the phenotypes are independent with each other; this is essentially equivalent to setting the non-diagonal elements of $\boldsymbol{\Omega}$ equal to zero, leaving only the non-diagonal σ_j^2s that correspond to the residual variance of each phenotype; in other words, now $|\boldsymbol{\Omega}| = \prod_{j=1}^{P} \sigma_j^2$, where σ_j^2 is the residual variance of each phenotype. The probability distribution for phenotypes $\mathbf{Y}$ becomes a product of P normal distributions, with the number of distributions equal to the number of phenotypes:

$$\begin{aligned}\Pr(\mathbf{Y}|\mathbf{X}, M) = &\int_{\boldsymbol{\beta}} \Pr(\boldsymbol{\beta}) d\boldsymbol{\beta} \int_{\boldsymbol{\Omega}} \Pr(\boldsymbol{\Omega}) d\boldsymbol{\Omega} \times \\ &\prod_{i=1}^{N} \prod_{j=1}^{P} (2\pi\sigma_j^2)^{-1/2} \times \\ &\exp[-(2\sigma_j^2)^{-1}(y_{ij} - \mathbf{x_i}\boldsymbol{\beta}_j)^{+}(y_{ij} - \mathbf{x_i}\boldsymbol{\beta}_j)], \qquad (2.2.4)\end{aligned}$$

where y_{ij} is the jth element of vector $\mathbf{y_i}$, and $\boldsymbol{\beta}_j$ denotes the jth column of matrix $\boldsymbol{\beta}$. Adopting the same asymptotic approximation for the log-probability using BIC yields

$$\begin{aligned}\ln \Pr(\mathbf{Y}|\mathbf{X}, M) &\sim \ln \Pr(\mathbf{Y}|\mathbf{X}, \widehat{\beta}, \widehat{\boldsymbol{\Omega}}) - (|M|/2) \ln N \\ &= -\frac{1}{2} \sum_{j=1}^{P} \ln(2\pi \widehat{\sigma_j}^2) - (N/2) \\ &\quad -(|M|/2) \ln N. \qquad (2.2.5)\end{aligned}$$

2.2.2 Model prior

The model prior probability, $\Pr(M)$, is represented as the product of two terms,

$$\Pr(M) \propto \Pr M_1^{\gamma} \Pr M_2^{(1-\gamma)}. \qquad (2.2.6)$$

This form corresponds to linear interpolation on a logarithmic scale,

$$\ln \Pr(M) = \gamma \ln \Pr M_1 + (1-\gamma) \ln \Pr M_2 + \text{constant}, \tag{2.2.7}$$

where $\gamma \in [0,1]$ and the constant term is model-independent and may be ignored. The prior $\Pr M_1$ penalizes each genotype-phenotype association individually, while the prior $\Pr M_2$ penalizes based on each different association pattern, as described below.

The prior $\Pr M_1$ models each possible SNP-phenotype pair as a binary random variable reflecting association with probability θ or no association with probability $1-\theta$ for $\theta \in [0,1]$; the single parameter θ is shared by each of the $T \times P$ possible genotype-phenotype associations. For any model with $|M| = K$ total associations, $\Pr M_1$ is

$$\Pr M_1 = \int_0^1 d\theta P(\theta) \theta^K (1-\theta)^{(TP-K)}, \tag{2.2.8}$$

where $P(\theta)$ is a possible prior on θ; we use the uniform prior $P(\theta) = 1$. While P varies with the number of phenotypes in different studies, T is set equal to the conventional number of independent effects in the genome for human GWAS, 10^6. The nuisance parameter θ is integrated out to yield the standard result,

$$\Pr M_1 = \frac{K!(TP-K)!}{(TP+1)!} = \text{Beta}(K+1, TP-K+1), \tag{2.2.9}$$

where '!' denotes the factorial function and 'Beta' is the standard Beta function extending the combinatorial factor to non-integer arguments.

The derivation of $\Pr M_2$ is similar to $\Pr M_1$, except that it considers patterns rather than individual associations. A pattern α is one of the $2^P - 1$ possible subsets of phenotypes, excluding the null pattern of no associations. The probability that a SNP belongs to pattern α is denoted θ_α, with $\theta_\alpha \in [0, 1]$ and $\sum_\alpha \theta_\alpha = 1$ defining a multinomial probability distribution. Denoting n_α as the total number of variants with pattern α, and $\sum_\alpha n_\alpha = n$, n being the total number of associated SNPs, the probability for a particular model M is

$$\Pr M_2 = (2^P - 1)! \int_{\{\theta\}} d\{\theta\} P[\{\theta_\alpha\}] \prod_\alpha \theta_\alpha^{n_\alpha}. \tag{2.2.10}$$

The integral is over all possible feasible parameters and $P[\{\theta_\alpha\}]$ is a possible prior distribution; we use the uniform distribution $P[\{\theta_\alpha\}] = 1$. The term $(2^P - 2)!$ is the standard normalization factor for a multinomial distribution. The nuisance parameters θ_α are removed by integration, yielding the standard result,

$$\Pr M_2 = \frac{(2^P - 1)!}{(n + 2^P - 1)!} \prod_{\alpha=1}^{2^P - 1} n_\alpha!. \tag{2.2.11}$$

Note that for $\Pr M_2$ only the occupied patterns contribute to the model probability, similar to latent Dirichlet allocation (LDA) in which only occupied states contribute (Blei et al., 2003). This probability model favors pleiotropy models in which variants share the same association patterns. The overall prefactor $(2^P - 1)!/(n + 2^P - 1)!$ is identical for all models and independent of the occupation numbers $\{n_\alpha\}$ for different patterns. Therefore, for the purpose of computational efficiency, we use

$$\mathrm{Pr} M_2 \propto \frac{\prod_{\alpha=1}^{2^P-1} n_\alpha!}{n!}. \tag{2.2.12}$$

2.2.3 Model score

The goal of SAPPHO is to identify the most likely model, $\widehat{M}$, defined as $\widehat{M} = \arg\max_M \mathrm{Pr}(M|\mathbf{Y}, \mathbf{X})$. The posterior probability of a model is defined by Bayes rule as

$$\begin{aligned} \mathrm{Pr}(M|\mathbf{Y}, \mathbf{X}) &= \mathrm{Pr}(\mathbf{Y}, \mathbf{X}, M)/\mathrm{Pr}(\mathbf{Y}, \mathbf{X}) \\ &= \mathrm{Pr}(\mathbf{Y}|\mathbf{X}, M)\mathrm{Pr}(\mathbf{X}, M)/\mathrm{Pr}(\mathbf{Y}, \mathbf{X}). \end{aligned} \tag{2.2.13}$$

We make the standard assumption that the model M is independent of the genotype data, $\mathrm{Pr}(\mathbf{X}, M) = \mathrm{Pr}(\mathbf{X})\mathrm{Pr}(M)$, giving

$$\mathrm{Pr}(M|\mathbf{Y}, \mathbf{X}) = \frac{\mathrm{Pr}(\mathbf{Y}|\mathbf{X}, M)\mathrm{Pr}(M)}{\mathrm{Pr}(\mathbf{Y}|\mathbf{X})}. \tag{2.2.14}$$

The conditional probability $\mathrm{Pr}(\mathbf{Y}|\mathbf{X})$ is independent of model and need not be calculated. Similarly, to avoid numeric overflow and underflow, model posterior probabilities are always calculated as log-likelihood ratios relative to the null model, defined as the model score S_M,

$$\begin{aligned} S_M &\equiv \ln \frac{\Pr(\mathbf{Y}|\mathbf{X}, M)\Pr(M)}{\Pr(\mathbf{Y}|\mathbf{X}, M_\emptyset)\Pr(M_\emptyset)} \\ &= (1/2)\ln(|V|/|\Omega|) - (|M|/2)\ln N \\ &\quad +\gamma \ln \mathrm{Beta}(K+1, TP-K+1) \\ &\quad +(1-\gamma)\sum_\alpha \ln\Gamma(n_\alpha+1) \\ &\quad -(1-\gamma)\ln\Gamma(\sum_\alpha n_\alpha + 1) \end{aligned} \tag{2.2.15}$$

In practice, all models, including the null model, typically include constant terms for phenotype mean and covariates that represent relevant clinical variables, including sex, age, height, weight, and body mass index, and possible additional covariates that describe population structures. Regression coefficients for these covariates are calculated in parallel with regression coefficients for genetic variants, but they make no net contribution when models are compared. For computational efficiency, SAPPHO regresses out the known covariates and then operates on the residuals.

The parameter γ is the single adjustable parameter in the SAPPHO method. While it could be set using cross-validation, this would require a gold-positive training set and depends on the genetic architecture. An architecture with no shared genetic factors would favor $\Pr M_1$, whereas an architecture with a small number of strong patterns would favor $\Pr M_2$. Instead, we relate γ to the effect size required to enter a new SNP-phenotype association into a model. To be more specific, we take the dominant term from the Beta penalty, together with the BIC penalty, giving the χ^2 threshold for adding a single association to the model,

$$\chi^2 = \ln(|V|/|\Omega|) \tag{2.2.16}$$

The value of γ is then calculated accordingly,

$$\gamma = \frac{\chi^2 - \ln N}{2\ln(TP)} \tag{2.2.17}$$

Behavior of SAPPHO under different values of this tuning parameter is discussed in the following paragraphs. In general, with real data, we found that setting $\gamma \ln(TP) = \ln(10^3)$ is a good value to control for type I and II error rate; with simulation, different tuning parameter will lead to different behaviors of SAPPHO, favoring different true underlying real association patterns.

2.2.4 Model search and variant ranking

Identifying $\widehat{M}$ is NP-hard and is not attempted directly. Instead, a greedy forward approach is employed. Given a current model, all models that may be reached by adding a single genetic association are considered. There are two possible cases. In one case, a variant with no associations gains an association to a single phenotype. In the second case, a variant associated with a subset of phenotypes gains an association with one additional phenotype. With T total variants and P total phenotypes, this procedure requires calculating the posterior probability for approximately $T \times P$ possible models. The model with the greatest increase in posterior probability is selected, and the procedure continues until any additional association decreases the posterior probability. The resulting model, locally optimal with respect to adding associations, is termed $\widetilde{M}$, distinct from the global optimum $\widehat{M}$. In principle, stepwise forward-backward selection would also be possible, but would require full matrix inverses that would vastly increase the computational cost. Backward steps removing a variant and all of its associations are used at the very end of the model search, however, for variant ranking.

For SAPPHO-I, computation is made more efficient by using successive orthogonalization rather than matrix inverses at each step to obtain the new maximum likelihood estimates for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Omega}}$.

To further speed the model search by SAPPHO-I and SAPPHO-C, the candidate list of variants is pre-filtered, as is common with with other approaches that build multivariate genome-scale models. A preliminary univariate test is conducted for each of the T total variants with each of the P total phenotypes, yielding $T \times P$ total p-values. Each variant is then assigned its minimum p-value across the P phenotypes, and variants are sorted from smallest to largest p-value. Candidate variants are taken from this list in increasing p-value order, with variants excluded if they are in strong linkage disequilibrium ($r^2 \geq 0.8$) with a better-ranked variant that has already been selected. Selection ends when the p-value of the best remaining variant is above a threshold. We used 1×10^{-4} as a threshold and found no differences in model selection for a more lenient and more computationally expensive threshold of 1×10^{-3}.

2.2.5 Test statistic and significance thresholds

Starting from the full final model, $\widetilde{M}$, a test statistic for each variant v in the model is obtained by removing that single variant from $\widetilde{M}$ to obtain a new model $\widetilde{M}_{/v}$. All of the variant's associations are removed; thus, if $\widetilde{M}$ has $|M|$ non-zero $\boldsymbol{\beta}$ parameters and a variant v in the model has associations with P_v phenotypes, the new model $\widetilde{M}_{/v}$ has $|M| - P_v$ non-zero $\boldsymbol{\beta}$ parameters. The score S_v for a particular variant v is calculated as expected,

$$S_v = \ln \Pr(\widetilde{M}|\mathbf{Y}, \mathbf{X}) - \ln \Pr(\widetilde{M}_{/v}|\mathbf{Y}, \mathbf{X}), \tag{2.2.18}$$

and serves as the test statistic for that variant. As mentioned above, the model $\widetilde{M}$ is optimal with respect to forward selection but not necessarily to backward selection. While the search biases S_v to positive values, negative values are possible and are observed, albeit infrequently.

Genome-wide randomizations were used to calibrate the value of S_v, controlling for either family-wise error rates (FWER), with FWER = 0.05 corresponding to genome-wide significance, or false discovery rate (FDR), controlling for FDR = 0.05 as the threshold. Thresholds for FDR require fewer permutations for estimation; FDR was therefore used for simulated data and CHARGE data, whereas FWER was used for analysis of ARIC data. Permutations of the original data were generated starting with the vector of genotypes and the vector of phenotypes for each individual, with covariates already regressed out. Permutations were performed by randomizing the pairing between a genotype vector and its phenotype vector. Elements within individual genotype and phenotype vectors were not permuted. These permutations maintain the genetic covariance and phenotype covariance structure of the data.

Filtering steps independent of genotype-phenotype pairing, primarily filtering based on allele frequency and Hardy-Weinberg equilibrium, were identical for all permutations. Subsequent processing of permuted data sets exactly matched processing of the original data, including the computationally expensive step of performing all the genome-wide univariate tests. For FWER, the null distribution of the test statistic was obtained from 100 genome-wide permutations. The score of the best variant was retained for the 100 permutations, and the 5th-best score was used to define the 0.05 FWER threshold. For FDR, 10 permutations were done in the same way. The expected number of false discoveries $\widehat{F}(S^\star)$ with scores greater than or equal to threshold $S^\star$ was estimated as

$$\widehat{F}(S^\star) = B^{-1} \sum_{b=1}^{B} \sum_{v} \Theta[S_v(b) \geq S^\star] \times P_{bv}, \tag{2.2.19}$$

where $B = 10$ is the number of permutations, $S_v(b)$ is the score of variant v in permutation b, $\Theta(u) = 1$ if logical argument u is true and 0 if false, and P_{bv} is the number of phenotypes associated with variant v in permutation b. The FDR for the unpermuted data for a given threshold was then calculated as

$$\mathrm{FDR}(S^\star) = \frac{\widehat{F}(S^\star)}{\sum_v \Theta[S_v \geq S^\star] \times P_v}. \tag{2.2.20}$$

Variants from the unpermuted data were arranged in decreasing order by score; the first SNP v' for which $\mathrm{FDR}(S_{v'}) > 0.05$ was identified; and the previous SNPs with $S_v > S_{v'}$ were retained as the predicted positives at FDR 0.05. As is standard for methods based on model scores, calibration by permutation test is done separately for each data set analyzed.

2.2.6 Other methods

Assessments involved representative available implementations of major classes of pleiotropy methods. In general, adjustable parameters were set using published recommendations. We restricted attention to methods with run times that were small multiples of the cost of performing all $T \times P$ univariate tests of individual variants and phenotypes.

For aggregating association signals over a set of phenotypes, the methods SHom and SHet, using homogeneous and heterogeneous test statistics, were selected (Zhu et al., 2015). Given summary statistics of all $T \times P$ univariate tests (regression coefficients and standard deviations), SHom uses a generalized inverse-variance weighted analysis to combine individual tests into a pooled z-score, which also considers the correlation between phenotypes and sample sizes of different studies. By comparing the test statistic with a standard normal distribution, SHom obtains one p-value for each tested variant. SHet has a scoring function similar to SHom, but also introduces a threshold τ. The p-value for the test statistic combining all t-scores greater than τ is maximized, and the corresponding p-value is assigned to the given SNP. Because of this selection process, SHet does not follow a standard normal distribution, and p-values are obtained through permutation.

For tests involving linear combinations of phenotypes, the Principal Components of Heritability (PCH) method (Klei et al., 2008) and MultiPhen (O'Reilly et al., 2012) were selected. With PCH, a phenotype loading vector w is first estimated for each SNP. The loading vector is selected to maximize the variance of the loaded phenotypes explained by the given SNP using a subset of the data. Next, a t-score is obtained by regressing the genotype data onto the loaded phenotypes using the remainder of the data. By adopting a bagging technique and running cross-validation, the t-score distribution is estimated and a p-value is obtained.

For MultiPhen, each genotype is treated as a response variable outcome, and phenotypes are predictors. MultiPhen uses proportional odds logistic regression to regress genotype on the hyperplane constructed by the phenotypes, which models genotype data as ordinal (O'Reilly et al., 2012). No distributional assumptions are required for phenotypes, allowing MultiPhen to accommodate both binary and continuous measurements in a single framework. While ordinal regression for integer-valued allele dosages is theoretically attractive, the computational cost is much greater than for linear regression (Table 2.1). Furthermore, allele dosages estimated from imputation are real-valued rather than integer-valued. MultiPhen includes a gaussian kernel for this reason, and gaussian regression was used instead of ordinal regression for some of the results reported here. MultiPhen has two additional modes, variable selection and variable non-selection. For variable non-selection, all phenotypes are used as predictors for the genotype data; for variable selection mode, backward selection were performed on the phenotypes in order to exclude the phenotypes that were not associated with the SNP, and then the selected phenotypes were regressed on the genotype data. Non-selected phenotypes have nominal p-values of 1 as output.

We attempted to assess a regularized regression method, using an available implementation of the graphical fusion LASSO (Kim and Xing, 2009). The shrinkage behavior of graphical fusion LASSO method requires optimization of regularization parameters λ and γ by cross-validation over a search grid. The cross-validation steps were computationally intensive, and for many (λ, γ) grid values the iterations did not converge. No other implementations for GWAS were readily available. Published results based on graphical LASSO are generally for smaller data sets; the available implementation was originally developed for a data set of 34 SNPs and 543 individuals (Kim and Xing, 2009). We therefore excluded LASSO-based methods from the comparison.

Calibration of methods for 0.05 FWER or 0.05 FDR were conducted as for SAPPHO using the same set of genome-wide permutations. This calibration was conducted even for methods providing a nominal p-value based on assumed parametric distributions to ensure accurate benchmarking.

2.3 Assessment with simulated data

2.3.1 Methods

Methods were first run on simulated data to gain insight into differences in performance due to controlled aspects of genetic and environmental architecture of complex phenotypes. Simulations followed previous protocols used to assess GWiS and other gene-based tests (Huang et al., 2011). Simulated data sets included 1,000,000 independent SNPs for 10,000 individuals. Minor allele frequencies for each SNP were generated uniformly between 0.01 and 0.5 to model common variants. Three sets of simulations were done using frameworks denoted ‘genes only’, ‘genes and environment’, and ‘genes only with random phenotypes’. Each phenotype y with genotype vector $\mathbf{x}$ was simulated as

$$y = \mu + (\mathbf{x} - 2\mathbf{p}) \cdot \beta + \epsilon\sqrt{1 - \sigma_G^2}, \tag{2.3.1}$$

where μ is the overall phenotype mean, $\mathbf{p}$ is the vector of minor allele frequencies, β is the vector of regression coefficients for the scenario, ϵ is a unit normal random variable, and σ_G^2 is the genetic variance,

$$\sigma_G^2 = \sum_k 2p_k(1 - p_k)\beta_k^2$$

where p_k denotes the minor allele frequency of SNP k. The entry $\beta_k = 0$ if SNP k is not associated with the phenotype. For an associated SNP,

$$\beta_k = \sqrt{\frac{V_k}{2p_k(1 - p_k)}}.$$

The term V_k is chosen based population size N and specified type I and type II error rates α_1 and α_2 for univariate tests as

$$V_k = \frac{(z_1 - z_2)^2}{N}.$$

For a two-tailed test, $z_1 = \Phi^{-1}(1 - \alpha_1/2)$. The term z_2 is set by the desired type II error rate $\alpha_2 \equiv 1 - \text{power}$ as $z_2 = \Phi^{-1}(\alpha_2)$. The function Φ^{-1} is the inverse of the standard normal cumulative distribution with $\Phi(z) \equiv \int_{-\infty}^{z} du \exp(-u^2/2)/\sqrt{2\pi}$. With this setting, a SNP with regression coefficient β_k will have the specified power at threshold α. All simulations were performed 5 independent times, including 10 sets of permutations each to determine the 0.05 FDR threshold.

2.3.1.1 Genes-only simulations

In the ‘genes-only’ scenarios, phenotypic correlations were due entirely to shared genetic variants with no environmental effects. These simulations considered 6 phenotypes and 24 SNPs with causal effects. Simulations were performed separately for 3 scenarios reflecting increased sharing of genetic factors: independent, in which 4 SNPs were associated with each phenotype and no SNPs shared between phenotypes (24 total pairwise SNP-phenotype associations); block, in which the 6 phenotypes were divided into 2 blocks, and each block was associated with 12 SNPs (72 total SNP-phenotype associations); and full, in which the 24 SNPs contributed to each of the 6 phenotypes (144 total SNP-phenotype associations). Effects for all SNPs were set to have 50% power at 5×10^{-8} threshold.

2.3.1.2 Genes-and-environment simulations

For ‘genes-and-environment’ simulations, phenotypic correlations were due to both genetic and environmental effects. Two scenarios were simulated in this case: weak environment and strong environment. For weak environment, 4 phenotypes were partitioned into 2 blocks of 2 phenotypes; phenotypes within the same block are correlated through environmental effects, while phenotypes across different blocks are not; for strong effects, the 4 phenotypes are all correlated through environmental effects. For these scenarios, the random variable ϵ for each phenotype p (Eq.2.3.1) follows a multivariate normal with $\text{Var}(\epsilon_p) = 1$ and with covariance $\text{Cov}(\epsilon_p, \epsilon_p')$ determined by a predefined environmental covariance structure. For the weak environment simulation, the 4×4 covariance matrix is a block matrix of two matrices of size 2, and for strong environmental simulation, all elements of the 4×4 matrix are non-zero. For this study, we chose to set the correlation for different environmentally correlated phenotypes to 0.5. For weak environment, SNPs were associated with the phenotypes in 4 modes: different-block-same-effect, where each SNP is associated with two phenotypes, one from each block, and the effects are in the same direction; different-block-different-effect, where each SNP is associated with two phenotypes, one from each block, and the effects are in opposite directions; same-block-same-effect, where each SNP is associated with two phenotypes from the same block, and the effects are in the same direction; same-block-different-effect, where each SNP is associated with two phenotypes from the same block, and the effects are in opposite directions. For strong environment, SNPs were associated with the phenotypes in 2 modes: same effect, where SNPs are correlated with all phenotypes with effects of the same direction; different effect, where SNPs are correlated with first two phenotypes with positive effects, and the last two phenotypes with negative effects. Different directions of effect were represented by different signs for the regression coefficients, with magnitudes defined based on type I and type II error rates exactly as in the 'genes-only' simulation. All effects were simulated to have 50% power at 5×10^{-8} univariate test threshold. For weak environment, 24 SNPs were simulated and divided equally into 4 modes, with 6 SNPs in each mode, ending up with 48 total associations; for strong environment, 24 SNPs were simulated and divided equally into two modes, with 12 SNPs in each mode, ending with 96 total associations. With this set of simulation assessed the capability of different methods to detect SNPs whose association patterns have the same or opposite direction from the environmental association patterns.

2.3.1.3 Mixture of genetic and non-genetic phenotypes

For the ‘mixture of genetic and non-genetic phenotypes’ simulations, we attempted to generate scenarios similar to the real ARIC data with known association patterns. Therefore, 13 total phenotypes were simulated, and the simulations were done with three scenarios: ONE association, where all active SNPs were associated with phenotype 1; TWO associations, where all active SNPs were associated with phenotypes 1 and 2; and THREE associations, where all active SNPs were associated with phenotypes 1, 2, and 3. For each scenario, all active SNPs follow the same association pattern, with the number of associated phenotypes differing between the scenarios. The effect of each association was simulated such that each SNP has 50% power at 5×10^{-8} threshold, with the two associations and three associations effect calculated as follows using P to denote power:

$$P_{\text{univariate}} = 1 - (1 - P_{\text{TWO}})^2 = \frac{1}{2} \tag{2.3.2}$$

which gives us $P_{\mathrm{TWO}} = 0.293$.

$$P_{\mathrm{univariate}} = 1 - (1 - P_{\mathrm{THREE}})^3 = \frac{1}{2} \tag{2.3.3}$$

which gives us $P_{\mathrm{THREE}} = 0.206$. 24 SNPs were simulated to be associating variants in each scenario, yielding 24, 48, and 72 total associations.

2.3.2 Results

2.3.2.1 Genes-only simulations

Methods were assessed with phenotypes with shared genetic factors but without shared environmental contributions (Fig. 2.2). Three scenarios were considered, with increased sharing of genetic factors: independent, with 4 independent SNPs contributing to each of 6 phenotypes; block, with 12 SNPs contributing to one block of 3 phenotypes, and 12 other SNPs contributing to a second block of 3 phenotypes; and full, with 24 SNPs contributing to all 6 phenotypes. In general, all pleiotropy-based methods performed better with increasing shared genetic factors. Univariate tests were run and real-positives were determined with three methods: UNI where the 0.05 FDR threshold was used; UNILE with LE standing for loose-empirical where the standard 5×10^{-8} threshold were used for each association; UNISE with SE standing for stringent-empirical where the 5×10^{-8} threshold was corrected with the number of phenotypes tested, which in this case is 6.

The performance of SAPPHO, assessed as power to detect at genome-wide significance, out-performed all other methods for the independent scenario. Methods other than SAPPHO have lower power than univariate tests when phenotypes do not share causal variants. For the full scenario, pleiotropy methods other than PCH and UNISE had power close to 100%, making the methods difficult to distinguish on this basis. For the block scenario, the MNS (MultiPhen-non-select) and MS (MultiPhen-select) methods were somewhat better than other methods, outperforming SAPPHO-I, SAPPHO-C, and SHet by 1 to 2 hits in 2 out of 5 runs.

2.3.2.2 Genes-and-environment simulations

As described in the method section, for the genes-and-environment simulations, phenotypic correlations were set to be due to both genetic and environmental effects. Simulations were run in two modes: strong environment and weak environment.

For strong environment same-effect SNPs, most methods performed well except for SHet and SAPPHO-C, which were unable to detect most associations for this group of SNPs. SHom was able to detect all SNPs because the underlying association pattern was same as the simulated pattern, namely the variant is associated with all phenotypes. Although the underlying assumption for SAPPHO-I did not match the simulation pattern, it performed well because all the SNPs followed the same association pattern; thus the pattern prior gives its power to detect all the variants.

Unexpectedly, SAPPHO-C performed poorly in the same-effect setting. The reason is likely because the variance explained by one association is compensated through correlation with other phenotypes, with the result that adding a variant to the model does not improve the model score. To explore this effect further, we calculated the SAPPHO-C score for the true model and found it to have a large negative score. We can also explain this effect by noting that the BIC penalty assumes that regression coefficients for a SNP have independent sign, whereas the architectures in this scenario force the regression coefficients to have the same sign, resulting in a penalty that is too large.

For strong environment different-effect SNPs, all methods performed well, except for SHom, due to the difference in the simulated data and its underlying assumption.

For weak environment simulations, SNPs were simulated to follow 4 different patterns: different-block-same-effect(dbse), different-block-different-effect(dbde), same-block-same-effect(sbse), same-block-different-effect(sbde). Different/same-block denotes whether the environmental correlation is identical with the genetic pleiotropic effects, and same/different effect denotes whether the effects for the simulated pleiotropic SNPs are of the same direction or not. For dbse, SAPPHO-C and MultiPhen performed the best by finding all associated SNPs; followed by SAPPHO-I, SHet, and PCH; for dbde, SAPPHO-C and MultiPhen again performed the best. This is somewhat similar to the genetic only simulation, where the correlation between phenotypes are only through genetic effect, and SNPs are correlated with both blocks. For sbse, SAPPHO-I performed the best, while SAPPHO-C, SHet, and SHom performed poorly; this situation is similar to the same-effect SNPs for strong environment simulation, where a SNP is associated with all phenotypes in a block, with positive effects. Similarly, the performance on sbde was same as different-effect for strong environment simulation, with all methods performing well. For both simulations, the original version of MultiPhen using ordinal regression was again performed, yielding outcomes similar to using Gaussian regression, but with much longer running time.

2.3.2.3 Mixture of genetic and non-genetic phenotypes

We performed simulations with genetic effects in 3 out of 13 total phenotypes to further investigate the relative performance of SAPPHO and MultiPhen and, in cases where SAPPHO performed less well, whether the cause was the underlying statistical model or the greedy rather than full model search. These simulations used three scenarios labeled ONE, TWO, and THREE. In scenario ONE, all associations were with phenotype 1. In scenario TWO, all associations were pleiotropic with phenotype 1 and phenotype 2. In scenario THREE, all associations were pleiotropic with phenotypes 1, 2, and 3. Phenotypes 4-13 were random in all simulations, with no genetic component. Other methods performed less well (Supplemental Table 3).

As described previously (see model score in the Method section), SAPPHO has a single tuning parameter defined by the least significant univariate p-value that can be added to an association model. Because model scores are calibrated by permutation, this parameter does not affect stringency in terms of FDR or FWER. It can affect power, however, because a more stringent threshold will reject weaker true associations. It also affects computational cost because a looser threshold yields a longer list of candidate variants. We used thresholds 5×10^{-4}, 2.5×10^{-4}, 10^{-4}, 10^{-5}, 10^{-6} (Supplementary Table 4).

Using the loose threshold of 5×10^{-4}, SAPPHO-I outperformed MultiPhen for scenario ONE, performed equivalently for scenario TWO, and had slightly less power than MultiPhen for scenario THREE (Fig. 2.5). To determine whether the power disadvantage was due to the greedy search or to the statistical model, we also calculated the score of model defined by the true associations. In this case, SAPPHO performed better than MultiPhen. We conclude that SAPPHO's performance could be improved using a more sophisticated model search, for example considering considering single and double associations in a single step, or adding backward steps.

We also performed simulations using the more stringent threshold of 10^{-6} for adding associations. In scenario ONE, SAPPHO still out-performed MultiPhen. In scenarios TWO and THREE, SAPPHO performed worse, in large part because the threshold prevented true associations from being considered.

At 0.05 FDR, both SAPPHO and MultiPhen returned false positive results. SAPPHO false positives tend to be spurious associations of a variant with an individual phenotype. MultiPhen tends to over-predict associations: given a true association to a phenotype, MultiPhen often predicts additional false-positive associations for the same SNP with additional phenotypes. These results support the hypothesis on the real data result that the additional pleiotropic associations found by MultiPhen in the ARIC data are false positives rather than true associations, consistent with the lack of significance for these associations in the larger CHARGE data set.

2.4 Assessment with real data and pleiotropy on ECG phenotypes

2.4.1 Electrocardiogram phenotypes

Electrocardiogram (ECG) phenotypes are considered to be measurements that reflects conditions of heart. To be more specific: PR interval (from beginning of the P wave to the beginning of the QRS interval) reflects atrial and atrioventricular nodal conduction (Pfeufer et al., 2010); QRS interval (from beginning of the Q wave to the end of the S wave) reflects depolarization of both ventricles (Sotoodehnia et al., 2010); QT interval (from beginning of the Q wave to the end of the T wave) reflects myocardial re-polarization (Arking et al., 2006; Pfeufer et al., 2009; Newton-Cheh et al., 2009; Arking et al., 2014). These phenotypes are considered to be risk factors for atrial fibrillation, sudden death, heart failure, and etc., and therefore have been extensively investigated into (Pfeufer et al., 2010; Sotoodehnia et al., 2010; Pfeufer et al., 2009; Arking et al., 2006; Newton-Cheh et al., 2009; Arking et al., 2014). To date, research have shown that the fraction of heritability explained by genome-wide significant associations for these traits ranges from 4% to 17%, with 44, 22, and 35 independent genome-wide associations detected for PR, QRS, and QT intervals, respectively. Therefore, these ECG phenotypes serve as ideal traits for us to test the methods on.

2.4.2 Methods

SAPPHO, together with other methods under comparison, were tested using individual-level ECG phenotype and genotype data from the the Atherosclerotic Risk In Communities (ARIC) study cohort (ARICInvestigators, 1989). This cohort includes approximately 8000 Caucasian ethnicity subjects and 2000 African-American ethnicity subjects. Assessments here use only the Caucasian ethnicity because power has been insufficient for African-American ethnicity.

Known positive phenotype-genotype associations were taken from meta-analyses conducted by the CHARGE consortium, which includes ARIC as a cohort. Recent meta-analyses have included 88,000 individuals for PR, 40,407 for QRS, and approximately 100,000 for QT (Pfeufer et al., 2010; Sotoodehnia et al., 2010; Arking et al., 2014). Covariates for the EKG phenotypes were selected to be identical to those used in meta-analysis: height, age, gender, center, BMI, centerm, and heart rate.

Genotypes for ARIC were imputed by pre-phasing with Shapelt (v1.r532) and then imputing to 1000 Genomes (1000GenomesProjectConsortium et al., 2010) using IMPUTE2 (Howie et al., 2009, 2011). Measured SNPs used for imputation were restricted to MAF > 0.005, $> 95\%$ complete, HWE > 0.00001, resulting in 711,589 SNPs in the final set used for the imputation. Final imputations from IMPUTE2 used the reference panel 1000 Genomes haplotypes – Phase I intergrated variant set release (v3) in NCBI build 37 (hg19) in chunks of size 5Mb. All 1092 individuals were used for imputation from the reference panel.

Analyses were done focusing on the overlapping variants between ARIC and CHARGE cohorts, and we have tested that this overlap set of variants essentially includes all SNPs reported as significant by previous GWAS for ECG traits. Variants were removed if the ARIC or CHARGE genotypes violated Hardy-Weinberg equilibrium ($P < 0.00001$), were poorly imputed (Qual < 0.3), or if the minor alleles were too low frequency to have power (MAF < 0.01), corresponding to fewer than 160 copies of the minor allele. These criteria and filtering variants to require univariate p-value $\leq 10^{-4}$ and low LD with more significant variants (see Methods) resulted in 620 total variants for the PR, QRS, and QT phenotypes. This filtered list was used for each method.

Assessments were performed by defining variant-phenotype associations present in the meta-analyses at genome-wide significance (p-value $\leq 5\times10^{-8}$ for each phenotype) as known positives. Assessments are complicated by linkage disequilibrium within the genome, which can lead to genome-wide significant findings for multiple variants within a linkage disequilibrium block. These multiple variants often correspond to a single causal variant, and for purposes of assessment they were grouped into a single known positive.

We performed the grouping as follows. For each phenotype, we linked together pairs of genome-wide significant SNPs within 500 kbp of each other. We then identified the distance-based connected components defined by these pairwise links. Each connected component in principle could contain multiple independent causal effects. To determine whether independent effects were present, we provided the SNPs in each connected component as a single locus to GWiS (Huang et al., 2011). GWiS was then run separately on each connected component to select candidate SNPs representing independent effects.

In regions with strong association signals, these candidate sets may still contain more SNPs than independent effects. Furthermore, independent effects must be matched across phenotypes. We therefore used linkage disequilibrium as defined by r^2 correlation to identify a final set of independent effects. We introduced correlation-based links between pairs of SNPs with $r^2 \geq 0.05$ and identified the connected components defined by the correlation-based links. This resulted with 107 gold-standard connected components, each connected component corresponding to a single effective known positive with one or more phenotypes.

We also investigated the robustness of the gold-standard connected components with respect to the r^2 threshold of 0.05. For a threshold of $r^2 \geq 0.01$ the number of connected components was 90, and for $r^2 > 0.1$ the number of connected components was 112. At the lower threshold, multiple effects are merged into a single connected component, while at higher threshold, one single effect may be divided into multiple connected components. With $r^2 \geq 0.05$, the SCN5A-SCN10A locus was assigned 3 different association signals, while at $r^2 \geq 0.1$ the SCN5A-SCN10A locus had 5 independent effects, which bracket the estimates from existing literature (Pfeufer et al., 2010; Sotoodehnia et al., 2010; Pfeufer et al., 2009; Arking et al., 2014). While the details of performance of individual methods depend somewhat on the r^2 threshold, the relative performance of different methods is stable with reasonable choices for the clustering threshold. Therefore for the current study all results were reported based on $r^2 \geq 0.05$ (Supplementary table 1).

Methods differ in their treatment of LD blocks and their attempt to identify the subset of phenotypes associated with each variant. The SAPPHO-I and SAPPHO-C methods attempt to provide a parsimonious list of associations, with only one SNP in each significant LD block. Other methods report each SNP within an LD block as a positive. The SAPPHO-I, SAPPHO-C, MultiPhen-Selection, and univariate methods identify the subset of phenotypes associated with a SNP, whereas other methods do not. For purposes of assessment, we calculated the r^2 for each SNP selected by a method to each SNP in the gold standard correlation-based connected components. We defined $r^2 \geq 0.1$ as the threshold for matching. Each correlation-based connected component with at least 1 matching SNP was counted as a true positive; the remaining connected components were counted as false negatives. We tried different threshold including $r^2 \geq 0.1$, $r^2 \geq 0.5$, $r^2 \geq 0.7$, and $r^2 \geq 0.8$, and found that while using the thresholds other than $r^2 \geq 0.1$ did not substantially increase the number of true positives, they yielded many false positives, primarily non-causal variants somewhat correlated with real effects. We therefore used $r^2 \geq 0.1$ as the threshold for reporting results.

To be more favorable to the non-parsimonious methods, we used a similar grouping strategy to define the number of false positives. SNPs reported by a method but with $r^2 < 0.1$ to any SNP in a gold-standard connected component were grouped into false-positive connected components using $r^2 \geq 0.05$, and each false-positive connected component was then counted as a single false positive. For SAPPHO, MultiPhen-Selection and univariate tests, the methods which provide the subset of phenotypes associated with each variant, we performed subsequent analyses to assess the ability to detect the correct variant-phenotype associations.

In addition to performing assessments with the original ARIC data, we also performed assessments in which the ARIC data was augmented with with random phenotypes generated as independent and identically distributed standard normal random variables. We performed tests with 3, 6, and 10 random phenotypes added to the 3 ECG phenotypes. These assessments were designed to identify robustness of methods when phenotypes with shared genetic factors are unknown.

SAPPHO was then run in summary mode using sufficient statistics from CHARGE analyses. The sufficient statistics included phenotype covariances, phenotype-genotype regression coefficients and standard errors, allele frequencies, genotype-genotype covariances, and sample numbers for each phenotype-genotype association. For SAPPHO-C, the difference in sample numbers complicated the likelihood ratio term in the score statistics, and the computational expense increased dramatically as more associations were included into the model; therefore, only SAPPHO-I was run in this case. Permutations were performed by randomly resampling 100,000 individuals from the ARIC primary data. To be more specific, individuals from ARIC data were resampled with replacement for 100,000 times to construct 100,000 'new' individuals. With this procedure, the genotype allele distribution for resampled population should be consistent with that of the ARIC population. Shuffling and subsequent steps were then performed exactly as for the ARIC primary data to preserve the underlying phenotype-phenotype correlation was preserved. Results for 0.05 FDR used 10 population-wide, genome-wide permutations. The methods for constructing the gold standard and assessing true and false positives were the same as for the ARIC primary data.

To assess the biological relevance of new associations identified as significant by SAPPHO for the CHARGE cohort, enrichment analysis was performed for genes detected by SAPPHO-I at 0.05 FDR. The analysis focused on the curated gene sets including BIOCARTA, KEGG, REACTOME, and GO pathways as aggregated by MSigDB (Liberzon and et al., 2011). For this analysis, a 2 × 2 contingency table was constructed for each pathway, with 0/1 columns denoting whether the gene was detected by SAPPHO-I at 0.05 FDR as the columns, and 0/1 rows denoting whether that gene was in the pathway. Fisher's exact test was then run on each contingency table to obtain a one-sided p-value for a one-sided test of enrichment. We performed this assessment first for all the loci reported by SAPPHO-I. We then modified the procedure to account for the possibility that some of the pathway assignments found in MSigDB may have been influenced by the GWAS contributing to CHARGE, whose data we are using. Our modification was to exclude all gold-standard loci from consideration, removing them both from the SAPPHO-I results and from the gene sets. We then performed 2×2 contingency enrichment analysis as before but restricted to the non-gold-standard loci.

2.4.3 Results

2.4.3.1 Assessment for different methods on ARIC data

For evaluation with primary data, both SAPPHO-I and SAPPHO-C, together with other pleiotropy methods, were applied to identify shared genetic contributions to ECG traits (Fig. 2.6). Methods were calibrated using permutations to identify the appropriate threshold corresponding to a 0.05 FWER for genome-wide tests of 3 phenotypes. The conventional threshold for univariate tests of 3 independent phenotypes would be $(5/3) \times 10^{-8} = 1.67 \times 10^{-8}$. Calibration by permutation for univariate tests gave a single-test threshold of 1.068×10^{-8}. With ARIC data, SAPPHO-C was able to recall 15 known loci, better than any other methods, followed by SAPPHO-I, finding 13 known loci (Fig. 2.6). The MultiPhen and SHet methods are next best in performance, returning 8 or 9 true positives, but no better than standard univariate tests. The methods PCH and SHom perform worse than univariate tests (Fig. 2.6, Supplementary Table 2).

We then investigated the variants identified by each method. The set identified by SAPPHO included the variants identified by all other methods, except for a single locus found by SHet. This SNP, rs1896312, has known positive associations with $P_{CHARGE\_PR} = 1.151 \times 10^{-34}$ and $P_{CHARGE\_QRS} = 2.626 \times 10^{-9}$. P-values from ARIC for univariate tests were $P_{ARIC\_PR} = 1.5 \times 10^{-6}$ and $P_{ARIC\_QRS} = 3.3 \times 10^{-3}$. The p-value returned by SHet was 2.94×10^{-9}, better than its FWER=0.05 threshold value 1.52×10^{-8}. The SNP was added to both SAPPHO-I and SAPPHO-C models, but did not pass the 0.05 FWER threshold. The reason that this loci was only detected by SHet is that as observed with ARIC data and simulated data, SHet is powered to find variants that have weak associations across all or most of the given phenotypes. For SAPPHO, given its stringency for adding associations to the model, only the PR association was detected, and thus, this SNP was not reported as being significant.

Two SNPs were identified by multiple methods, yet were not in the gold-standard: rs7638275 and rs17608766. rs7638275 has $P_{CHARGE\_PR} = 0.845$, $P_{CHARGE\_QRS} = 0.41$, and $P_{CHARGE\_QT} = 0.4667$, and was therefore not included in the gold-standard. However, it has $P_{ARIC\_PR} = 5.27 \times 10^{-13}$ and $P_{ARIC\_QRS} = 1.97 \times 10^{-6}$, strong evidence for a true association within the ARIC subpopulation. We note that rs7638275 was reported as a rare variant with low imputation quality for most cohorts in CHARGE, and therefore it was not detected for any of the three ECG traits. In ARIC, however, rs7638275 was well imputed with a 1.5% minor allele frequency, and therefore detected by methods including SAPPHO-I, SHet, MultiPhen, PCH, and univariate tests. SAPPHO-C did not identify rs7638275 because its effect was partially explained by correlated SNPs already in the model. We reached this conclusion by attempting to add rs7638275 to the SAPPHO-C model; we found that its p-values were much less significant, 2.11×10^{-5} for PR and 0.01745 for QRS.

The other SNP, rs17608766, had p-values $P_{CHARGE\_PR} = 1.7 \times 10^{-7}$ and $P_{CHARGE\_QRS} = 1.2 \times 10^{-5}$ for CHARGE, and $P_{ARIC\_PR} = 1.4 \times 10^{-7}$ and $P_{ARIC\_QRS} = 3.0 \times 10^{-4}$ in ARIC cohort. It was not included the gold standard because none of its associations passed the p-value 5×10^{-8} threshold. This SNP was found by both SHom/SHet and SAPPHO-I using data from ARIC cohort. This SNP was also later found by SAPPHO-I run on CHARGE meta-analysis results, which is strong evidence for its association with ECG traits (Supplementary Table 3). This finding was supported by a recent study which reported rs17608766, located in the gene *GOSR2*, to be associated with cardiac structure and function (Wild et al., 2017), which exhibits SAPPHO's capability to identify novel pleiotropic associations.

We also investigated performance of each method as noise was introduced through addition of null phenotypes (Fig. 2.6). This assessment models a collection of phenotypes where only small subsets share genetic factors, and these subsets are unknown at the outset. The least robust method is SHom, which makes the assumption that all traits share genetic factors. Other methods also lose power when noise phenotypes are presented, though to a lesser degree. For real-world application, variants are not likely to be associated with all inputed phenotypes, making robustness when noise phenotypes are presented crucial.

2.4.3.2 Pleiotropy of ECG traits in ARIC and CHARGE

SAPPHO-I was run in summary mode on CHARGE meta-analysis summary results to see whether any additional pleiotropic SNP could be identified. The other methods were not run with CHARGE summary results, for two different reasons: PCH and MultiPhen require primary phenotype and genotype data which are not available for CHARGE, and SHom/SHet perform poorly based on results from ARIC cohort and simulations. These results, together with the association pattern obtained from ARIC and the gold-standard, are shown in (Fig. 2.7).

The genetic architecture of ECG traits includes SNPs contributing to distinct subsets of PR, QRS, and QT phenotypes. Given that ARIC is a subset of the CHARGE cohort, the power to detect true associations using ARIC data is smaller compared to using the entire CHARGE data. Therefore, for SAPPHO and univariate tests run on ARIC, in some cases gold-standard SNPs were not detected at all; in other cases, the strongest associations of a SNPs are retained but other gold-standard associations are lost. As is seen in (Fig. 2.7), the predicted number of associations is smaller than the expected number of associations, and the numbers denoting count of real hits lie below the diagonal line. For MultiPhen run on ARIC, however, the number denoting real hit counts all lie on or above the diagonal line, indicating that more associations were found for certain loci in the gold standard. Given the much smaller power of ARIC compared to CHARGE, these are likely to be over-predicted associations from MultiPhen rather than true pleiotropic loci; this over-predicting behavior of MultiPhen was later observed in the simulation studies as well (Supplementary Table 4, association pattern).

With ARIC data, SAPPHO methods were able to retrieve more hits than either MultiPhen or univariate tests. For CHARGE data, SAPPHO-I run in summary mode was able to retrieve all the real hits in the gold standard. Additional associations were found with SAPPHO-I for some gold standard loci, yielding additional loci with pleiotropic effects (Table 2.3). The number of additional associations added depends on the tuning parameter; for this test we set the γ parameter to allow for associations with $p < 10^{-4}$ to be added.

For loci already part of the gold standard, two types of new associations were added (Table 2.3): (1) a variant already associated with at least one trait was associated with at least one additional trait; (2) a variant not previously associated with any trait was associated with a new trait not yet associated with that locus. In loci not part of the gold standard, SAPPHO detected 124 new hits at 0.05 FDR. Most were associated with single traits, but the above-mentioned SNP rs17608766 in *GOSR2* was detected as pleiotropic. Pleiotropic effects at the locus level were observed more frequently. For example, *SLC12A7* contains rs2334955 and rs4285270, which were associated with QT and PR respectively, and *KLHL38* contains rs4871397 and rs16898685, which were associated with PR and QRS respectively. The linkage disequilibrium for each of these pairs of SNPs is weak, $R^2 = 0.064$ for rs2334955 and rs4285270 and $R^2 = 0.0017$ for rs4871397 and rs16898685, suggesting two independent effects within each locus. These associations reveal new genetic connections between different ECG traits.

Our analysis of CHARGE has no known ground truth; therefore, we used biological annotations to assess performance and gain insight. These assessments tested for enrichment of genes identified at 0.05 FDR for membership in annotated gene sets (see Methods). A total of 7246 gene sets were analyzed, corresponding to a nominal p-value of 6.9×10^{-6} for conventional significance. Genes detected by SAPPHO at 0.05 FDR show strong enrichment signals for several pathways involved with cardiac physiology and activities (Supplemental Table 5). The three most significant gene sets represent regulation of heart contraction ($p = 2.5 \times 10^{-19}$), muscle systems processes ($p = 3.6 \times 10^{-19}$), and cardiac conduction ($p = 3.7 \times 10^{-19}$). Additional notable categories include regulation of heart rate ($p = 3.0 \times 10^{-17}$), heart process ($p = 1.1 \times 10^{-14}$), and cardiac muscle cell action potential ($p = 4.3 \times 10^{-14}$). To ensure that these findings were robust, we repeated the analysis but excluded the significant findings from previously published GWAS. Gene sets specific to cardiac electrophysiology remain highly significant, including regulation of heart contraction ($p = 3.5 \times 10^{-8}$), muscle systems processes ($p = 2.3 \times 10^{-9}$), cardiac conduction ($p = 1.6 \times 10^{-7}$), as well as heart rate ($p = 5.9 \times 10^{-6}$), heart process ($p = 5.9 \times 10^{-6}$), and cardiac muscle cell action potential ($p = 3.1 \times 10^{-6}$). These findings support the conclusion that the novel loci with statistical associations detected by SAPPHO increase are indeed causal for ECG traits.

Given that pleiotropy is observed for ECG traits, we asked whether variants that affect the same subset of traits typically have the same direction of effect, and whether this depends on the observed phenotypic correlation. For ECG traits, $\mathrm{Cor}(\mathrm{PR}, \mathrm{QRS}) = 0.051$, $\mathrm{Cor}(\mathrm{PR}, \mathrm{QT}) = -0.026$, $\mathrm{Cor}(\mathrm{QRS}, \mathrm{QT}) = 0.168$, where $\mathrm{Cor}(x, y)$ stands for the correlation between the two traits x and y. The consistency of direction of effects and phenotypic correlations for pleiotropic SNPs detected by SAPPHO on CHARGE are shown in Table 2.4, where '+/+' denotes the SNPs whose effects are in the same direction, and '+/-' denotes the count of SNPs whose effects are in different directions. Out of the 23 pairs of association effects, 16 were consistent with the phenotypic correlations. The probability that association effects were consistent with phenotype correlation was 0.70 with a 95% binomial distribution confidence interval of $[0.49, 0.84]$. We conclude that SAPPHO is able to detect variants whether or not the direction of genetic effect matches the overall direction of correlation. We further conclude that variants that contribute to the same pair of phenotypes can often show different directions of effect, suggesting that distinct biological mechanisms connect the variants to their downstream effects on ECG traits.

2.5 Application to UK Biobank cancer phenotypes

2.5.1 UK Biobank project and cancer phenotypes

Cancer is the second leading cause of death in the United States, which develops when abnormal cells undergo uncontrolled growth and spread. In contrast to benign tumors, cancer cells usually invades to other parts of the body, impacting the normal functioning of these tissues and organs, thus becomes life threatening. In 2017, 1.7 million new patients were diagnosed with cancer, and about 600,000 patients died because of it (AmericanCancerSociety, 2017). Large twin studies of cancer have been performed in order to provide insight into the relative contribution of inherited factors and characterize familial cancer risk by leveraging the genetic relatedness of monozygotic and dizygotic pairs of twins (Mucci et al., 2016). With a recent study which consists of 27,156 incident cancers happened in 23,980 individuals, it was found that for most cancer types, there were significant familial risks and the cumulative risks were higher in monozygotic than dizygotic twins. Heritability of cancer overall was 33%(95% CI, 31%-37%), with significant heritability detected for skin melanoma(58%; 95% CI, 43%-73%), prostate cancer(57%; 95% CI, 51%-63%), non-melanoma skin(43%; 95% CI, 26%-59%), ovary(39%; 95% CI, 23%-55%), kidney(38%; 95% CI, 21%-55%), breast (31%; 95% CI, 11%-51%), and corpus uteri(27%; 95% CI, 11%-43%) (Mucci et al., 2016).

GWAS has been performed extensively on cancer phenotypes. To date, over 450 genetic variants have been proved to associate with increased risk of cancer (Sud et al., 2017). Some of the well-proved cancer susceptibility genes include those for breast (BRCA1/2), colorectal cancer(CRC), and melanoma (CDKN2A) (Sud et al., 2017). Besides revealing novel pathways important in carcinogenesis, GWAS studies have also shown that some common genetic variations contribute substantially to the heritable risk of many common cancers (Wu et al., 2017). This makes looking into the pleiotropic effect across different cancer phenotypes meaningful, and discovering such genetic variants would provide opportunities for drug discovery and repositioning as well as for cancer prevention.

In this project, we decided to look into cancer susceptibility, as well as the pleiotropic effect for genetic variants, using data from the UK Biobank (UKB) project (Bycroft et al., 2017). The UK Biobank project is a large prospective cohort study consisting of 500,000 individuals from across the United Kingdom, aged between 40-69 at recruitment. A rich variety of phenotypic and health-related information was collected on each participant, including self-reported information like basic demographics, diet, and exercise habits; as well as other sources of health-related information such as medical records and cancer registers being integrated and followed up over the course of the participants' lives. 14 distinct cancer phenotypes were provided among all phenotypes given by UKB project, with in total 40,000 cancer incidents reported. Genotypes for more than 98% of all participants were collected, which provides great potential to exploit the interaction between cancer phenotypes and genotypic variants. For the study, both traditional univariate logistic regression test and SAPPHO were run on the cancer phenotypes; hits were collected at 0.05 FWER for univariate tests and 0.05 FDR for SAPPHO, and first compared with genome-wide associations in GWAS catalog to see if any new associations were detected. Then, phenotypes associated with each locus were compared between the two methods, to see whether any additional pleiotropic effects were discovered.

2.5.2 Methods

UKB project participants were genotyped using two arrays that are very similar with each other. The UKB Axiom array was used to genotype $\sim 450,000$ of the $\sim 500,000$ individuals, and the other $\sim 50,000$ samples were genotyped using the UK BiLEVE Axiom array. Both arrays include $\sim 825,000$ markers, including both single nucleotide polymorphisms (SNPs) and small insertions and deletions (Indels). 95% of common markers for the two arrays are overlapped, with UKB Axiom array including additional novel markers such as cancer-related markers, which replaced a small fraction of markers used for genome-wide coverage. The position of markers in are reported in coordinates of the genome build Genome Reference Consortium Human Reference 37 (GRCh37). In order to increase power of the study, number of markers were boosted by first pre-phasing the directly genotyped markers, followed by haploid imputation. Pre-phasing was done by first removing markers having more than 5% missingness or having minor allele frequency $< 10^{-4}$, which resulted in 670,739 autosomal markers in 487,442 samples; then phasing on the autosomes were carried out using SHAPEIT3 (O'Connell et al., 2016) in chunks of 15,000 markers, with an overlap of 250 markers between chunks. The 1000 Genome Phase 3 dataset (1000GenomesProjectConsortium et al., 2010) was used as reference panel, predominantly to help with the phasing of sample with non-European ancestry. Imputation using IMPUTE2 (Howie et al., 2009, 2011) was carried out in chunks of approximately 50,000 imputed markers with 250kb buffer region on 5,000 samples per compute job, which resulted in 92,693,895 autosomal SNPs in 487,442 individuals.

For GWAS, it is usually preferred that effects of strong population structures are reduced and the samples are unrelated, yielding a set of independent samples with relatively homogenous ancestry, to satisfy the independent identical distributed assumption for most tests. Therefore, two more processes were performed on the samples to get a maximal set of unrelated individuals with homogeneous ancestry. First, a white British ancestry subset was selected; this was done by first selecting 431,059 individuals who reported their ethnic background as "British"; then a Bayesian outlier detection algorithm implemented in the R package called aberrant (Bellenguez et al., 2012) was run on the set to isolate the largest cluster of samples from the rest, using the first 40 principle components. Then, a maximal set of unrelated individuals was found by using the i-graph(Csardi and Nepusz, 2006) package in R, where for each family (network of nodes joined by edges), the largest subset of individuals (vertices) were found such that there is no relatedness (edges) between them. After doing these two sample filtering procedures and excluding samples that have missing values, we were left with $\sim 330,000$ samples. Out of the $\sim 330,000$ samples, the number of cases for each cancer phenotype is shown in Table 2.5. It could be observed that bone, articular cartilage of limbs has number of cases of 58, which is considered too small a number to perform associations tests; for other cancer phenotypes, the number of cases range from a few hundreds to $\sim 15,000$, which guarantees power to detect meaningful associations.

To test associations between genetic variants and cancer phenotypes, univariate logistic regressions were first performed on each cancer phenotype. To avoid running logistic regressions on highly unbalanced data, given the large population of UKB individuals and relatively small number of cases for each cancer phenotype, control samples were sampled randomly from the above mentioned unrelated British ancestry population, such that the case/control has an approximately 1:3 ratio. For gender-specific cancer types, including breast cancer and female genital organs for female, and male genital organs for male, controls were sampled from only female/male sub-population of the unrelated British ancestry population for more detailed stratification. Covariates of the tests included BMI, current smoker, age, and 40 genetic principal components; sex was included as a covariate only for non-gender-specific cancer types, to avoid collinearity with the intercept term of the linear component of logistic regression, which will lead to singularity when taking inverse. SNPs were excluded for $MAF < 0.01$, $HWE < 10^{-5}$.

SAPPHO was run on summary statistics obtained from logistic regression. The candidate SNP list selection, filtering procedures, and hit calling were performed exactly the same as for ARIC ECG phenotypes. Since SAPPHO is implemented as a linear model, the regression coefficients and standard errors obtained from logistic regression could not be directly used as inputs; therefore, these two terms were back calculated from the z-score of each univariate test, under the assumption that when effect size is small, power for logistic regression and linear regression will be similar. To be more specific, assume each phenotype has variance equal to 1, then for any SNP with minor allele frequency p:

$$se \sim \sqrt{\frac{1}{N \times 2p(1-p)}}$$

and

$$beta = se \times Z$$

where N stands for the effective sample size for that specific association, and Z stands for the z-score for the association. Each phenotype were assumed to have unit variance, and covariance between any pair of phenotypes were set to be 0. Genotypic covariances and minor allele frequency for each SNP were calculated using the unrelated British ancestry mentioned above. Permutations were first performed for each phenotype individually, and then followed by exact same steps as for ARIC ECG permutations. In total 10 permutations were performed to control for SAPPHO result at 0.05 FDR.

2.5.3 Results

Numbers of associations detected on locus level for both methods are shown in Table 2.6. It could be observed that no association was detected using either method for 4 out of 13 cancer phenotypes. Univariate test detected 54 associations with 5 cancer phenotypes, and SAPPHO detected 66 associations with 9 phenotypes. More than ten associations were detected by both methods for skin cancer, breast cancer, and cancer for male genital organs, with fewer associations detected for other cancer phenotypes. All associations detected by univariate tests were also detected by SAPPHO at 0.05 FDR. Comparing the detected hits with known associations from GWAS catalog, it could be observed that most hits detected by either methods have been well established to be associated with the corresponding phenotypes (Supplementary Table 6). For example, many detected associations with skin cancer were found in GWAS catalog to be associated with “basal cell carcinoma”, “melanoma”; and most detected associations with male genital organs were found in GWAS catalog to be associated with “prostate cancer”, and etc.

Number of different loci detected by univariate tests and SAPPHO was 51 and 59, respectively (Supplementary Table 6). For both methods, the same four loci were detected to be pleiotropic, with the remaining 47 and 55 loci associated with one single phenotype (Table 2.7). For two out of all four pleiotropic loci, SAPPHO was able to detected one extra association for each one of them: cancer for digestive organs for the *SRRM1P1-POU5F1B* locus, and cancer for respiratory intrathoracic organs for the *TERT-MIR4457-CLPTM1L* locus. Both extra associations were verified by comparing with known associations from GWAS catalog, which further proved SAPPHO's advantage over univariate test. 8 novel cancer related loci were detected by SAPPHO (Supplementary Table 6). Unlike the 51 loci detected by both SAPPHO and univariate test, some of these 8 additional loci did not have explicit associations with corresponding cancer phenotypes in GWAS catalog. For some of the novel hits detected, like *SNRK*(rs76161159), *GLIS3* (rs12378136), and *PCDH15* (rs118113166), it was observed that the detected loci were associated with more general underlying biological processes that might contribute to cancer in a more indirect way; for other novel hits detected, like *ZNF646P1* (rs7997746), associations from GWAS catalog indicate that they are associated with explicit phenotypes that doesn't seem to be correlated with cancer phenotypes, like the association with BMI in the case of rs7997746. Another reason for these detections might be because for SAPPHO, correction is performed to allow 5% false discoveries on the SNP association level, which given the ~ 90 total associations between SNP and cancer phenotypes, the expected number of false discoveries will be around 5. A way to further eliminate the false discoveries will probably be to run more permutations and control for family-wise error rate at desired level, say 5%. Nonetheless, these novel hits detected by SAPPHO could provide potential cancer associated loci that could be further investigated into.

2.6 Discussion

We have presented a Bayesian-motivated method, SAPPHO, designed to detect pleiotropic effects in GWAS. SAPPHO exploits previously observed association patterns to identify additional variants following the same pattern. Representative methods were selected for comparison: SHom and SHet, which pool summary statistics of all variant-phenotype associations to define a combined test-statistic; PCH, which constructs linear combinations of phenotypes for genotype data to be regressed on; and MultiPhen, which performs reverse regression such that the phenotypes are treated as predictors to explain the variance of genotypes.

Simulated phenotype and genotype data sets were first used to compare the methods. These studies suggested that the version of SAPPHO that models the full phenotype covariance matrix, SAPPHO-C, can actually perform poorly when phenotypes are strongly correlated. A simplified version in which phenotypes are modeled as independent when conditioned on genetic effects, SAPPHO-I, retains robust performance. SAPPHO-I is also more computationally efficient and more amenable to use with summary data from meta-analysis.

Simulations also demonstrated that SAPPHO can have an advantage over univariate tests even applied to a mixture of phenotypes in which some lack genetic effects. The Bayesian prior learns this pattern and is able to boost associations with the phenotypes that have genetic effects.

The MultiPhen method also performed well. It out-performed SAPPHO in some simulation settings involving weaker effects, although it had a drawback of over-predicting spurious associations for variants with true associations for a subset of phenotypes.

SAPPHO depends on a single adjustable parameter, which in effect determines the minimum effect strength that can be entered into the genetic model. Permitting weaker effects, expressed as a looser univariate p-value, improved SAPPHO's performance. We found, however, that the greedy forward search implemented by SAPPHO occasionally yields a local rather than global optimum, as assessed by calculating the score of the true model. Improving the search heuristic, for example by permitting the model to add two associations simultaneously, may improve the performance.

In applications to a real data set, ECG phenotypes from ARIC, with known positives available from the much larger CHARGE study, SAPPHO performed better than other pleiotropy methods in discovering the true associations at genome-wide significance. SAPPHO also performed best when additional random phenotypes augmented the true phenotypes, an assessment of performance when pleiotropy involves only a subset of the phenotypes in a study.

SAPPHO uses an association model that is in the exponential family, which makes it amenable to use with summary statistics rather than individual phenotype-genotype data in the context of meta-analyses. In applications to meta-analysis data from CHARGE, SAPPHO identified 295 loci at 0.05 FDR, corresponding to 171 loci in the genome-wide significance gold standard and 124 novel loci. Gene sets corresponding to cardiac electrophysiology are highly enriched for these novel loci, supporting a conclusion that SAPPHO has identified many additional relevant loci beyond those previously reported. Similar results were observed for SAPPHO run on data from UK Biobank project, where 2 pleiotropic loci with additional associations and 8 novel loci were detected for cancer phenotypes. While some of the additional loci may arise from using a less stringent 0.05 FDR threshold compared with the 0.05 FWER threshold, we also note that there are no established methods to define significant loci for pleiotropic tests. Investigating the direction of effect for pleiotropic variants, we find that variants affecting pairs of traits often have relative directions of effect that are different and that often do not match the overall phenotypic correlation. These findings suggest that multiple independent biological mechanisms connect pleiotropic variants to downstream phenotypes.

We conclude that SAPPHO, and particularly the SAPPHO-I implementation for summary statistics, is a powerful method for discovering pleiotropic patterns of association in the context of single studies, with access to individual genotype and phenotype data, and also to meta-analyses. Application to large compendiums of GWAS results, for example dbGaP or the UK BioBank, could lead to new discoveries of genetic associations and patterns of shared genetic architecture for human phenotypes and disease.

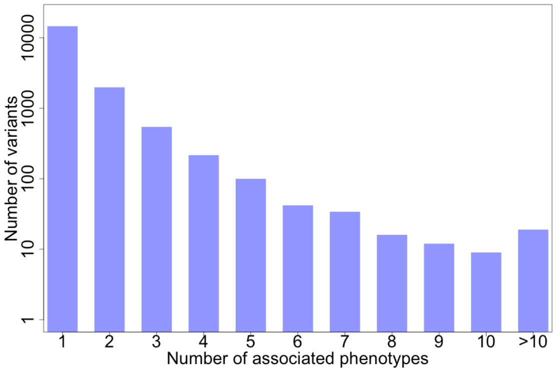

Figure 2.1: Histogram of pleiotropic variants. Counts are for variants associated with at least one trait at genome-wide significance ($p < 5 \times 10^{-8}$) (Welter et al., 2014); no attempt was made to correct for correlation between variants (linkage disequilibrium) or between traits. The number of total variants is 17607, and total number of phenotypes is 785.

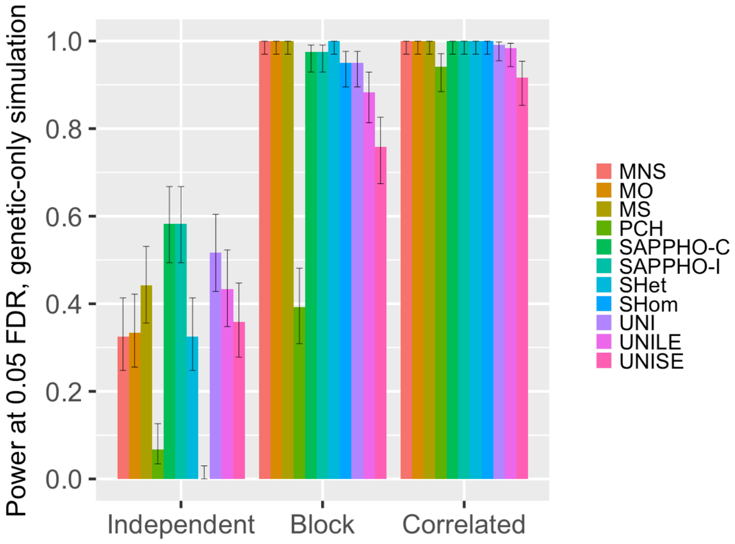

Figure 2.2: Power for associations of 6 phenotypes simulated to be correlated through genetic effects only. Scenarios are Independent (6 phenotypes, no pleiotropy), Block (2 blocks of 3 phenotypes each), and Correlated (all 6 phenotypes correlated). Error bars indicate 95% confidence intervals estimated from 5 repeated runs and a binomial distribution. For the independent scenario lacking pleiotropy, SAPPHO methods performed the best. For the block correlation scenario, MultiPhen leading performed best, followed by SAPPHO. For a single correlated block, all pleiotropy methods perform well. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI); univariate test at loose empirical threshold $p < 5 \times 10^{-8}$ (UNILE); univariate test at stringent empirical threshold $p < 5 \times 10^{-8}/6$ for 6 phenotypes (UNISE).

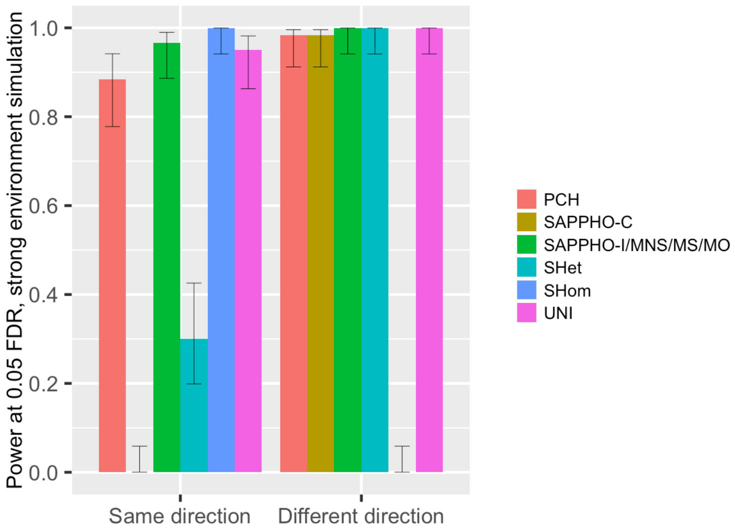

Figure 2.3: Power to detect associations for scenarios involving strong environmental correlations. Two sets of SNPs each containing 12 variants were simulated to contribute to 4 phenotypes. For the same-direction scenario, all 12 SNPs have positive effect with all phenotypes; for the different-direction scenario, the 12 variants had positive effects for the first 2 phenotype and negative effects for the second 2 phenotypes. SAPPHO-C and SHet were unable to detect same-effect SNPs; SHom could not detect different-direction SNPs. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI).

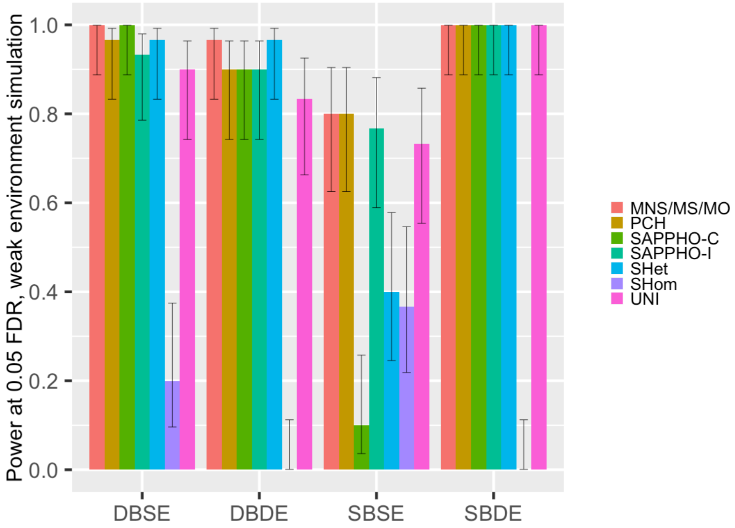

Figure 2.4: Power to detect associations for scenarios involving weak environmental correlations. Four sets of SNPs each containing 6 variants were simulated to contribute to 4 phenotypes. Scenarios as described in the main text were DBSE, different-block-same-effect; DBDE, different-block-different-effect; SBSE, same-block-same-effect; and SBDE, same-block-different-effect. SAPPHO-I and MultiPhen performed the best over all scenarios. SAPPHO-C experiences dramatic loss of power under the sbse scenario, similar to its loss of power for similar scenarios involving strong environmental correlations. Methods were MultiPhen-NonSelection (MNS); MultiPhen Ordinal regression (MO); MultiPhen-Selection (MS); univariate test corrected with permutation (UNI).

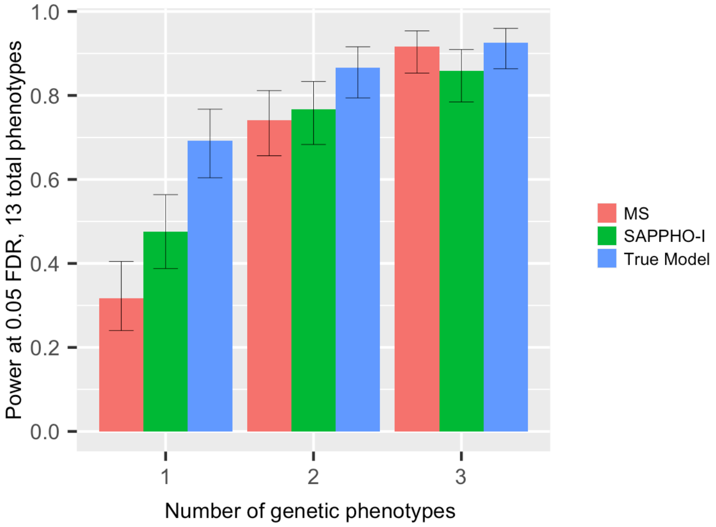

Figure 2.5: Simulations of a mixture of genetic and non-genetic phenotypes, with 1, 2, or 3 genetic phenotypes embedded as part of 13 total phenotypes. The remaining non-genetic phenotypes generated as standard normal random variables. The SAPPHO parameter was set to include associations with p-value $< 5 \times 10^{-4}$. Methods were MultiPhen in selection mode (MS), SAPPHO-I with a greedy forward (SAPPHO-I), and the SAPPHO-I score for the true model (True Model). The greedy forward search limits the power of SAPPHO-I; a more sophisticated strategy could improve its power.

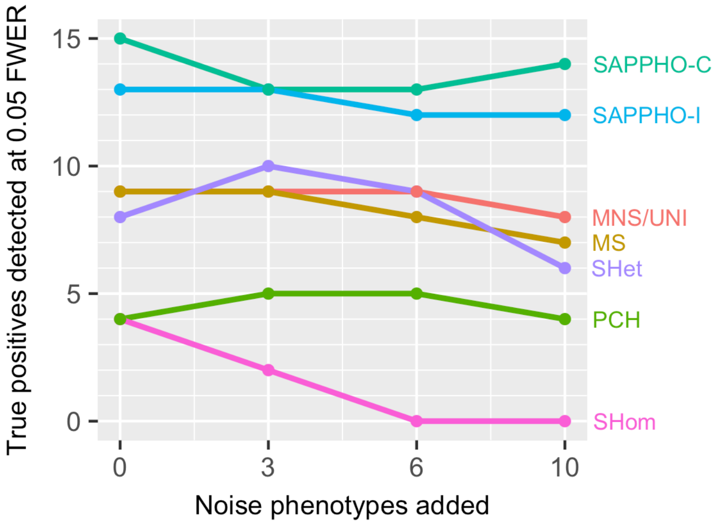

Figure 2.6: Pleiotropy methods were used to detect associations with the PR, QRS, and QT phenotypes in the ARIC cohort. The three measured phenotypes were then augmented with 3, 6, and 10 noise phenotypes. SAPPHO had the greatest power to detect associations regardless of noise phenotypes were present. The pleiotropy methods SHet, MS, MNS had power similar to standard univariate tests. PCH and SHom had lower power than univariate tests, and the performance of SHom degraded further as noise phenotypes were added. All methods were controlled for type I error at FWER = 0.05. Methods were MultiPhen-Selection (MS), Mutiphen-NonSelection (MNS), Principal Components of Heritability (PCH), Homogeneous Test Statistics (SHom), Heterogeneous Test Statistics (SHet), and univariate tests corrected using permutations (UNI).

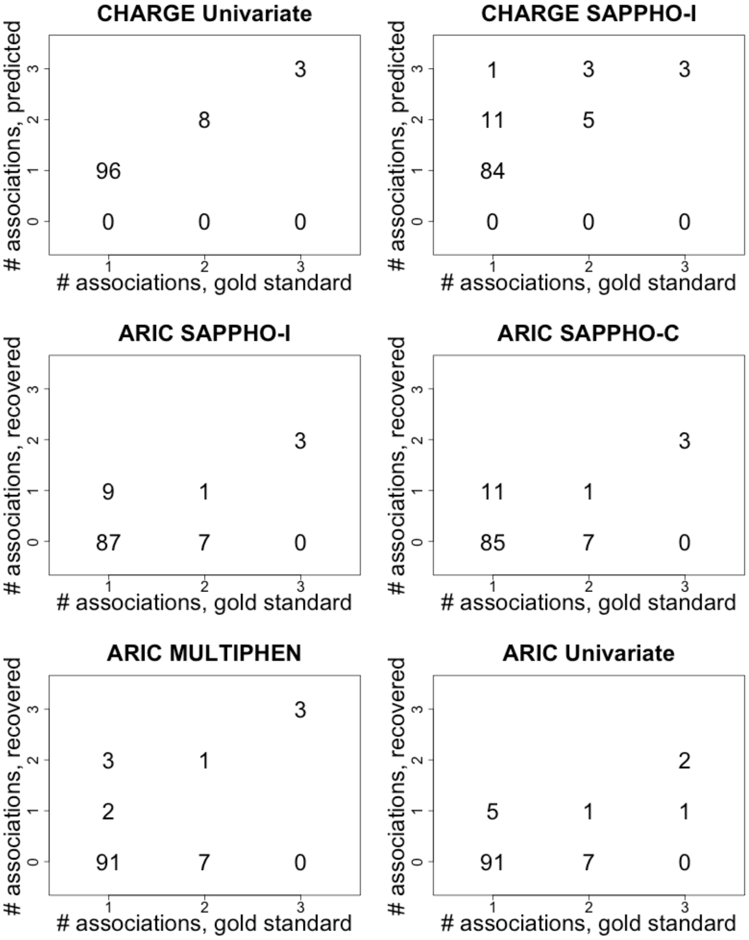

Figure 2.7: Number of associations recovered versus number previously known ($p < 5 \times 10^{-8}$ in CHARGE for univariate single phenotype test). First row: CHARGE meta-analysis data as analyzed by univariate tests for each phenotype ($p < 5 \times 10^{-8}$) and by SAPPHO-I (0.05 FDR). The first row denotes univariate test (with $p < 5 \times 10^{-8}$ cut-off) and SAPPHO-I run on CHARGE meta-analysis results; Second row: ARIC data as analyzed by SAPPHO-I and SAPPHO-C using 0.05 FWER threshold. Third row: ARIC data as analyzed by MultiPhen and univariates test using 0.05 FWER threshold, equivalent to $p < 5 \times 10^{-8}/3$ for univariate tests. More discoveries are made with CHARGE (top row) because it is a larger cohort and because 0.05 FDR is a less stringent threshold than 0.05 FWER. SAPPHO has greater power for the ARIC cohort than MultiPhen or univariate tests. Pleiotropic associations discovered by MultiPhen in the ARIC cohort (bottom left panel) may be over-predictions as these associations were not genome-wide significant in the much larger CHARGE cohort.

Method	Run Time
SHom	0m2s
SHet	0m13s
SAPPHO-I	1m1s
MultiPhen NonSelection Gaussian	2m2s
PCH	3m38s
SAPPHO-C	4m39s
MultiPhen Selection Gaussian	4m55s
MultiPhen NonSelection Ordinal	30m15s
MultiPhen Selection Ordinal	115m28s

Table 2.1: Running time for different methods for a simulation with 10,000 individuals, 6 phenotypes, and 638 SNPs of which 24 had associations, with Core i5 2.9 GHz CPU, 8 GB RAM. 12 SNPs are associated the first three phenotypes, while the other 12 SNPs are associated with the second three phenotypes.

	Parsimonious	Non-Parsimonious
Phenotype selection	SAPPHO-I SAPPHO-C	Univariate MultiPhen-Selection
No phenotype selection		SHom, SHet, PCH MultiPhen-NonSelection

Table 2.2: Assessed methods provide qualitatively different types of predictions. 'Parsimonious' indicates that a single SNP is selected from an LD block, whereas 'Non-Parsimonious' indicates that all SNPs in an LD block are reported. 'Phenotype selection' indicates that the subset of associated phenotypes is reported, whereas 'No phenotype selection' indicates that the results do not specify which phenotypes are associated with a SNP and which are not.

Locus	Univariate GWS	SAPPHO 0.05 FDR	SAPPHO, SNPs	P_{PR}	P_{QRS}	P_{QT}
LRIG1	QRS	PR, QRS	rs2242285ă	1.3×10^{-6}	2.0×10^{-8}	0.094
HERPUD2	PR	PR, QRS	rs11763856	4.5×10^{-10}	1.82×10^{-6}	0.018
SIPA1L1-C14orf56	PR, QRS	PR, QRS, QT	rs17767398	6.4×10^{-13}	1.3×10^{-10}	3.1×10^{-5}
LAP3P2	QRS	PR, QRS	rs9470361	5.6×10^{-7}	8.8×10^{-29}	0.0048
EPS15	PR	PR, QRS	rs17106627	2.7×10^{-8}	2.2×10^{-7}	0.49
SLC35F1-C6orf204	QRS,QT	PR,QRS,QT	rs11153730 (plus 3 additional)	1.2×10^{-6}	1.3×10^{-19}	5.2×10^{-67}
LOC401324	QRS	PR, QRS	rs340389	6.4×10^{-7}	2.6×10^{-8}	0.011
		PR	rs12673438	9.3×10^{-8}	0.21	0.53
NOS1AP-OLFML2B	QT	QT	rs12143842	0.12	1.3×10^{-4}	8.97×10^{-210}
		QRS	rs4656349 (plus 10 additional)	0.22	3.1×10^{-6}	5.23×10^{-134}
CAV1-CAV2-TES	PR	PR	rs3807989	8.7×10^{-69}	5.8×10^{-6}	6.6×10^{-5}
		QRS	rs6867	4.0×10^{-20}	1.3×10^{-6}	2.0×10^{-4}
		QT	rs7801180	1.8×10^{-15}	1.7×10^{-6}	1.4×10^{-6}
SLC8A1	QT	QT	rs12997023	3.1×10^{-6}	0.06	5.4×10^{-14}
		PR	rs4993292	4.0×10^{-6}	0.06	2.3×10^{-13}
HETR5B-STRN	QRS	QRS	rs2160411	5.0×10^{-3}	1.3^{-9}	2.3×10^{-4}
		PR	rs6744560	1.2×10^{-7}	6.5×10^{-4}	0.45
FADS2	QT	QT	rs174577	5.0×10^{-4}	2.2×10^{-5}	1.2×10^{-10}
		PR	rs2727270	3.1×10^{-6}	0.34	0.22
SKI	PR	PR	rs4648819	4.7×10^{-10}	1.1×10^{-3}	1.8×10^{-2}
		QRS	rs12045693	1.5×10^{-3}	6.7×10^{-6}	0.12
SMARCAD1	QT	QT	rs183993	1.9×10^{-4}	0.98	8.8×10^{-9}
		PR	rs2639793	3.6×10^{-6}	0.34	0.11
CCDC141-TTN	PR,QT	PR	rs922984	1.8^{-11}	0.11	3.6×10^{-3}
		PR	rs10497523	2.6×10^{-8}	0.42	0.18
		QT	rs7600330	0.045	0.068	3.2×10^{-8}
		QRS	rs17362588	1.01×10^{-6}	2.2×10^{-7}	0.087

Table 2.3: Additional novel associations detected by SAPPHO-I in known gold-standard loci. **Locus**: Gene symbol or symbols spanning an associated region. **Univariate GWS**: phenotypes detected at 5×10^{-8} genome-wide significance threshold. **SAPPHO 0.05 FDR**: phenotypes detected by SAPPHO at 0.05 FDR. **SAPPHO, SNPs**: SNPs detected by SAPPHO, with rows corresponding to the previous *SAPPHO 0.05 FDR* column. **P_{PR}, P_{QRS}, P_{QT}**: univariate p-values from CHARGE meta-analysis. New associations added to a locus are in two categories: (1) a variant already associated with at least one trait is associated with a new trait or traits (the first 7 loci in the table); (2) a new variant is introduced and is associated with a trait not previously associated with the locus (the last 8 loci in the table).

Phenotypes	Correlation	+/+	+/−
PR-QRS	0.051	9	3
QRS-QT	0.168	3	4
PR-QT	−0.026	0	4

Table 2.4: Consistency of direction of effects and phenotypic correlations for each pair of ECG phenotypes. 'Correlation' denotes the correlation of each phenotype-phenotype pair; '+/+' denotes the count of SNPs whose effects are in the same direction; '+/−' denotes the count of SNPs whose effects are in different directions. Of the 23 pairs of association effects, 16 were consistent with the phenotypic correlation, corresponding to a probability of 0.70 (binomial parameter 95% confidence interval $[0.49, 0.84]$) that the direction of effect agreed with the phenotypic correlation. These results demonstrate that SAPPHO is able to detect variants whether or not the direction of genetic effect matches the overall direction of phenotypic correlation.

Cancer type	ICD9	ICD10	Number of cases
Melanoma, malignant neoplasm of skin	172-173	c43-c44	14637
Breast cancer	174-175	c50	8847
Male genital organs	185-187	c60-c63	5913
Digestive organs	150-157,159	c15-c26	4760
Primary lymphoid	200-208	c81-c96	2623
Female genital organs	179-184	c51-c58	2343
Urinary tract	188-189	c64-c68	1492
Respiratory, intrathoracic organs	160-165	c30-c39	1441
Lip, oral cavity, pharynx	140-149	c00-c14	649
Central never system	190-192	c69-c72	446
Peripheral nerve tumors	158,171,176	c45-c49	421
Endocrine glands	193-194	c73-c75	400
Unspecific sites	195-199	c76-c80	320
Bone, articular cartilage of limbs	170	c40-c41	58

Table 2.5: Information for cancer phenotypes on UKB data. Columns are cancer phenotypes, cancer identification codes according to ICD9 coding, cancer identification codes according to ICD10 coding, and number of cases for each type of cancer.

Cancer type	Univariate GWS	SAPPHO 0.05 FDR
Melanoma, malignant neoplasm of skin	29	29
Breast cancer	12	12
Male genital organs	12	12
Unspecific sites	1	4
Endocrine glands	1	4
Central nerve system	0	1
Digestive organs	0	1
Respiratory intrathoracic organs	0	1
Peripheral nerve tumors	0	1

Table 2.6: Number of associations for each cancer phenotype for univariate test at $p = 5 \times 10^{-8}$ and SAPPHO at 0.05 FDR on locus level. No association was detected for either method for 4 out of 13 cancer phenotypes. Univariate test detected in total 54 associations with 5 cancer phenotypes, and SAPPHO detected 66 associations with 9 phenotypes. More than ten associations were detected by both methods for skin cancer, breast cancer, and cancer for male genital organs, with a few associations detected for other cancer phenotypes. All loci detected by univariate tests were also detected by SAPPHO at 0.05 FDR.

Locus	Univariate GWS	SAPPHO 0.05 FDR	GWAS catalog
MYEOV,IFITM9P	mgo, breast	mgo, breast	breast cancer, prostate cancer
RPL7P41	mgo, mmns	mgo, mmns	basal cell carcinoma, cancer (pleiotropy)
SRRM1P1, POU5F1B	mgo, breast	mgo, breast, do	prostate cancer, colorectal cancer, breast cancer, etc
TERT,MIR4457,CLPTM1L	mgo, mmns	mgo, mmns, rio	prostate cancer, lung adenocarcinoma, basal cell carcinoma, etc

Table 2.7: Pleiotropic loci detected by univariate test at $p = 5 \times 10^{-8}$ and SAPPHO at 0.05 FDR. The same four loci were detected to be pleiotropic by both methods. For two out of all four pleiotropic loci, SAPPHO was able to detect one extra association. All associations were verified by comparing the detected associations with known associations from GWAS catalog.

Description for supplementary tables

Supplementary Table 1 : Gold-standard from CHARGE meta-analysis results at different thresholds for generating the connected components.

Supplementary Table 2 : SNPs detected by pleiotropy methods on ARIC data at 0.05 FWER. For each table, multiple SNPs can belong to the same locus. The number of loci detected by each method is shown in Fig. 2.6.

Supplementary Table 3 : All variants detected by SAPPHO-I on CHARGE meta-analysis results at 0.05 FDR.

Supplementary Table 4 : Results for simulation with noise phenotypes at 0.05 FDR. Sheet 1: Number of true positives and false positives for all methods; Sheet 2: Number of true positives for SAPPHO-I fed with true associations at different thresholds, and MultiPhen in select mode; Sheet 3: association patterns for SAPPHO-I and MultiPhen in select mode: *Missed entirely* denotes the number of variants missed entirely; *Under-predicted* denotes the number of variants detected but with a subset of true associations found; *Exactly-predicted* denotes variants detected with the correct association pattern; *Over-predicted* denotes variants detected with a mix of true associations and spurious associations; *False positives* denotes variants reported but lacking any true associations.

Supplementary Table 5: Gene sets enriched for loci reported by SAPPHO-I at 0.05 FDR for CHARGE ECG meta-analysis, calculated for all loci and for novel findings defined by excluding the ECG gold-standard loci. Meaning of the columns: ‘Pathway’ denotes the pathway name; ‘pathway gene counts’ denotes the number of genes in that pathway; ‘SAPPHO gene counts’ denotes number of all genes detected by SAPPHO; ‘SAPPHO 11’ denotes the number of genes detected by SAPPO and in that corresponding pathway; ‘SAPPHO 10’ denotes the number of genes detected by SAPPHO but not in the pathway; ‘SAPPHO 01’ denotes the number of genes in the pathway but not detected by SAPPHO; ‘SAPPHO 00’ denotes number of genes not in the pathway and not detected by SAPPHO; ‘SAPPHO pval’ denotes the pvalue of Fisher’s exact test; ‘SAPPHO loci’ denotes the genes in the ‘SAPPHO 11’ column. The following columns are of the same meaning for the SAPPHO loci excluding the gold-standard.

Supplementary Table 6: Results for cancer phenotypes using traditional univariate test and SAPPHO in summary mode on UK Biobank cancer phenotypes. Results were corrected to control for 0.05 FWER for univariate test and 0.05 FDR for SAPPHO.

Chapter 3

The paltry power of priors versus populations

CHAPTER 3. THE PALTRY POWER OF PRIORS VERSUS POPULATIONS

Abstract

Biological experiments often involve hypothesis testing at the scale of thousands to millions of tests. Gene-based tests for human RNA-Seq data, for example, involve approximately 20,000; genome-wide association studies (GWAS) involve about 1 million effective tests. The conventional approach is to perform individual tests and then apply a Bonferroni correction to account for multiple testing. This approach implies a single-test p-value of 2.5×10^{-6} for RNA-Seq experiments, and a p-value of 5×10^{-8} for GWAS, to control the false-positive rate at a conventional value of 0.05. Alleviating the multiple testing burden has been a goal of many methods designed to boost test power by focusing tests on the alternative hypotheses most likely to be true. Very often, these methods either explicitly or implicitly make use of prior probabilities that bias significance for favored sets thought to be enriched for significant finding. At the extreme limit, only the candidate set is tested, corresponding to a decreased multiple testing burden. Despite decades of methods development, prior-based tests have not been generally used; most genomics experiments, and in particular genome-wide association studies (GWAS), still use traditional univariate tests rather than more sophisticated approaches. Here we use GWAS to demonstrate

why unbiased tests remain in favor. Here we compare the power increase possible with a prior with the increase possible with a much simpler strategy of increasing a study size. We calculate test power assuming perfect knowledge of a prior distribution and then derive the population size increase required to provided the same boost without a prior. We show that increasing the population size is exponentially more valuable than increasing the strength of prior, even when the true prior is known exactly. The results provide a rigorous explanation for the observed avoidance of prior-based methods, and support for the continued use of simple, robust methods rather than more sophisticated approaches.

3.1 Introduction

Genomics experiments involve testing thousands to millions of hypotheses. In functional genomics and proteomics, each gene or protein usually corresponds to a single test, with 20,000 or more tests required for an RNA-Seq or proteomics experiment. In human genetics, the number of independent tests accounting for linkage disequilibrium in a single ethnicity is usually assumed to be about 1 million. To maintain a family-wise error rate (FWER) controlled at 0.05, a long-standing approach has been to apply a Bonferroni correction, requiring a single-test p-value of 0.05 divided by the number of hypotheses tested. This multiple-testing correction from this stringent approach is seen as a burden for identifying genome-wide significant findings.

A current direction of GWAS is to incorporate prior knowledge about functional effects of SNPs, in order to increase the power to detect SNPs with true associations or to identify which SNP in an linkage disequilibrium (LD) region is most likely to be the causal variant (Hindorff et al., 2009; Schork et al., 2013; Petersen et al., 2013; Sveinbjornsson et al., 2016). A representative approach incorporated 450 different annotations into GWAS analysis of 18 human traits; the number of loci with high-confidence associations was increased by around 5% (Pickrell, 2014). Despite the intuitive value of incorporating pre-existing biological knowledge, it remains unclear whether this roughly 5% increase in genome-wide significant findings is the best that could be obtained, and additionally whether the increase comes at the cost of false negatives for true positives that lack similar annotations.

Other groups, including our own, have developed methods that incorporate priors based on patterns learned from the data (Chanda et al., 2013; Huang et al., 2011; Zhan et al., 2018). These patterns may include multiple independent effects found within a single genes, or patterns of pleiotropic variants that contribute to a shared subset of traits in a multi-phenotype data set. While these methods have value in providing a clearer view of genetic architecture than available through univariate tests, the number of new significant findings has been small (Sveinbjornsson et al., 2016; Meyer et al., 2018).

Still other methods introduce prior distributions for model parameters, or equivalently regularizations, which implicitly define a prior favoring candidate variants with the largest observed effects. These methods have usually not been used in practice for GWAS because the computational expense has not been justified by improved results.

In this paper, we use theoretical models and derivations to investigate into the dependency of power on population size and incorporating priors. We consider two types of priors: hard prior, which is an idealized prior that only a fraction of total hypotheses are tested; soft prior, for which all hypotheses are divided into two classes, and a higher prior value is given to the favored class which is believed to be enriched with true associations. For hard prior, we proved analytically that the dependence of power on population size is linear, whereas the dependence on prior strength is logarithmic, which indicates the importance of having larger population size over bigger prior strength when doing association tests. For soft prior, we provide numerically exact results showing that the power gains for the favored class are large only for limited circumstances; the gains for the favored class imply power loss for the non-favored class, and the average power gain considering both classes is only 5-10%. These gains require exact knowledge of the true priors; in practice, gains with estimated priors should be smaller.

3.2 Methods

3.2.1 Hypothesis testing

We consider tests of association between a feature of the data, x, and an observed phenotype or response variable, y, assumed to be scalars for simplicity. For a population of size N, these are aggregated into vectors $\mathbf{x}$ and $\mathbf{y}$. An association test compares a null model M_0, to an alternative, M_1, which for a linear model takes the form

$$M_0 : y \sim \mathrm{Norm}(\mu_0, \sigma_0^2);$$

$$M_1 : y \sim \mathrm{Norm}(\mu_1 + \beta x, \sigma_1^2).$$

One such M_1 exists for each possible feature to be tested. With A total possible alternatives to be tested, these could be denoted $\{M_a\}$, $a \in \{1, 2, \ldots, A\}$. We consider one such alternative at a time and for simplicity denote it M_1. Model parameters are $\Theta_0 = \{\mu_0, \sigma_0^2\}$ for the null model and $\Theta_1 = \{\mu_1, \sigma_1^2, \beta\}$ for the alternative model. These models correspond to a null hypothesis H_0 and alternative hypothesis H_1,

$$H_0 : \beta = 0;$$
$$H_1 : \beta \neq 0.$$

For nested models, the hypothesis test is usually performed by a likelihood ratio test or its equivalent. Denote the maximum likelihood parameters as $\widehat{\Theta}_0$ and $\widehat{\Theta}_1$, and assume independence of the model and data. A test statistic τ is defined as

$$\begin{aligned} \tau &= 2 \ln \frac{\Pr(\mathbf{y}|\mathbf{x}, \widehat{\Theta}_1)\Pr(M_1)}{\Pr(\mathbf{y}|\mathbf{x}, \widehat{\Theta}_0)\Pr(M_0)} \\ &= q^2 + 2\ln[\Pr(M_1)/\Pr(M_0)]. \end{aligned} \tag{3.2.1}$$

According to Wilks' Theorem, under the null hypothesis, q^2 is a random variable distributed as χ_1^2, or more generally as a χ_d^2 random variable where the null model is nested inside an alternative model with d additional parameters (Wilks, 1938). Under the alternative hypothesis, q^2 is distributed as a non-central χ^2 with non-centrality parameter q_1^2,

$$q_1^2 = NR^2/(1 - R^2), \tag{3.2.2}$$

where R^2 is the fraction of variance explained by the alternative hypothesis, and $1 - R^2$ is the residual fraction of variance.

For a conventional test, the prior $\Pr(M)$ is identical for the null and each alternative; it does not contribute to the test statistic. To control the type I error (false-positive rate) at family-wise error rate FWER α, the Bonferroni method requires a single-test p-value of α/A for A total tests. Define the quantile of the uniform normal distribution corresponding to a two-tailed test at this stringency z_I. More formally, if $\Phi(z)$ is the cumulative lower tail probability distribution for standard normal random variable z, then $\Phi(-z_I) = \alpha/2A$. For true effect q_1, the power is $\Phi(|q_1| - z_I)$, or equivalently

$$(z_I - z_{II})^2 = \frac{NR^2}{1 - R^2}. \tag{3.2.3}$$

This key expression relates the type I error (false-positive rate), the type II error (false-negative rate or complement of power), the population size N, and the effect size R^2.

3.2.2 Hard prior

A hard prior is an idealized prior in which only hypotheses corresponding to a faction $1/S$ of the total are tested. Larger S corresponds to a stronger prior. For

20,000 gene-based tests, testing 10% of the total corresponds to $S = 10$, and testing 20 genes corresponds to $S = 1000$. Realistically, priors stronger than $S = 100$, corresponding to 200 genes tested, are unlikely.

The effect of a hard prior is to reduce the multiple-testing burden. To maintain FWER α, each two-tailed test is performed at stringency $S\alpha/2A$ rather than $\alpha/2A$. This reduces the quantile z_I required for significance and increases the power to detect an association with a smaller effect R^2. Equivalently, Eq. 3.2.3 can be solved for R^2 to calculate the critical effect size to achieve desired power at stated type I error,

$$R^2 = \frac{(z_I - z_{II})^2}{N + (z_I - z_{II})^2}. \tag{3.2.4}$$

The effect of a hard prior on z_I may also be estimated analytically. A steepest descents approximation relates the quantile $z > 0$ to its upper-tail area ϵ,

$$\begin{aligned}
\epsilon &= (2\pi)^{-1/2} \int_z^\infty du e^{-u^2/2} \\
&= (2\pi)^{-1/2} e^{-z^2/2} \int_z^\infty du e^{-(u+z)(u-z)/2} \\
&\approx (2\pi)^{-1/2} e^{-z^2/2} \int_z^\infty du e^{-z(u-z)} \\
&= \frac{1}{\sqrt{2\pi} z} e^{-z^2/2}.
\end{aligned}$$

Equivalently,

$$z^2 \approx -2 \ln[\sqrt{2\pi} z \epsilon].$$

In terms of the quantile z_I for prior strength S and a two-tailed test, we have approximately

$$z_I^2 \approx -2\ln[\sqrt{2\pi} z_I S\alpha/A]. \tag{3.2.5}$$

Define ζ as the value of z_I for no prior, $S = 1$, with $\Phi(-\zeta) = \alpha/2A$ and

$$\zeta^2 \approx -2\ln(\sqrt{2\pi}\zeta\alpha/A). \tag{3.2.6}$$

For GWAS with a p-value threshold of 5×10^{-8}, $\zeta = 5.45$ and $\zeta^2 = 29.7$. Because the dependence of Eq. 3.2.5 on $\ln z$ is weak, we replace $\ln z$ with $\ln\zeta$,

$$z_I^2 \approx -2\ln[\sqrt{2\pi}\zeta S\alpha/A] \approx \zeta^2 - 2\ln S = \zeta^2(1 - 2\zeta^{-2}\ln S).$$

Keeping terms of order $1/\zeta$,

$$\begin{aligned}
z_I &\approx \zeta(1 - \zeta^{-2}\ln S)\\
z_I - z_{II} &\approx \zeta - z_{II} - \zeta^{-1}\ln S\\
(z_I - z_{II})^2 &\approx (\zeta - z_{II})^2 - \frac{2(\zeta - z_{II})}{\zeta}\ln S\\
&= (\zeta - z_{II})^2\left[1 - \frac{2}{\zeta(\zeta - z_{II})}\ln S\right].
\end{aligned}$$

According to Eq. 3.2.3, the critical effect size depends only on the ratio $(z_I - z_{II})^2/N$. Consider two scenarios with equal critical effect size, one with population size N_1 and prior strength S_1, and the second with population size N_2 and prior strength S_2. For these to have equal critical effect size,

$$(\zeta - z_{II})^2 \left[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S_1\right] / N_1 \approx (\zeta - z_{II})^2 \left[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S_2\right] / N_2.$$

Cancelling constant terms $\zeta - z_{II}$ and noting that $2\zeta^{-1}(\zeta - z_{II}) \ln S$ is small,

$$\begin{aligned}
\frac{N_1}{N_2} &\approx \left[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S_1\right] / \left[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S_2\right] \\
&\approx \left[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S_1\right] \times \left[1 + \frac{2}{\zeta(\zeta - z_{II})} \ln S_2\right] \\
&\approx 1 + \frac{2}{\zeta(\zeta - z_{II})} \ln \frac{S_2}{S_1}.
\end{aligned}$$

The dependence on population size is linear, whereas the dependence on prior strength is logarithmic. Equivalently, population size is exponentially more important that prior strength. Again for GWAS with z_{II} selected for 80% power, $\zeta(\zeta - z_{II})/2 = 17.15$, and only a small fractional population increase is required to obtain the equivalent power increase for a strong prior. An extremely strong prior with $S_2 = 1000$, with effectively only 20 genes selected for testing, can be matched by a population increase of about 40%.

Contours of N and S with equal critical effect size can be estimated by returning to the approximate result

$$NR^2/(1-R^2) \approx (\zeta - z_{II})^2[1 - \frac{2}{\zeta(\zeta - z_{II})} \ln S].$$

Noting that for small ϵ, $1 + \epsilon \ln S \approx S^{\epsilon}$, contours are given by

$$NS^{2/\zeta(\zeta - z_{II})} \approx (\zeta - z_{II})^2 R^2/(1-R^2). \tag{3.2.7}$$

On a log-log plot of $\log S$ versus $\log N$, these contours would have steep negative slope equal to $-\zeta(\zeta - z_{II})/2$.

3.2.3 Soft prior

Soft priors are incorporated into association analysis such that sequence variants like loss-of-function and missense variants, which are more likely to affect protein function and therefore more likely to be causative, are given higher prior belief to have true signals before data was analyzed. For simulation, this is done by first dividing all possible associations into two classes, a favored class F and a non-favored class NF, and assuming true associations are enriched in the favored class and depleted in the non-favored class. For Eq. 3.2.1, if we normalize $\Pr(M_0) = 1$, the expression could be simplified into the following form:

$$\tau = q^2 + 2\ln[\Pr(M_1)] \tag{3.2.8}$$

Denote the model for variants in the favored class as M_F and in the non-favored class as M_{NF}, if priors for the two classes $Pr(M_F)$ and $Pr(M_{NF})$ are known exactly, the probability distribution for the test statistic will now depend on the classes, with their individual test statistics being:

$$\tau_F = q^2 + 2\log\left[\Pr(M_F)\right]$$

$$\tau_{NF} = q^2 + 2\log\left[\Pr(M_{NF})\right]$$

Based on assumptions given above, we are able to simulate the change of power in detecting real effects after incorporating the two classes. For simulation, we first fixed the association effective size corresponding to 50% power to detect a true association in a GWAS study at genome-wide significance ($p = 5 \times 10^{-8}$) assuming 1 million effective tests. Using the same notation as above, this is equivalent to solving for first the critical value τ_c such that

$$\Phi(-\tau_c) = 5 \times 10^{-8}/2 = 2.5 \times 10^{-8}$$

then solve for true effect q_c such that $\Phi(|q_c| - \tau_c) = 0.5$, which gives us $|q_c| = \tau_c$. To avoid calculation on both tails of the standard normal distribution, the expression could be simplified by introducing $\Psi(z^2; q^2)$ to denote $\Pr(t > z^2)$ where t follows a $1df$ non-central χ^2 distribution with non-centrality parameter q^2. The relationship between Φ and Ψ thus satisfies

$$2 \times \Phi(z - |q|) = \Psi(z^2; q^2)$$

for $z < |q|$ and

$$2 \times \Phi(|q| - z) = \Psi(z^2; q^2)$$

for $z \geq |q|$.

Power to detect SNPs with true associations for two classes combined could then be calculated as a function of two variables:

- S, power strength, which is defined as inverse of the fraction of variants in the favored class;
- $[\Pr(M_F)/\Pr(M_{NF})]$ the relative priors of the two classes.

The exact steps of simulation are as following:

- For a pair of values for fraction of variants in the favored class 1/S and prior enrichment fold-enrichment $[\Pr(M_F)/\Pr(M_{NF})]$, the critical threshold for genome-wide significance τ_c' could be solved using the following equation:

  $$(1 - 1/S)\Pr(q^2 > \tau_c') + (1/S)\Pr(q^2 + 2\ln\left[\frac{\Pr(M_F)}{\Pr(M_{NF})}\right] > \tau_c') = 5 \times 10^{-8}$$

  With q^2 following an $1df$ χ^2 distribution, with the denotations defined above, this equation simplifies to:

$$(1-1/S)\Psi(\tau_c';0)+(1/S)\Psi(\tau_c'-2\ln\left[\frac{\Pr(M_F)}{\Pr(M_{NF})}\right];0)=5\times10^{-8} \tag{3.2.9}$$

- Now with the critical threshold τ_c' and the non-centrality parameter q_c calculated above, power for the favored class could be calculated as:

$$power(M_F)=\Psi(\tau_c'-2\ln\left[\frac{\Pr(M_F)}{\Pr(M_{NF})}\right];q_c^2) \tag{3.2.10}$$

And the power for the non-favored class could calculated as:

$$power(M_{NF})=\Psi(\tau_c';q_c^2) \tag{3.2.11}$$

- The average power for true associations could be calculated as:

$$\begin{aligned} power(Avg) = & \frac{\Pr(M_F)}{\Pr(M_F)+(S-1)\Pr(M_{NF})}\times power(M_F) \\ & +\frac{(S-1)\Pr(M_{NF})}{\Pr(M_F)+(S-1)\Pr(M_{NF})}\times power(M_{NF}) \end{aligned} \tag{3.2.12}$$

- Population size fraction increase to achieve the same average power could then be calculated as $N_1/N_2=q_c'/q_c$, where q_c' correspond to the same average prior without using any prior. Here N_2 = population size to achieve the specific power and 0.05 FWER using a prior, and N_1=population size to achieve the same power and 0.05 FWER without a prior. This is similar to the exploration between S and N in the hard prior case.

3.3 Results

3.3.1 Hard prior

Figure showing contour lines for GWAS.

Figure showing left panel as (population factor equivalent) vs (prior strength), right panel as (exponent) vs (prior strength), demonstrating that slope estimate from analytical expression is reasonably good.

Fig. 3.1 shows contours for critical R^2 for $p = 5 \times 10^{-8}$ at power $= 0.8$ as a function of prior strength and population size. As could be observed, the color corresponding to critical R^2 changes rapidly as population size changes, and doesn't change much as a function of prior strength. This indicates that power is much more sensitive to population size compared to using a hard prior, namely restricting tests to a subset of variants. On the log-log scale, given a fixed value of R^2, the prior strength and population size exhibits a clear linear pattern, which leads to the derivation on the Hard prior part in the Methods section, and yielding a slope of 17.15 at 80% power. Fig. 3.2 shows the analytical solution of the relationship between population size and prior strength. The left panel denotes the population increase ratio versus prior strength increase to achieve the same increase in power. If we have a prior strength of 100 for example, which correspond to testing 1/100 of all variants, we could get the same power increase by increasing the population from N to fN where f is the factor increase. Reading from the figure, this would be a factor about 1.35, which correspond to a 33% increase in the cohort size. At the very extreme, a prior strength of 1000, which corresponds to testing SNPs in only 20 genes, will only do as well as increasing the cohort size by 70%. Panel on the right shows the linear relationship between population size exponent and prior strength for a fixed effect size, which quantifies the linear relationship as described in Eq. 3.2.2. Both figures provide numerically exact result of the relationship between population size and prior strength, which further strengthen the conclusion that population size is a much more important factor to gain power than incorporating priors.

3.3.2 Soft prior

Figure showing left panel as (power increase for favored class), right panel as (power increase for unfavored class) as function of enrichment vs fraction of favored class. Maybe use 50% power for this one with no prior?

Figure showing average power as a function of enrichment vs fraction of favored class. Again use 50% power with no prior?

Figure showing population size fraction increase to achieve the same average power. x-axis is fraction of favored class. y-axis is enrichment. values are (N1/N2) where N2 = population size to achieve the specified power and FWER using a prior, N1 = population size to achieve power and FWER without a prior.

3.3.2.1 Power for the two classes

Fig. 3.3 shows power for the two classes at different fold-enrichment for favored class versus prior strength. For power of the favored class, when prior strength is small, corresponding to a large fraction of variants in the favored class, power is not sensitive to the fold enrichment. For example, when $S = 2$, corresponding to the favored class consisting around 50% of the total tested variants, power boost for the favored class is fixed at around 2% to 4% regardless of fold enrichment. This is because, a large favored class fraction is essentially equivalent to a less well-defined subset of variants, which usually fails to provide much valuable information regarding prior beliefs. Therefore, giving the big favored class a higher prior value only results in the decreased power for the non-favored class, as is shown in the right panel of Fig. 3.3.

As S increases, which corresponds to decreasing the fraction of favored class, power becomes more sensitive to fold enrichment, and assigning it with a bigger prior enhances its power. Specifically, a prior 2 to 5 folds as big as the non-favored class prior gives a 5% to 10% power boost. This power increases even more as the fold-enrichment enhances, as long as the fraction is fixed at the same level. This is because, as the favored class gets smaller, the subset of variants becomes more informative, thus giving the class higher prior greatly boosts its power. From the calculation perspective, as S becomes bigger, τ_c' term in Eq. 3.2.9 is mostly determined by the non-favored class and remains mostly unchanged; thus little change to the fold enrichment would result in big difference of the power of favored class.

Also note that, since the overall power is fixed at 50%, gain in power for the favored class implies a loss of power for the non-favored class, which corresponds to the "no-free-lunch theorem". Power for the non-favored class remains around 50% when fraction of favored class is low or fold enrichment is low, as in these two situations the impact of the favored class is small; its power decreases, yielding more power for the favored class, when both fold enrichment is high and fraction of favored class is high.

3.3.2.2 Average power for combining both classes

Average power for combining both classes is shown in Fig. 3.4. For same fold-enrichment and class fraction, the average power is relatively smaller than power for the favored class alone, and higher than the non-favored class alone. This is because gain for power of the favored class leads to loss for the non-favored class, as discussed above, thus a combination of the two will result a value of power in the middle.

Shape of the contour is determined by weights of the two classes: for small prior strength and large fold enrichment, shape of contour is more similar to the favored class in Fig. 3.3; for large prior strength and small fold enrichment, shape is similar to the non-favored class. This explains the curve which tilts up towards large prior strength, as when fraction for favored class is small, the power is mostly determined by the un-favored class. Fig. 3.5 shows the population size fraction increase in order to achieve the same power increase. As could be observed, the maximum population increase is 1.3 fold to obtain the maximum power gain fulfilled through incorporating a prior, further strengthening the conclusion from the hard prior part that, population size is of a more crucial factor compared to prior incorporation as for association involving large number of hypothesis testings.

3.4 Discussion

Despite the efforts on developing methods that incorporate priors into association hypothesis tests, traditional unbiased univariate tests combined with Bonferroni correction to control for FWER remains the rule of thumb method to test for associations. In this paper, we exploited the relationship of power to detect true associations on increasing study size and incorporating priors. Two scenarios were considered in this study: hard prior, for which only a fraction of all variants are tested to lower the burden coming from multiple testing; soft prior, for which a fraction of variants are given a higher belief a prior to doing the association analysis. For hard prior, the dependence of heritability on population size and prior strength was analytically derived, and it was proved that the dependence on population size is linear, whereas the dependence on prior strength is logarithmic. Soft prior was able to boost power with very specific requirement on class fraction and fold-enrichment, and even so, its maximum boost of power could be achieved by increasing population size by approximately 30%. For both scenarios, it was concluded that increasing population size is a better strategy to boost power compared to incorporating priors. With recent developments in high throughput biology, immense amount of data is being generated, making improving power through increasing population size possible; in the meantime, a lack of prior-based methods with extraordinary performance on association tests has been observed, which further strengthen the favor of population size from an practical perspective. These results give valuable insights into what strategy should be taken towards future directions for establishing associations between biological variants and traits.

The usefulness of having larger sample size is not restricted to association tests in the biological field. Huge success has been achieved in the application of machine learning and deep learning into image analysis, nature language processing fields and etc. While big credit has been given to the design and implementation of sophisticated learning structures like convolution neural networks, recurrent neural networks and etc., from results in this project, the role that the tremendous amount of data for training these networks might have been well under-estimated.

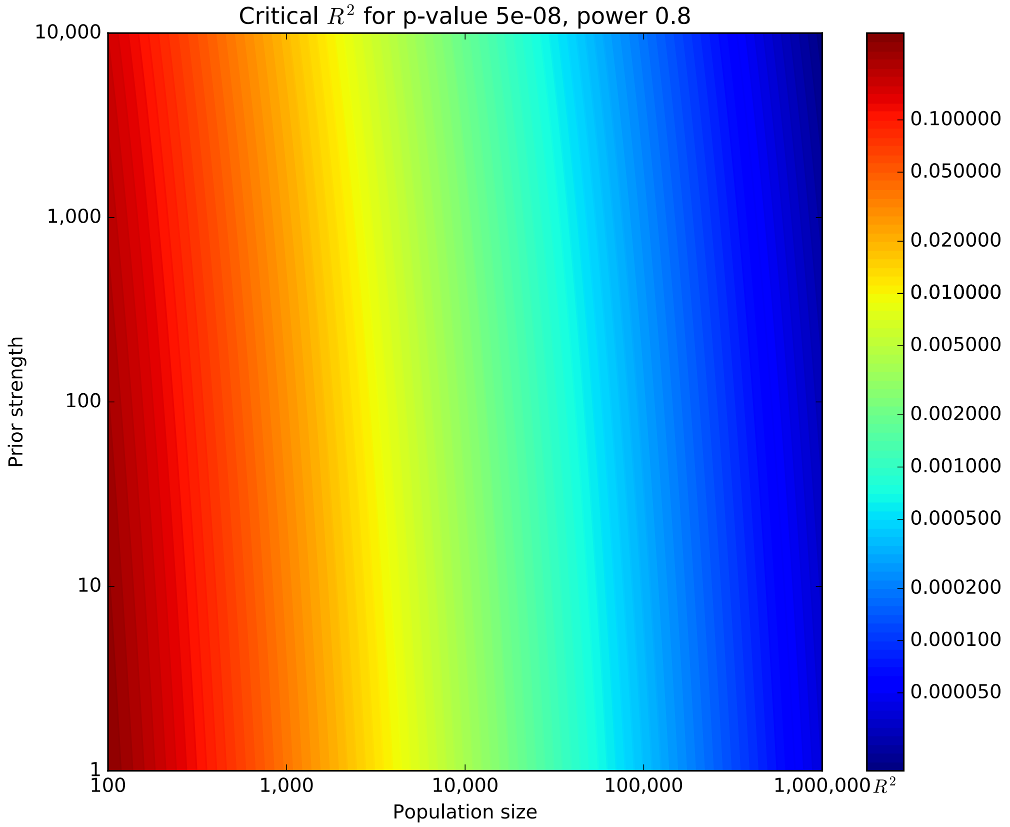

Figure 3.1: **Contour plot for critical R^2 for $p = 5 \times 10^{-8}$ at power $= 0.8$ as a function of prior strength and population size.** The color corresponding to critical value R^2 changes rapidly as population size changes, and doesn't change much as a function of prior strength, indicating that power is much more sensitive to population size compared to restricting tests to a subset of variants. Relationship between prior strength and population size appears to follow a linear relationship on the log-log scale, and the slope of the linear relationship given a fixed value of R^2 indicates the relative importance of the two factors.

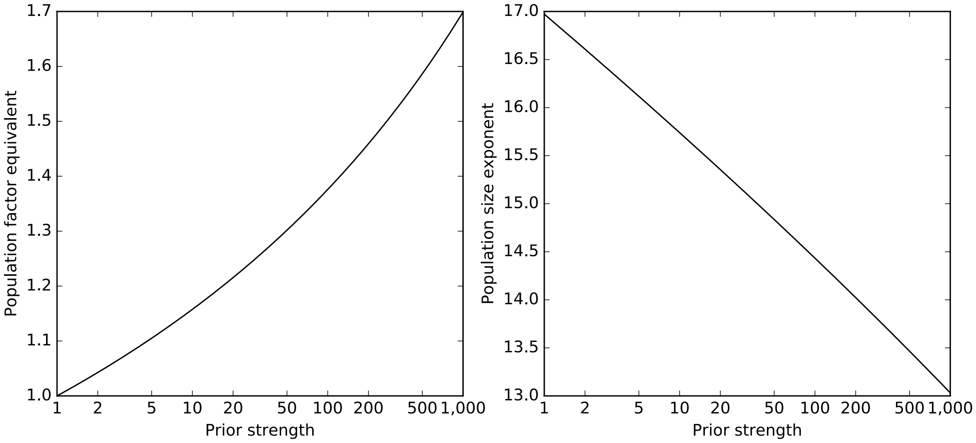

Figure 3.2: **Population size and prior strength to achieve same power.** X-axis denotes the prior strength S; y-axis of the left panel denotes population fraction increase; y-axis of the right panel denotes population size exponent for a fixed effect value. The left panel shows the population size and prior strength to achieve the same power; a prior strength at 1000 will have equivalent power if the cohort size increases by 70%. The right panel shows the relation between population size exponent and prior strength given a fixed effect size. Both figures shows numerically the linear relationship between population size and logarithmic of prior strength.

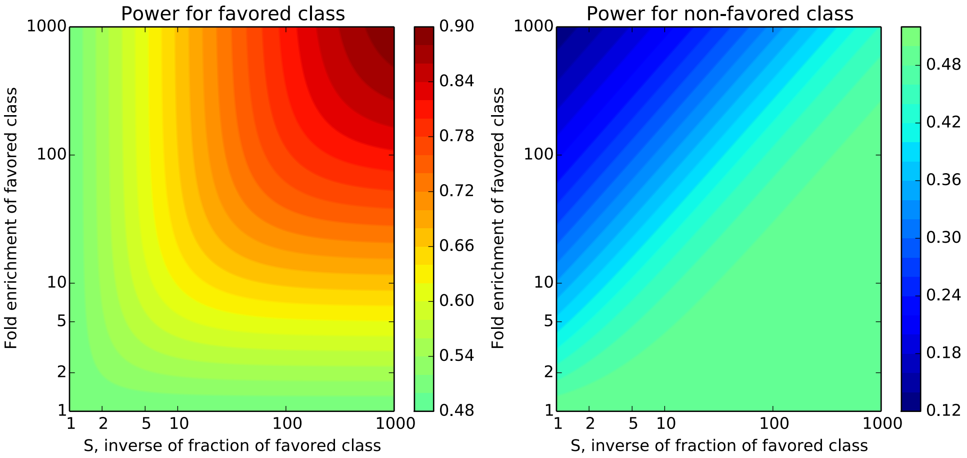

Figure 3.3: **Contour plot of power for the favored class and non-favored class at $p = 5 \times 10^{-8}$ threshold for 50% power.** The left panel shows power for the favored class, and the right panel shows power for the non-favored class. X-axis denotes prior strength S, which is equal to inverse of fraction of SNPs in the favored class; larger S value denotes smaller group of favored class and a more focused subset of variants with higher prior. Y-axis denotes fold enrichment of the favored class $\Pr(M_F)/\Pr(M_{NF})$. For power for the favored class, it could be observed that when prior strength is small, the power is insensitive to prior fold enrichment; this is because large fraction for the favored class is essentially equivalent to a less well-defined subset of variants, thus effect of prior enrichment becomes less obvious. When the favored class is more well-defined, corresponding to larges S values and smaller fractions of favored class, the effect of incorporating prior becomes more obvious; this is reflected by the power gain at large S values, and the power increases with higher fold-enrichment. Power of the non-favored class remains at around 50% when fraction of favored class is small or fold-enrichment is small, because small fold enrichment or small fraction of favored subset is unlikely to make big impact to the non-favored class; when the favored class is given a large prior and consists of large proportion of the total variants, the non-favored class begins to loose power.

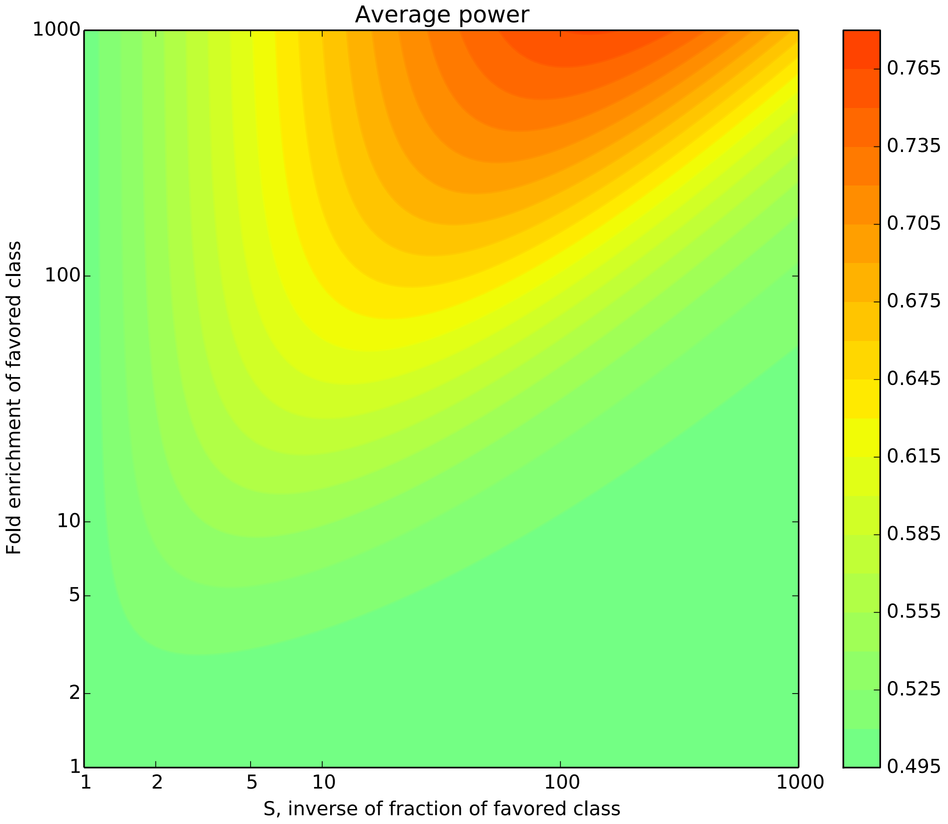

Figure 3.4: **Average power for both classes combined at $p = 5\times10^{-8}$ threshold for 50% power.** Axes are defined in the same way as for Fig.3.3. Average power is smaller than power for the favored class and larger than power for the non-favored class given the same prior strength and fold enrichment, as a balance between gain of power for the favored class and loss of power for the non-favored class. Shape of the contour is determined by weights of the two different classes: for small prior strength and large fold enrichment, shape of contour is more similar to the favored class in Fig. 3.3; for large prior strength and small fold enrichment, shape is similar to the non-favored class.

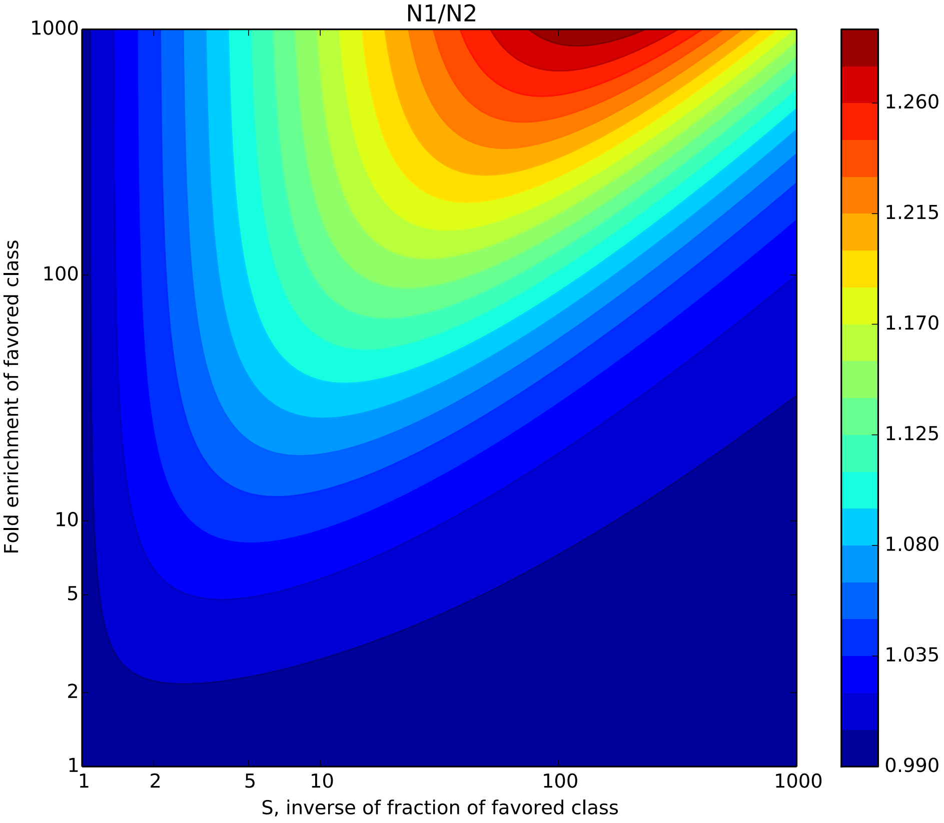

Figure 3.5: **Population size fraction increase to achieve the same average power for at** 5×10^{-8}**.** Axis are defined in the same way as for Fig. 3.3. Values are N_1/N_2 where N_1 is the population size without prior, and N_2 is population size to achieve the same power using prior. The maximum population size fraction increase is 1.3 fold to obtain the maximum power gain fulfilled through incorporating a prior, further prove the point that population size is of a more crucial factor for association testing.

Chapter 4

Discussion

CHAPTER 4. DISCUSSION

Two main projects were presented in this thesis: development and assessment of SAPPHO, a Bayesian-based GWAS method aimed at detecting pleiotropic associations; an analytical simulation that explores the dependency of Bayesian-based GWAS method on different factors including prior strength, population size, and fold-enrichment.

Pleiotropy is wide-spread. At 5×10^{-8} genome-wide significance, $\sim 2,000$ out of 17,607 SNPs were pleiotropic, consisting about 10-15% of the population(Welter et al., 2014). The number is consistent with our analysis using univariate test on ECG and cancer phenotypes: for CHARGE ECG meta-analysis results, 11 out of 107 loci were pleiotropic; for UK Biobank cancer phenotypes, 4 out of 51 loci were pleiotropic. However, if more sophisticated pleiotropy detection methods, like SAPPHO, were applied to the same set of data, it could be observed that the pleiotropic effect using traditional univariate tests could be potentially under-estimated: for CHARGE results, 15 loci were detected to have additional associations; and for UK Biobank cancer phenotypes, 2 loci were associated with more phenotypes. Although part of the reason that more pleiotropic effects were observed might be because of the usage of 0.05 FDR instead of a more stringent control over FWER, the observation on two real data sets still makes us wondering if more sophisticated methods were used to test on all associations from GWAS catalog, how many more pleiotropic effects could be observed. Moreover, if instead of human diseases and traits, such approaches are used to explore a broader range of phenotypes, for example gene expressions and expression quantitative trait loci (eQTLs), these methods could potentially help us to better understand the underlying cellular biological processes, and build a more completed gene regulatory network.

Despite the efforts spent on developing SAPPHO for detecting pleiotropy, we found that the performance of SAPPHO was not as great as expected. For example, it was shown through simulations that SAPPHO performed the best only when the association patterns were simple: either the association is with only one single phenotype, or all phenotypes are associated; when associations are for a subset of all phenotypes, SAPPHO sometimes have difficulty in beating other non-Bayesian pleiotropic methods, for example MultiPhen. On real data, application of SAPPHO on ARIC and UK Biobank cancer phenotypes detected only 1 and 2 additional pleiotropic effects, respectively, at 0.05 FDR. Given that for one set of simulation, SAPPHO performed best when it is fed with the true model, we inferred that a more sophisticated searching strategy could potentially improve the performance of SAPPHO. Therefore, we tried another searching strategy, called SAPPHO local search, to see whether it could improve the performance of SAPPHO.

The idea of SAPPHO local search is simple: with the SAPPHO scoring function, the association that increases the score the most is added to the model; then, instead of adding the next association which increases the score the most, we look at all other associations for the same SNP added in the previous step, and add any other association to the model if the score increase is greater than 0; associations are added until none of them could increase the score any more. This strategy was tried on UK Biobank cancer phenotypes and results were compared with those obtained from greedy-forward heuristic searching. It was observed that the two results were exactly the same, except that the order of some associations added to the model was changed. By looking at the marginal p-values of all associations for each SNP, we concluded that the reason this local search failed is because: in one locus, a SNP with best p-value for one phenotype is unlikely to have a great p-value for another phenotype, and the pleiotropic SNPs usually have relatively good p-value for all phenotypes, but not the best for any one of them. Therefore, given the parsimonious behavior of SAPPHO, it tends to find the best SNP for each phenotype, instead of finding one SNPs that have relative good p-value for all phenotypes. Given this, another possible searching strategy that could potentially work better is as following: at each step, for each SNP we keep adding associations until SAPPHO score does not improve any more, and then for the current step, we add the SNP and its associations that increases SAPPHO score the most. Given that the implementation of this strategy involves lots of re-coding of the current version of SAPPHO, this part of the work is left open as a future work for people to explore.

Another future work for SAPPHO is with score correction. Since SAPPHO uses an heuristic selection process, the score itself does not follow any empirical distribution. Because of this, permutations are needed to correct for the SAPPHO scores, in order to get p-values and control for false discovery rate or family-wise error rate. Permutations are easily done when primary data are present; with summary statistics, how to do permutations has been less intuitive. Currently, the way permutations are done with summary statistics has been to sample individual genotypes from a reasonable population, and run permutations using the sampled primary data. The sampling and running univariate tests has been very computationally intensive: for UKB cancer phenotypes, running univariate tests alone took about 3 weeks to finish. An easier and much faster way to run such permutations using summary statistics, as suggested in a previous work (Chanda et al., 2013), is to directly sample z-scores for all phenotypes from haplotype with the corresponding LD structures. Yet again due to the efforts in implementation of such method, this part of the work is also left open for future people to explore.

The performance of SAPPHO on detecting pleiotropic effects, as well as GWiS on single phenotype, was not as satisfying as expected, despite the efforts devoted to developing these methods. Therefore, this analytical study was performed to answer the question “is Bayesian a good idea for GWAS”, the answer to which essentially determines whether we should still focus on GWAS method improvement. Two conclusions drawn from the simulations were the most important: 1. population size is exponentially more important than prior; 2. on average the power gain obtained through prior incorporation is only around 5-10%. The former conclusion analytically explained why in practice progress on finding GWAS hits has been majorly coming from increasing sample size; and the later conclusion provides numerical explanation on the observed gain of power obtained through more advanced methods, including GWiS and SAPPHO. Therefore, the insight that we could gain from the simulations is simple: since power gain obtained from any method improvement would not beat the effect of doubling the sample size, working on GWAS method development might probably not be the best idea from now on. The simulation also indirectly indicates why machine learning methods based on big data, like deep learning, has been a success in the recent years: it might not have much to do with the details of the method; the only thing which is important is the tremendous amount of data, and a model complicated enough so that the data wouldn't be under-fitted.

Bibliography

1000GenomesProjectConsortium, Abecasis, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R., Hurles, M., and McVean, G. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

AmericanCancerSociety (2017). Cancer facts and figures 2017. *American Cancer Society*.

ARICInvestigators (1989). The atherosclerosis risk in communities (aric) study: design and objectives. *Am. J. Epidemiol.*, 129(4):687–702.

Arking, D., Pfeufer, A., Post, W., Kao, W., Newton-Cheh, C., Ikeda, M., West, K., Kashuk, C., Akyol, M., Perz, S., and et al. (2006). A common genetic variant in the nos1 regulator nos1ap modulates cardiac repolarization. *Nat. Genet.*, 38(6):644–651.

Arking, D., Pulit, S., Crotti, L., van der Harst, P., Munroe, P., Koopmann, T., Sotoodehnia, N., Rossin, E., Morley, M., Wang, X., and et al. (2014). Genetic association study of qt interval highlights role for calcium signaling pathyways in myocardial repolarization. *Nature Genetics*, 46(8):826–836.

Bellenguez, C., Strange, A., Freeman, C., and et al. (2012). A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioninformatics*, 28(1):134–135.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bycroft, C., Freeman, C., Petkova, D., Band, G., and et al. (2017). Genome-wide genetic data on 500,000 uk biobank participants. *bioRxiv*, doi: https://doi.org/10.1101/166298.

Chanda, P., Huang, H., Arking, D., and Bader, J. (2013). Fast association tests for genes with fast. *PLoS ONE*, 8(7):e68585.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *Inter Journal complex systems.*

Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., and Manolio, T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases. *Proceedings of the National Academy of Sciences*, 106(23):9352–9367.

Howie, B., Donelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *Genes, Genomics, Genetics*, 1(6):457–470.

Huang, H., Chanda, P., Alonso, A., Bader, J., and Arking, D. (2011). Gene-based tests of association. *PLoS Genet.*, 7(7):e1002177.

Kim, S. and Xing, E. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genetics*, 5(8):e1000587.

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32:9–19.

Liberzon, A. and et al. (2011). Molecular signature database (msigdb) 3.0. *Bioinformatics*, 27(12):1798–1812.

Liu, J., Mcrae, A., Nyholt, D., Medland, S., Wray, N., Brown, K., Investigators, A., Hayward, N., Montgomery, G., Visscher, P., Martin, N., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, 87(1):139–145.

Meyer, V., Francesco, P., Oliver, S., and Ewan, B. (2018). Limmbo: a simple, scalable approach for linear mixed models in high-dimensional genetic association studeis. *bioRxiv*, doi:10.1101/255497.

Mucci, L., Hjelmborg, J., Harris, J., Czene, K., Havelick, D., and et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. *JAMA*, 315(1):68–76.

Newton-Cheh, C., Eijgelsheim, M., Rice, K., de Bakker, P., Yin, X., Estrada, K., Bis, J., Marciante, K., Rivadeneira, F., Noseworthy, P., and et al. (2009). Common variants at ten loci influence qt interval duration in qtgen study. *Nat. Genet.*, 41(4):399–406.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., and et al. (2016). Hapltype estimation for biobank-scale data sets. *Nature Genetics*, 46(7):817–820.

O'Reilly, P., Hoggart, C., Pomyen, Y., Calboli, F., Elliott, P., Jarvelin, M., and Coin, L. (2012). Multiphen: Joint model of multiple phenotypes can increase discovery in gwas. *Plos One*, 7(5):e34861.

Petersen, A., Spratt, J., and Tintle, N. (2013). Incorporation prior knowledge to increase the power of genome-wide association studies. *Methods in Molecular Biology*, 1019:519–541.

Pfeufer, A., Sanna, S., D.E., A., Muller, M., Gateva, V., Fuchsberger, C., Ehret, G., Orru, M., Pattaro, C., Kottgen, A., and et al. (2009). Comman variants at ten loci modulate the qt inteval duration in the qtscd study. *Nat. Genet.*, 41:407–414.

Pfeufer, A., van Noord, C., Marciante, K., Arking, D., Larson, M., Smith, A., Tarasov, K., Muller, M., Sotoodehnia, N., Sinner, M., and et al. (2010). Genome-wide association study of pr interval. *Nat. Genet.*, 42(2):153–159.

Pickrell, J. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94(4):559–573.

Pontes, B., Giraldez, R., and Aguilar-Ruiz, J. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180.

Schork, A., Thompson, W., Pham, P., and et al. (2013). All snps are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLOS Genetics*, 9(4):e1003449.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Silva, C., Kors, J., Amin, N., Dehghan, A., Witteman, J., Willemsen, R., Oostra, B., van Duijn, C., and Iassacs, A. (2015). Heritabilities, proportions of heritability explained by gwas findings, and implications of cross-phenotype effects on pr interval. *Human Genetics*, 134:1211–1219.

Sotoodehnia, N., Isaacs, A., de Bakker, P., Dorr, M., Newton-Cheh, C., Nolte, I., van der Harst, P., Muller, M., Eijgelsheim, M., Alongso, A., and et al. (2010). Common variants in 22 loci are associated with qrs duration and cardiac ventricular conduction. *Nat. Genet.*, 42(12):1068–1076.

Sud, A., Kinnerseley, B., and Houlston, R. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nature Review Cancer*, 17(11):692–704.

Sveinbjornsson, G., Alberechtsen, A., Zink, F., Gudjonsson, S., Oddson, A., and et al. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetcis*, 48(3):314–317.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkingson, H. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Research*, 42:D1001–D1006.

Wild, P., Felix, J., Scheillert, A., Teumer, A., Chen, M., Leening, M., Volker, U., Grobmann, V., Brody, J., Irvin, M., Brody, J., Irvin, M., and et al. (2017). Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *The Journal of Clinical Investigation*, 127(5):1798–1812.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypothesis. *The Annals of Methmatical Statistics*, 9:60–62.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89:82–93.

Wu, T., Chen, Y., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Wu, Y., Graff, R., Passarelli, M., Hoffman, J., Ziv, E., and et al. (2017). Identification of pleiotropic cancer susceptibility variants from genome-wide association studeis reveals functional characters. *Cancer Epidemiology Biomarkers and Prevention*, 27(1):75–85.

Zhan, J., Arking, D., and Bader, J. (2018). Discovering patterns of pleiotropy in genome-wide association studies. *bioRxiv*, doi:10.1101/273540.

Zhu, X., Feng, T., Tayo, B., Liang, J., Young, J., Franceschini, N., Smith, J., Yanek, L., Sun, Y., Edwards, T., Chen, W., Nalls, M., Fox, E., Sale, M., Bottinger, E., Rotimi, C., Consortium, T. C. B., Liu, Y., McKnight, B., Liu, K., Arnet, D., Chakravati, A., Cooper, R., and Redline, S. (2015). Meta-analysis of corelated traits via summary statistics from gwass with an application in hypertension. *American Journal of Human Genetics*, 96(1):21–36.

CURRICULUM VITAE

Jianan Zhan

Email: *jiananzhan@gmail.com* Cell: *(443)-721-6168*
Address: *MRB 311, 733 N Broadway, Baltimore, MD, 21205*

Education

PhD candidate in Biomedical Engineering *August 2012 - May 2018*
JOHNS HOPKINS SCHOOL OF MEDICINE, BALTIMORE, MD, USA
Advisor: Dr. Joel S. Bader, Professor
Field: Computational Biology, Human Genetics

M.S.E in Computer Science *August 2014 - May 2016*
JOHNS HOPKINS UNIVERSITY, BALTIMORE, MD, USA
Advisor: Dr. Alexis Battle, Assistant Professor

M.S.E in Biomedical Engineering *August 2010 - May 2012*
JOHNS HOPKINS UNIVERSITY, BALTIMORE, MD, USA
Advisor: Dr. Jennifer H. Elisseeff, Professor
Field: Cell & Tissue Engineering

Exchange Student in Biomedical Engineering *Summer 2009*
NORTH CAROLINA STATE UNIVERSITY, RALEIGH, NC, USA
Advisor: Dr. Elizabeth G. Loboa, Professor
Field: Tissue Engineering

B.S.E in Biomedical Engineering *August 2006 - July 2010*
ZHEJIANG UNIVERSITY, HANGZHOU, ZHEJIANG, P.R.CHINA
Advisor: Dr. Xiaoxiang Zheng, Professor
Field: Cell & Molecular Biology

Publications

J Zhan, DE Arking, JS Bader. Discovering patterns of pleiotropy in genome-wide association studies. *Submitted to Nature Communications.*

J Zhan, DE Arking, JS Bader. The Structure of human genetics associations permits efficient optimization of NP-hard Bayesian gene-based tests. *Paper in preparation.*

J Zhan, DE Arking, JS Bader. Effect enrichment of SNPs changes power of likelihood-ratio test based GWAS methods. *Paper in preparation.*

S Gupta, SE Ellis, FN Ashar, A Moes, JS Bader, **J Zhan**, AB West, DE Arking. Transcriptome analysis reveals dys-regulation of innate immune response genes and neuronal activity-dependent genes in autisum. *Nature Communication 2014; 5:5748*

J Zhan, A Singh, Z Zhang, L Huang, JH Elisseeff. Multifunctional aliphatic polyester nanofibers for tissue engineering. *Biomatter 2012; Volume2, Issue 4, 201-212.*

A Singh, **J Zhan**, Z Ye, JH Elisseeff. Modular multifunctional PEG hydrogels for stem cell differentiation. *Advanced Functional Materials, 2013; Volume 23, Issue 5, 575-582.*

SD McCullen, **J Zhan**, ML Onorato, SH Bernacki, EG Loboa. Effect of varied extracellular calcium on human adipose derived stem cell mineralization. *Tissue Eng Part A 16(2010). pp. 1971-1981.*

K Xu, Y Yang, M Yan, **J Zhan**, X Fu, X Zheng. Autophagy plays a protecting role during free cholesterol overloading induced smooth muscle cell death. *Journal of Lipid Research, 2010; 51(9):2581-2590.*

Work Experience

Predictive Biomarker Intern *Summer 2016*

MEDIVATION(ACQUIRED BY PFIZER), SAN FRANCISCO, CA, USA

Supervisor: Dr. Hirdesh Uppal, Mr. Heng Wang

Developed machine learning algorithms to predict Xtandi (Enzalutamide, MDV3100) response for triple-negative breast cancer patients.

Teaching Experience

Systems Bioengineering III (EN580.429) *Fall 2015*

DEPARTMENT OF BIOMEDICAL ENGINEERING, JOHNS HOPKINS UNIVERSITY

Supervisor: Dr. Joel S. Bader

Teaching Assistant. Responsibilities: TA sessions, office hours, and grading homeworks/exams.

Honor & Award

Excellent Graduate of Zhejiang Province	*2010*
First Prize Excellent Scholarship, Zhejiang University	*2007, 2009*
Second Prize Excellent Scholarship, Zhejiang University	*2008*
Foreign Exchange Program Scholarship, Zhejiang University	*2008*
Toshiba Excellent Scholarship, Zhejiang University	*2009*
Outstanding Student Leadership, Zhejiang University	*2008, 2009*

Leadership & Extracurricular Activities

Student Ambassador, 100K Strong Program.	*October 2013*
President, Johns Hopkins Chinese Scholar & Student Association	*May,2012 - 2013*