Bias, Precision and Power of Some Techniques in Genome-Wide Association Analysis

by

Pranav C Yajnik

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2020

Doctoral Committee:

Professor Michael Boehnke, Chair
Professor Goncalo Abecasis
Professor Jun Li
Professor Rod Little
Research Professor Laura Scott
Professor Sebastian Zoellner

Pranav C. Yajnik

pyajnik@umich.edu

ORCID iD: 0000-0001-5294-4036

© Pranav C. Yajnik
2020

Dedication

This dissertation is dedicated to my parents, Pippin and Chatura.

May all beings be safe, happy and peaceful.

Acknowledgements

The work presented in this document is the result of the collective effort and support of many people (and at least a few non-human animals). The attribution of its authorship to me merely reflects my role as the person nominated to express the ideas contained in it. Although I know all the people responsible for the successful completion of this work (and, indeed, my successes) do not expect acknowledgement in any form, I will, nevertheless, thank as many as possible given the constraints of space and time.

I would like to express my deepest gratitude to my advisor Mike Boehnke. It goes without saying that your technical expertise played an integral role crafting the ideas presented in this thesis. Of equal importance was your immense patience, kindness, good humor and, genuine concern for my success and well-being.

I would also like to express my sincere thanks and appreciation to Laura Scott, who has essentially been a second advisor. Your careful analyses and critiques of the ideas we discussed always egged me on to broaden the scope of my narrow outlook. Your warmth and concern for my well-being were of immense help when I was down. Your concern for my success and willingness to tell-it-as-it-is were necessary ingredients in my completion of this grand PhD project.

I would also like to thank my committee: Goncalo Abecasis, Jun Li, Rod Little and Sebastian Zoellner for their feedback, support, encouragement and patience. Goncalo, your precise comments and critiques were immensely helpful. Your careful

analysis and takes on the broader implications of this work were most enlightening. Rod, your technical rigor and philosophical acuity have hopefully rubbed off on me. I immensely enjoyed sitting in on your classes; you have turned me into a Bayesian! Sebastian, you've contributed greatly not only to the ideas in this thesis but also towards my training as a statistical geneticist through your class and the many enjoyable (and technically trying) conversations we've had.

I feel extremely privileged to have been a part of the Boehnke-Scott group and to have gotten to know it's wonderful members. Dawn Keene and Peggy White, your ever-present helpfulness in matters bureaucratic (and otherwise) was essential. Anne Jackson and Heather Stringham, you're the queens of data and analysis; your kindness, helpfulness and friendship is much appreciated. Tanya Teslovich, you've been a great mentor, friend and host of many a wonderful get-together at your house with your family; I've learned a lot from you. Shyam Gopalakrishnan and Matthew Zawistowski, you showed me the ropes, introduced me to Sporcle and were the embodiment of pragmatic sanity in the midst of apparent insanity. Mark Reppell and Clement Ma, your soft-spoken kindness and warm friendship is inspiring. Ryan Welch and Robert Weyant, you were awesome office-mates; along with Kraig Stevenson, your impeccable taste in music and enthusiasm for live-shows exposed me to the lively Detroit area metal music scene. Matthew Flickinger, your acting, improv and R-coding skills continue to amaze me; I was honored and humbled to accept the mantle of longest-serving PhD student when you graduated. Corbin Quick, your mathematical acuity was much appreciated. Christian Fuchsberger, Xueling Sim, Jeroen Huyghe, Adam Locke, Debashree Ray, Daniel Taliun, Sarah Gagliano and Xianyong Yin: thank you for all the help and your warm friendship. The number of people to thank in the Computational Statistical Genetics group is too large to list here. I'll make my peace by giving a shout out to Sayantan Das, Alan Kwong, Keng-Han Lin and Adrian Tan who've been with me on this ride for most of the time; you guys certainly made all those ASHG meetings a lot more fun!

Finally, I would like to acknowledge all my friends and family, spread out all over the globe (and, also too numerous to list individually). A person could not hope for dearer friends (and chai drinking buddies) than the Tapri Gang. My extended family- my mother and father in law have been very supportive and patient with me these past few years and for that I am eternally grateful. My parents have been a constant source of support whether it was encouraging every book I wanted to read or introducing me to the world of science. My mother especially went above and beyond in making sure that we were always put first, even though she was a working mother. My father has been a constant source of inspiration through his work ethic and through his research, which he has been conducting with very little infrastructural support/personnel and yet paving the way for putting Indians on the map and for ensuring that Indians from every walk of society have a voice through his work. My brother and sister-in-law lent continuous support from afar while navigating through their own, important life events.

Finally, the most direct and important contribution has been made by the two resident family members here in the US- my wife, Chatura and, our dog, Pippin. Pippin made sure that my sanity was maintained during these past few years by making me get out of the house for long walks and keeping me entertained through his unending antics. Chatura, you have gone above and beyond what any normal human being would do! Without your unconditional support, I would never be here! This achievement is truly as much your success as it is mine.

Table of Contents

List of Figures

Abstract

Genome-wide association studies (GWAS) have successfully identified thousands of genetic loci associated with a wide variety of human phenotypic traits. In this thesis, we evaluate the bias, precision and power of three statistical techniques employed in GWAS.

In Chapter 2, we assess bias and power for adjusted-trait regression (ATR). ATR is a modification to the traditional ordinary least-squares estimation and F-test hypothesis testing techniques for quantitative trait multiple linear regression models. ATR involves performing bivariate correlation analysis between a genetic variant (or set of genetic variants) and a covariate-adjusted trait, obtained by regressing the trait on covariates. We show that ATR effect size estimates for single variant analysis are biased towards the null by a factor equal to coefficient of determination obtained from the regression of genetic variant onto covariates. We derive the exact distributions of ATR test statistics and show that ATR is less powerful than traditional methods when the genetic variant are correlated with covariates. The loss of power increases as stringency of Type 1 error control increases. The maximum possible power loss for the ATR multi-variant test is completely characterized by the canonical correlation between genetic variants and covariates. We show that, for typical covariates like genetic principal components, the loss of power will likely be low in practice.

In Chapter 3, we assess three genetic imputation quality scores (allelic-RSQ, MACH-RSQ and INFO) as predictors for realized imputation quality (squared correlation between true genotypes and imputed dosages) for low-frequency and rare variants. We assess the impact of using different imputation algorithms (Beagle 4.2, minimac3 and IMPUTE 2) and reference panels (1000 Genomes [1KG] and Haplotype Reference Consortium [HRC]) on the relationship between imputation quality scores and realized quality.

We imputed genotypes into 8,378 participants using each imputation algorithm with the 1KG panel and minimac3 with the HRC panel. We show that MACH-RSQ and INFO are identical when calculated on the same data. We observe that allelic-RSQ predicts realized quality less well than MACH-RSQ/INFO for low-frequency and rare variants. Realized quality decreases as minor allele frequency (MAF) decreases. The mean absolute difference (MAD) between quality scores and realized quality increases as MAF decreases. Imputation with HRC resulted in better realized quality for low-frequency and rare variants compared to imputation with 1KG. However, the MAD between quality scores and realized quality for low-frequency and rare variants was similar for both panels.

In chapter 4, we assess the efficiency gained or lost by adding an external sample with missing case-control status to an (internal) case-control study sample. We propose a method for estimation and testing that accounts for the known (or presumed) proportion of cases in the external sample. Misspecification of the external sample case proportion leads to biased estimation; in particular, treating the external sample as a control sample leads to underestimation of the effect size. However, the proposed test controls Type 1 error regardless of the particular value chosen for the presumptive external sample case proportion. When treating the external participants as controls, addition of external participants improves power if the proportion of cases in the internal sample is at least twice that in the external sample.

Chapter 1 Introduction

The publishing of the draft sequence of the human genome (Lander et al., 2001) and availability of a large catalog of human genetic variation (International HapMap Consortium, 2005) set the stage for two decades of unprecedented discovery in human genetics. Since the (arguably) first genome-wide association study (GWAS) identified a single locus associated with myocardial infarction (Ozaki et al., 2002; Ikegawa, 2012), more than 4500 GWAS have identified nearly two hundred thousand associations between genetic variants and thousands of phenotypic traits (GWAS catalog: https://www.ebi.ac.uk/gwas/).

In contrast to Mendelian or monogenic traits, most common diseases and phenotypic traits are complex (Timpson et al., 2018; Watanabe et al., 2019) with numerous genetic and environmental causative factors, each typically having modest influence. GWAS assess association (correlation) between a phenotypic trait and large numbers (currently, up to millions) of genetic variants sampled approximately uniformly across the genome. This design and analysis strategy was proposed as an alternative to linkage analysis, an approach that works well for Mendelian traits but not complex traits (Risch and Merikangas, 1996; Hirschhorn and Daly, 2005) and candidate gene studies that showed poor replication in practice (Hirschhorn et al., 2002).

Despite the success of GWAS at identifying previously unknown associations, currently identified associations typically account for only a modest fraction of the total genetic influence (heritability) of most complex traits (Boyle et al., 2017; Watanabe et al., 2019). Large sample sizes are required to identify common variants with low effect sizes and low-frequency and rare variants with moderate to large effect sizes with appropriate correction for the number of hypothesis tests conducted in a GWAS. Current study sizes for large GWAS range from hundreds of thousands of participants up to a million participants; large samples sizes are primarily obtained via meta-analysis of multiple studies and less often, through joint-analysis of participants pooled across studies.

Conducting large genetic studies presents multiple challenges in addition to the primary challenge of recruiting a large number of participants in a systematic and principled manner. The storage and analysis of large datasets is computationally challenging. Large studies are expensive with one of the major expenses for GWAS being the cost of assaying genotypes for large numbers of genetic variants. Differences in sampling protocols and measurement techniques across different batches of data (or studies) need to be addressed with statistical rigor.

Regression models are the most commonly used framework for conceptualizing the relationship between genetic variants and phenotypes, testing for association and estimating variant effect sizes. Regression models accommodate both quantitative and disease phenotypes (and retrospective designs like case-control studies), and single and multiple variant tests, while allowing adjustment for confounding factors including age, sex, and population stratification. Mixed-effects regression models accommodate presence of related individuals (Kang et al., 2010; Loh et al., 2015). The score (or Lagrange-multiplier) test (Cox & Hinckley, 1979) is commonly employed to test for association in GWAS and other genome-wide analyses like expression quantitative traits locus (eQTL) studies.

GWAS analysis involves performing a large number of tests with the same phenotype but different variants (or sets of variants). The score test requires maximum likelihood estimates only under the null; since the null model doesn't include the genetic variant being tested, the computationally expensive maximization procedure needs to be performed only once for a given sample. In this setting, the score test is computationally more efficient than the Wald and likelihood ratio tests which require maximization of the likelihood for every variant. Thus, in chapters 2 and 4, we focus attention to the score test for linear (chapter 2) and logistic (chapter 4) regression models.

Next-generation sequencing platforms have the capacity to assay nearly the complete spectrum of human genetic variation and can detect rare and previously unknown variants. Despite a substantial reduction in cost over the last decade, sequencing large numbers of study participants remains prohibitively expensive. Most GWAS assay genetic variants through microarray genotyping technology. Microarrays typically assay a few hundred thousand to a few million genetic variants. Although this represents a small fraction the total number of known genetic variants, microarrays are designed to include variants that optimally tag (or predict) nearby unmeasured variation. To extend the array-assayed information, GWAS routinely employ the strategy of imputing unmeasured variants (Li et al., 2010; Marchini & Howie, 2010; Das et al., 2018) with the aid of external (and often freely available), comprehensively sequenced genotype imputation reference panels (McCarthy et al., 2016). Genotype imputation improves power for GWAS based on microarray genotyping by facilitating meta-analysis of studies utilizing different microarrays and leveraging information from multiple nearby variants to predict unmeasured variants with greater accuracy than single tagging variants (Marchini & Howie, 2010; Das et al., 2018). Given current sequencing costs, microarray genotyping followed by imputation is a cost-effective and powerful strategy for GWAS (Quick et al., 2019).

Recently, large databases like the UK Biobank (Bycroft et al., 2018), Genome Aggregation Database (gnomAD), and Exome Aggregation Consortium (ExAC) (Lek et al., 2016) have emerged. These databases contain comprehensive genetic information on tens to hundreds of thousands of participants and were created as a resource for the scientific community. It is likely that the number and size of such databases will increase over the next decade. Such datasets present an inexpensive resource for both aggregated, and in some cases, participant level genetic information. For disease GWAS with relatively rare diseases, genetic and phenotypic measurements for participants from these databases could be added to existing or new GWAS (as external controls, in case disease information is unavailable in the database) enabling larger sample sizes at relatively lower cost. Challenges to this approach include bias due to differential sample ascertainment/selection and differential measurement error (for genetic and other variables) between the GWAS sample and database sample.

This thesis focuses on bias, precision and power for three statistical techniques used in GWAS. Chapter 2 focuses on a variant of standard estimation and testing techniques for linear regression for quantitative trait association. Chapter 3 focuses on predicted quality of genotype imputation. Chapter 4 focuses on misclassified or missing case-control status in a logistic regression model of case-control data.

In Chapter 2, we assess bias and power for adjusted-trait regression (ATR), an often-used variant of the traditional ordinary least-squares and F-test techniques for linear regression (Randall et al., 2013; UK10K Consortium, 2015; Tachmazidou et al., 2017; Kanai et al., 2018; Styrkarsdottir et al., 2019; Niarchou et al., 2020). ATR proceeds by first creating a covariate adjusted trait by regressing the trait onto covariates. In the second step, bivariate correlation analysis (simple regression) is performed between each variant (or each set of variants for gene-based tests) and the adjusted trait. Although this strategy looks superficially similar to the score test, previous work (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012) has shown that ATR estimates are biased towards the null (in case of single variant tests, by a factor equal to the coefficient of determination, R^2, for the regression of variant onto covariates) and used simulations and approximations to show that ATR based single variant tests are less powerful than traditional tests like the F-test. We show that this approach is less powerful than traditional techniques by deriving the exact relationships between ATR and traditional test statistics for the single variant test and inequalities for the omnibus gene-based test. We show that the loss of power increases as the canonical correlation between variants and covariates increases but power loss is unlikely to be large in typical situations.

In Chapter 3, we assess three imputation quality scores (allelic-RSQ, MACH-RSQ and INFO) as predictors of realized imputation quality (squared correlation between imputed dosages and observed genotypes) for low-frequency and rare variants by imputing genotypes in 8,378 Finnish participants in the METSIM study for which “gold-standard” genotypes for a subset of imputed variants were assayed with the Illumina ExomeArray. We also assess the impact of reference panel size (1000 Genomes [1KG] Phase 3 versus Haplotype Reference Consortium [HRC] reference panels) and choice of imputation algorithm (Beagle 4.2, minimac3 and IMPUTE 2 with 1000 Genomes reference panel).

We show that MACH-RSQ and INFO yield the same value when calculated on the same data. Allelic-RSQ is a poor predictor of observed quality for rare and low-frequency variants; hence we recommend using MACH-RSQ/INFO as a quality score. It is well known that realized quality decreases with decreasing minor allele frequency (MAF). We observe that imputation quality score predictions (of the realized quality) also become worse as minor allele frequency decreases. When imputation quality scores are used as a classifier high-vs-low realized quality, more stringent thresholds are required for rare and low-frequency variants as compared to common variants. Imputation based on the HRC reference panel yields higher realized qualities for low-frequency and rare variants compared to imputation based on the 1KG panel. With rare variant imputation based on HRC, less stringent thresholds (e.g. MACH-RSQ/INFO=0.3) yield the same classification operating characteristics as substantially more stringent thresholds (MACH-RSQ/INFO=0.6) with 1KG imputation.

In chapter 4, we assess the efficiency gained or lost by adding an external sample with missing case-control status to an (internal) case-control study sample. For simplicity, we assume that other potential sources of bias, like selection bias and differential measurement error of genotypes between internal and external samples, are absent or can be controlled prior to association analysis (for example, by stringent quality control). We propose a likelihood based method that models the external sample as a mixture of cases and controls based on a known (or presumed) external sample case proportion. This method is a generalization of the naïve strategy of treating external participants as controls (achieved by assuming that external sample case proportion is zero). We derive a closed form non-centrality parameter for the score test and asymptotic approximations for bias when the external sample case proportion is incorrectly specified.

We show that estimates of effect size are unbiased under the null but biased under the alternative. In particular, setting the external sample case proportion to a value less than its true value (for example, treating all external participants as controls) results in underestimation under the alternative; setting it to values larger than the true value may result in either under or overestimation under the alternative. We show that the score test controls Type 1 error regardless of the particular value assumed for the external sample case proportion. For analysis that treats external participants as controls, including the external sample is more powerful than avoiding it when the *internal* sample case proportion is at least twice as large as the external sample case proportion. In situations where the external sample case proportion is large relative to the internal sample case proportion, power can be gained upon inclusion of the external sample if the external sample case proportion is known accurately.

In Chapter 5, we present a brief summary of our results and discuss possible extensions and future directions.

REFERENCES

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, *169*(7), 1177-1186.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Cortes, A. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203-209.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Che, R., Motsinger-Reif, A. A., & Brown, C. C. (2012). Loss of power in two-stage residual-outcome regression analysis in genetic association studies. *Genetic epidemiology*, *36*(8), 890-894.

Demissie, S., & Cupples, L. A. (2011). Bias due to two-stage residual-outcome regression analysis in genetic association studies. *Genetic epidemiology*, *35*(7), 592-596.

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype imputation from large reference panels. *Annual review of genomics and human genetics*, *19*, 73-96.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics*, *6*(2), 95-108.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in medicine*, *4*(2), 45-61.

Ikegawa, S. (2012). A short history of the genome-wide association study: where we were and where we are going. *Genomics & informatics*, *10*(4), 220.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., ... & Kubo, M. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature genetics*, *50*(3), 390-400.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, *42*(4), 348-354.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Funke, R. (2001). Initial sequencing and analysis of the human genome.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285-291.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, *34*(8), 816-834.

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., ... & Patterson, N. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, *47*(3), 284.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499-511.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... & Luo, Y. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, *48*(10), 1279-1283.

Niarchou, M., Byrne, E. M., Trzaskowski, M., Sidorenko, J., Kemper, K. E., McGrath, J. J., ... & Wray, N. R. (2020). Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits. *Translational Psychiatry*, *10*(1), 1-11.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., ... & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, *32*(4), 650-654.

Quick, C., Anugu, P., Musani, S., Weiss, S. T., Burchard, E. G., White, M. J., ... & Fuchsberger, C. (2019). Sequencing and imputation in GWAS: Cost‐effective strategies to increase power and genomic coverage across diverse populations. *Genetic Epidemiology*.

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., ... & Workalemahu, T. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet*, *9*(6), e1003500.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516-1517.

Styrkarsdottir, U., Stefansson, O. A., Gunnarsdottir, K., Thorleifsson, G., Lund, S. H., Stefansdottir, L., ... & Ivarsdottir, E. V. (2019). GWAS of bone size yields twelve loci that also affect height, BMD, osteoarthritis or fractures. *Nature communications*, *10*(1), 1-13.

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R., Steinberg, J., Walter, K., ... & McCarthy, S. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *The American Journal of Human Genetics*, *100*(6), 865-884.

Timpson, N. J., Greenwood, C. M., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), 110.

UK10K consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82-90.

Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., ... & Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics*, *51*(9), 1339-1348

Xing, G., Lin, C. Y., & Xing, C. (2011). A comparison of approaches to control for confounding factors by regression models. *Human heredity*, *72*(3), 194-205.

Chapter 2 Power loss due to testing association between covariate adjusted traits and genetic variants

INTRODUCTION

Multiple linear regression and the associated ordinary least-squares and F-test methodologies are effective and widely used approaches to test for association between genetic variants and quantitative traits and to estimate genetic effect sizes while controlling for the effects of other variables (covariates). Covariates may be included to account for confounding (e.g. due to population structure or assay batch effects), to reduce trait variability and consequently increase power, or to exclude associations that are driven primarily through the action of the variants on an intermediate trait.

Current genome-wide association studies (GWAS) typically assay hundreds of thousands to millions of genetic variants. Single-variant association tests are performed separately on each variant to test whether the variant is associated with the trait. Multi-variant, gene-, or region-based tests are performed to address the omnibus hypothesis that one or more in a set of variants are associated with the trait. Since the dependent variable and covariates are typically the same across all tests, some analysts use a two-stage approach for quantitative trait GWAS (Randall et al., 2013; UK10K Consortium, 2015; Tachmazidou et al., 2017; Kanai et al., 2018; Styrkarsdottir et al., 2019; Niarchou et al., 2020 are some examples of studies employing this methodology). In the first stage, an 'adjusted' trait is obtained as the residuals from the regression of the trait on covariates. In the second stage, association analyses are performed to test for association between the adjusted trait and each variant (or set of variants) without inclusion of other covariates. We term this strategy “adjusted-trait regression (without covariates)” (ATR).

Although ATR can be conceptualized as a two-stage method, we note that it bears no relation to the “two-stage least-squares” method used in structural equations modeling and estimation of causal effects using instrumental variables. We assume that the target of inference is the conditional association between the unadjusted trait and variants given the covariates rather than the association between the adjusted trait and variants unconditional on the covariates. Thus, we view ATR as a numerical technique to conveniently approximate the results that would have been obtained from analysis of the unadjusted trait (with covariates included). The strategy of analyzing a covariate-adjusted trait may be used for any statistical method that deals with linear models, including gene/region based tests like burden or SKAT (Lee et al., 2014) or methods for linear mixed-models.

We have not found any methods papers that recommend the use of ATR. Indeed, the research articles cited above make use of ATR without comment or justification. ATR results are not identical to results obtained from modeling the unadjusted trait along with covariates. Previous investigations of single-variant models showed that the ordinary least-squares ATR estimator of genetic effect is biased towards zero by a factor of $1 - R^2$ (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012), where R^2 is the sample coefficient of determination obtained by regressing the genetic variant onto the covariates. These investigations used approximations and simulations to assess power and Type 1 error of the ATR-based tests assuming a Type

1 error rate of *α=0.05* and showed that ATR is typically less powerful than multiple linear regression when the sample correlation between a genetic variant and covariates is non-zero. More recently, Sofer et al. (2019) showed that the ATR-based single-variant score and multi-variant SKAT test statistics are numerically (deterministically) dominated by the corresponding test statistics obtained from analyzing the unadjusted trait with covariates leading to deflated p-values and loss of power.

We extend these previous results by deriving the exact relationship between ATR and multiple linear regression score, likelihood ratio, Wald, and F test-statistics for single-variant analysis. We use these relationships to derive (1) the exact finite sample distributions of the ATR test-statistics (hence, exact power and Type 1 error) under the assumption of independent and identically normally distributed errors and (2) the asymptotic relationship between the test-statistics for situations where the assumption is suspect. In addition, we derive the asymptotic distributions of ATR based analogs of two gene/region-based tests: the burden test and the (omnibus) score test, and show that these tests applied in the ATR framework may also suffer from loss of power compared to their multiple linear regression analogs. In particular, we show that the maximum possible power loss for gene-based ATR score tests depends on the maximum canonical correlation between the set of variants and the set of covariates, so that we expect power loss to be modest in typical GWAS with low to moderate population structure.

METHODS AND RESULTS

Definition of the ATR approach

We assume a model of the form:

$$Y_i = \alpha + \sum_{j=1}^{m} \beta_j g_{ij} + \sum_{l=1}^{k} \gamma_l c_{il} + \epsilon_i$$

(M1)

Here $Y_i, 1 \leq i \leq n$ is the trait value for the i^{th} study participant, g_{ij} the genotype (or genotype-imputation-based dosage) for the j^{th} variant for this study participant, β_j the effect of the variant on the trait (conditional on the other $m-1$ variants and covariates), c_{il} the value of the l^{th} covariate, γ_j the (conditional) effect of the covariate, and ϵ_i a random error. We assume the errors are independent and identically distributed across observations with $\mathbb{E}(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. For single-variant models, $m = 1$ and β is the conditional effect of the variant on the trait given the covariates, but unconditional on any other variant.

The above model can be represented as $Y = G\beta + C\gamma + \epsilon$ where $Y \text{ and } \epsilon$ are $n \times 1$ vectors, G is an $n \times m$ matrix, β is a $m \times 1$ vector, C is an $n \times (k+1)$ matrix (including a column of ones for the intercept), and $\gamma = (\alpha, \gamma_1, \dots, \gamma_k)'$ is a $(k+1) \times 1$ vector. We have $Var(Y|G,C) = Var(\epsilon) = \sigma^2 I_n$ where I_n is the n-dimensional identity matrix. We wish to test $H_0: \beta = 0$. Further, we assume that the test statistic T has the form $T = f(Y, G, C)$. We note that the distribution of T under the null may depend on G and C and on parameters that need to be estimated from the data. We assume that the (possibly estimated) parameter value $\hat{\theta}$ required to define the distribution of T under the null (for example, degrees of freedom for the F-statistic) also has the form $\hat{\theta} = g(Y, G, C)$.

Let $H_C = C(C'C)^{-1}C'$. Then $Y_r = Y - C\hat{\gamma} = (I_n - H_C)Y$ is the vector of residuals obtained by regressing Y onto C using ordinary least squares (with $\hat{\gamma} = (C'\square)^{-1}C'Y$). We define the ATR analog of T to be $T_{ATR} = f(Y_r, G, J_n)$ where $J_n = (1, \dots, 1)$ is the $n \times 1$

vector of ones denoting the intercept. Further, we assume that the parameter θ for ATR is calculated as $\hat{\theta}_{ATR} = g(Y_r, G, J_n)$. This definition of the ATR analog implies that inference based on T_{ATR} can be performed by using existing software designed for inference with T simply by replacing Y and C with Y_r and J_n. We note that if the parameter of the null distribution for a method depends on Y and/or C, we may have $\hat{\theta} \neq \hat{\theta}_{ATR}$, and the ATR analog may reference a null distribution that differs from the one used by the unadjusted method to calculate p-values.

Ordinary least-squares estimation with ATR

The ordinary least-squares estimator of β is given by $\hat{\beta} = (G_r'G_r)^{-1}G_r'Y_r$ where $G_r = (I_n - H_C)G$ is the matrix of residuals of variants regressed onto C. This result is often referred to as the Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933; Lovell, 2008). In the appendix, we show that

$$\hat{\beta}_{ATR} = (I_m - R_{GC}^2)\hat{\beta}$$

where $R_{GC}^2 = [G'(I_n - H_1)G]^{-1}G'(H_C - H_1)G$ and $H_1 = J_nJ_n'/n$. Note that the eigenvalues of R_{GC}^2 are the sample canonical correlations between the set of genetic variants and the set of covariates. In particular, $R_{GC}^2 = 0$ (the zero matrix) if and only if every genetic variant is uncorrelated with all covariates. Further, we have $\mathbb{E}(\hat{\beta}_{ATR}) = (I_m - R_{GC}^2)\beta$ and, consequently, $\mathbb{E}(\hat{\beta}_{ATR}) = 0$ if and only if none of the genetic variants are associated with the trait (conditional on covariates). Thus, any test that is valid for testing the omnibus hypothesis $H_0: \mathbb{E}(\hat{\beta}_{ATR}) = 0$ is also valid for testing $H_0: \beta = 0$.

In the case of single-variant analysis ($m = 1$), the above relationship simplifies to $\hat{\beta}_{ATR} = (1 - R^2)\hat{\beta}$ and we recover the result obtained previously (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012; Sofer et al., 2019). Thus, for single-variant analysis, the ATR ordinary least-squares estimator can only be biased towards the null. This is not true for individual elements of $\hat{\beta}_{ATR}$ when $m > 1$. Indeed, $\mathbb{E}\left(\hat{\beta}_{ATR}\right)_j$ is a linear combination of all the elements of the vector β. In particular, $\beta_j = 0$ does not necessarily imply that $\mathbb{E}\left(\hat{\beta}_{ATR}\right)_j = 0$. Thus, a test that is valid for $H_0: \mathbb{E}\left(\hat{\beta}_{ATR}\right)_j = 0$ is not necessarily valid for $H_0: \beta_j = 0$ (unless all remaining elements of β are also 0).

Single-variant association testing with ATR

Xing et al. (2011) showed that $W_{ATR} \leq W$ where W is the Wald test statistic. Che et al. (2012) refined an approximation proposed by Demissie and Cupples (2011) for the F test statistic (F) to $F_{ATR} = \frac{n-2}{n-k-2}\left(1 - \frac{R^2}{1+R^2 r^2(Y_r,G_r) - r^2(Y_r,G_r)}\right)F$ where $r^2(Y_r, G_r)$ is the sample squared correlation between Y_r and G_r and F is the F statistic. Xing et al. (2011) and Che at al. (2012) used simulations to estimate power and Type 1 error rate for $\alpha = 0.05$.

We show that $S_{ATR} = (1 - R^2)S$, where S is the score test statistic for the above linear model when $m = 1$. For linear models, the test statistics for the score, Wald, likelihood ratio, and F tests bear simple, deterministic relationships to each other (Vandaele 1981). Combining $S_{ATR} = (1 - R^2)S$ with these known relationships yields the following set of equalities:

$$F_{ATR} = \frac{n-2}{n-\square-2} \times \frac{(1-R^2)F}{1+R^2F/(n-k-2)}$$

$$W_{ATR} = \frac{(1-R^2)W}{1+R^2W/n}$$

$$LR_{ATR} = LR - n\log(1 + R^2[e^{LR/n} - 1])$$

where LR denotes the likelihood ratio test statistic. We see that $S, W,$ and LR are always strictly greater than their ATR anologs if $R^2 > 0$ and equal to them if $R^2 = 0$. P-values for the score, Wald, and likelihood ratio tests are standardly computed assuming the test statistics follow a chi-square distribution with $\theta = s = 1$ degree of freedom (χ_1^2 distribution). The ATR analogs of these methods also assume this same distribution and are less powerful than their counterparts if $R^2 > 0$.

In contrast, $F_{ATR} > F$ if $F < \frac{k}{R^2} - (n-2)$ and the ATR analog of the F-test uses the F-distribution with 1 and $n-2$ degrees of freedom while the F-test assumes a distribution with 1 and $n-k-2$ degrees of freedom; in this case, $\hat{\theta}_{ATR} \neq \hat{\theta}$ since the denominator degrees of freedom depends on the number of covariates. Thus, the ATR analog of the F-test may be slightly anti-conservative if $R^2 \approx 0$ and/or the number of covariates is large relative to the sample size. This is quite unlikely given the large sample sizes of current GWAS, the large values of the test statistic required to reject the null, and the fact that the expected value of the sample coefficient of determination increases with increasing number of predictors, even when the variant is independent of the predictors at the population level, in which case $\mathbb{E}(R^2) \approx k/n$ for large samples.

For a fixed number of covariates, the score, Wald, likelihood ratio, and F test statistics asymptotically converge to the same random variable T (almost surely) under the null and local alternatives ($\beta = O(n^{1/2})$ i.e. when the effect size tends to zero asymptotically). Similarly, their ATR analogs each converge to $(1 - R^2)T$. Asymptotically, each of the ATR test statistics follows a scaled χ_1^2 distribution whose scaling factor is less than or equal to one and are, thus, conservative when $R^2 > 0$. The exact finite sample distribution of the F statistic is known in the case where errors are normally distributed; the exact distributions of all the other test statistics can be derived easily given the above relationships.

For simplicity, we illustrate the conservative nature of ATR for single-variant tests under asymptotic conditions. Here, we have $P(T_{ATR} < \alpha) = P(T < \alpha/(1 - R^2))$. The relationship between the p-values generated by the score test and its ATR analog is non-linear; the ATR test becomes more conservative as the p-value threshold for declaring significance (α) becomes more stringent. Figure 2-1 shows power of the ATR test with $R^2 = 0.05$ for α values ranging from 10^{-1} to 10^{-10} where the effect size for each α value is chosen to yield 80% power for the score test. At the usual GWAS threshold of $\alpha = 5\times10^{-8}$, the power of the ATR test is

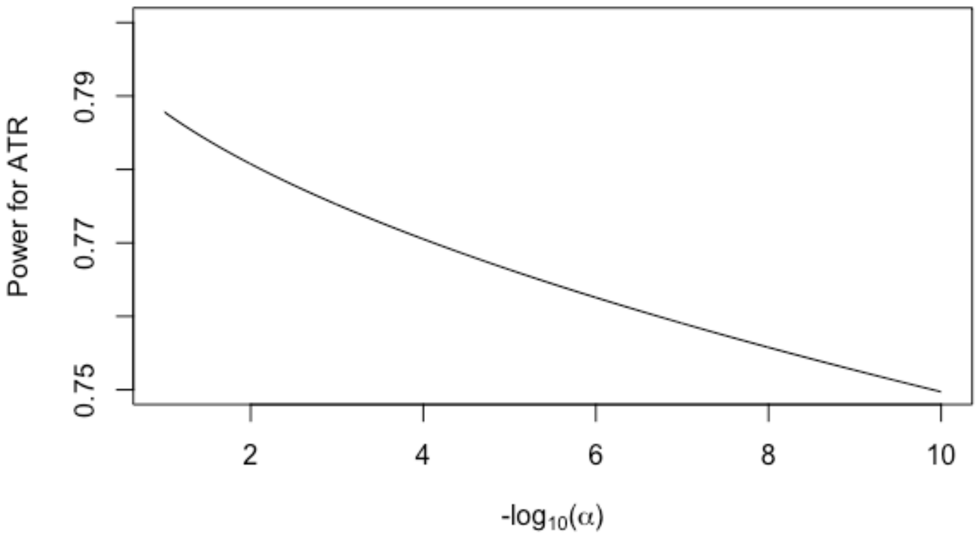

Figure 2-1 Power of ATR analog of single-variant score test when $R^2 = 0.05$ with varying stringency of statistical significance α displayed in the negative log ten scale. Effect sizes vary as a function of α to yield 80% power for the score test.

about 76%. Figure 2-2 shows how, for fixed $\alpha = 5\times10^{-8}$, the ATR test becomes less powerful as R^2 increases (again, with effect size chosen to yield 80% power for the score test).

Burden tests with ATR

The relationships derived for the single-variant tests are directly applicable to burden tests. Burden tests typically assume the same multiple linear regression model presented in the previous section with G replaced by $B = \sum_{i=1}^{m} w_i G_i = GW$ where $G_i, \dots, G_m$ are m genetic variants (columns of G), w_i are weights (and $W = (w_1, \dots, w_s)'$), and B is the (weighted) burden of alternate alleles (or genotype imputation-based dosages) from the m variants. For burden tests, R^2 is the sample coefficient of determination obtained by regressing B onto C. Given G and C, the maximum possible value for R^2 is obtained when the weight vector W is a scalar multiple of the eigenvector of R_{GC}^2 corresponding to the maximum eigenvalue and the maximum R^2 is equal to the maximum eigenvalue.

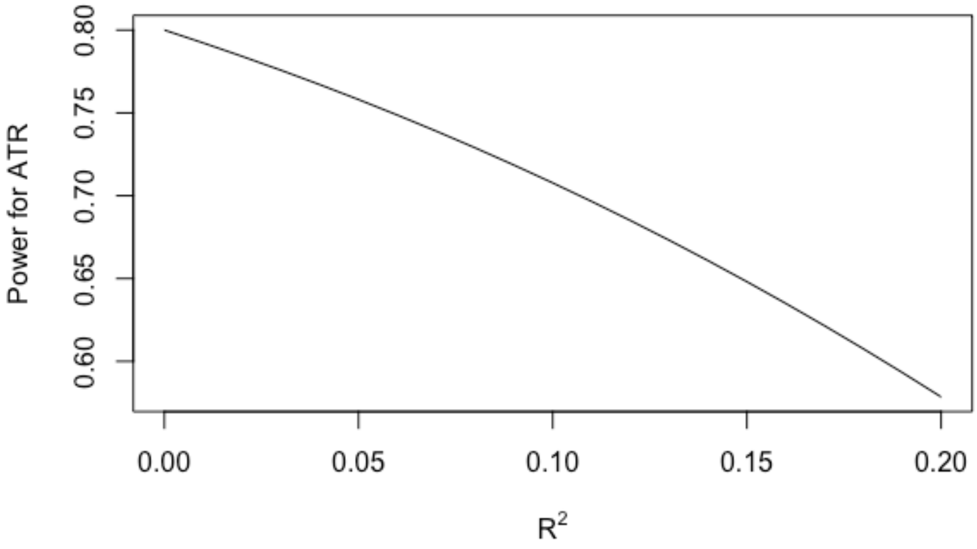

Figure 2-2 Power of ATR analog of single-variant score test with increasing R^2 for $\alpha = 5\times10^{-8}$. The effect size was chosen to yield 80% power for the score test.

Classical omnibus tests with ATR

The omnibus null hypothesis that none of the m variants are associated with trait (conditional on covariates) can be tested with the omnibus/multivariate score, Wald, likelihood ratio, and F tests. As before, these tests are asymptotically equivalent and we consider the score test as an exemplar. Unlike the single-variant case, no deterministic relationship exists between S_{ATR} and S when $m > 1$ (that is, S_{ATR} can take multiple values for any given value of S). However, we show that

$$(1 - R_{max}^2)S \leq S_{ATR} \leq (1 - R_{min}^2)S$$

where R_{max}^2 and R_{min}^2 are the maximum and minimum canonical correlations between the variants and covariates. Recall that S asymptotically follows a $\chi_m^2(\delta^2)$ distribution with non-centrality parameter $\delta^2 = \frac{1}{\sigma^2}\beta' G'(I_n - H_C)G$. Under the null, the distribution of S depends only on the parameter $\hat{\theta} = m$. Asymptotically, S_{ATR} follows the same distribution as the random variable $\sum_{i=1}^{p}(1-R_i^2)Z_i$ where $R_1^2,\dots,R_p^2$ are the distinct eigenvalues of R_{GC}^2 (in decreasing order so that $R_1^2=R_{max}^2$ and with p possibly smaller than m) and the random variables Z_i are mutually independent with $Z_i \sim \chi_{\nu_i}^2(\lambda_i^2), \sum_{i=1}^{p}\nu_i = m$ (see Appendix). Since $\hat{\theta}$ is independent of C, we have $\hat{\theta}_{ATR}=\hat{\theta}$ and p-values for S_{ATR} are calculated assuming a central χ_s^2 distribution.

Note that the score test yields the same power for all effect size vectors β such that $\beta' G'(I_n - H_C)\beta = c$ where $c \geq 0$ is a constant. Although the actual difference in power between S and S_{ATR} depends on the true value of β, we show that, amongst all β that yield the same power for the score test, the ATR analog achieves *minimum* power when β is a scalar multiple of the eigenvector of R_{GC}^2 corresponding to the maximum eigenvalue (see Appendix). Here, $\lambda_1^2=\delta^2$ and $\lambda_i^2=0$ for $i=2,\dots,p$. Thus, the *maximum possible* power loss of the ATR analog of the score test (relative to the score test) is completely characterized by the set of canonical correlations between the variants and covariates.

Figure 2-3 shows, for fixed $\alpha = 5\times10^{-8}$ and $s=10$ variants, the power of ATR analog across a range of R_{max}^2 with effect size chosen to yield 80% power for the omnibus score test. We calculated tail probabilities for the distribution of S_{ATR} using Davies' method as implemented in the R package *CompQuadForm* (de Micheaux, P. L., & de Micheaux, M. P. L., 2017). We consider two situations. First, if the remaining canonical correlations are zero, the maximum possible power loss is slightly larger than that for the single-variant case for $m=10$ and power loss increases as m increases ($m=100$ shown in Figure 2-3). Second, if all canonical correlations are equal to R_{max}^2, S_{ATR} follows the scaled chi-squared distribution $(1-R_{max}^2)\chi_m^2(\delta^2)$, and the maximum possible power loss is equal to the minimum possible power loss; thus, for a given value of R^2_{max}, this constitutes the worst-case scenario for ATR (Figure 2-3). Note that the maximum number of non-zero canonical correlations cannot exceed $\min(m, k)$. Thus, the second scenario is unlikely to occur in practice.

DISCUSSION

The ATR approach is often used in genetic association studies (Randall et al., 2013; UK10K Consortium, 2015; Tachmazidou et al., 2017; Kanai et al., 2018; Styrkarsdottir et al., 2019; Niarchou et al., 2020), and several papers have used simulation to assess its properties at modest significance thresholds (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012). However, to our knowledge no papers have presented analytic evaluations of ATR or considered significance thresholds appropriate for GWAS. The Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933; Lovell, 2008) demonstrates that when the target of inference is confined to a

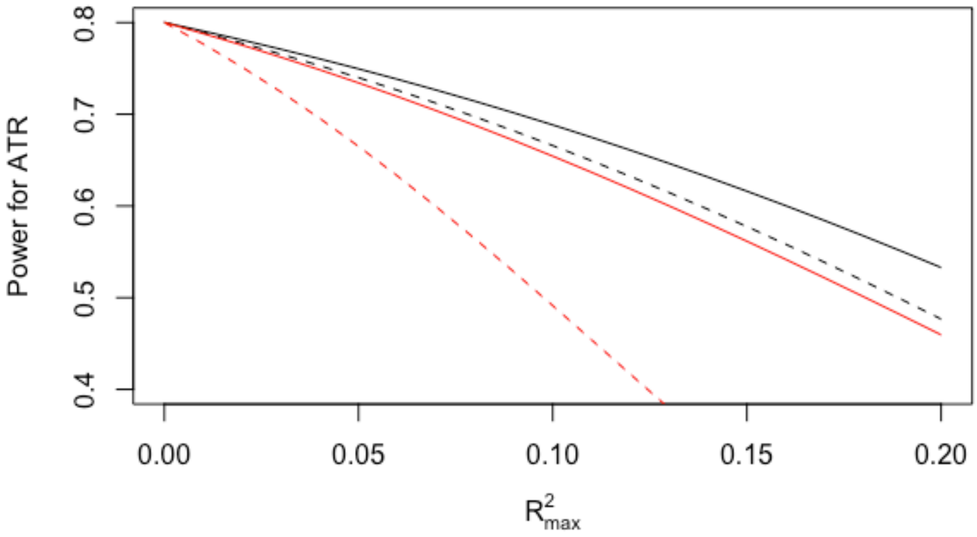

Figure 2-3 Power of ATR analog of the multi-variant (omnibus) score test (Y-axis) with $m = 10$ *(black) and* $m = 100$ *(red) variants. X-axis shows the maximum canonical correlation between the variants and covariates. Solid line: power when the other canonical correlations are 0. Dashed line: power when other canonical correlations are equal to the maximum correlation. Effect sizes for the set of variants are chosen to yield 80% power for the omnibus score test and minimum power for the ATR analog (see text) with* $\alpha = 5{\times}10^{-8}$.

subset of predictors in the multiple linear regression model (e.g. genetic variants), OLS analysis can be achieved as a two-stage method by regressing the covariate adjusted trait onto the covariate adjusted variants. Thus, the ATR strategy of adjusting the trait but not the variants is formally justified in the context of multiple linear regression only when variants and covariates are uncorrelated.

It may seem that score-tests like those presented above or SKAT employ the same strategy as ATR. Indeed, for single-variant analyses the *score-statistic* for linear models ($G'Y_r$) is based on the adjusted trait and *unadjusted* variant. However, the *score test-statistic* (calculated by squaring the score-statistic and dividing by its estimated variance) does depend on the adjusted variants. Indeed, it can be shown that ATR over-estimates the variance of the score-statistic by a factor of $(1 - R^2)^{-1}$ due to using unadjusted variants in the variance calculation. Our derivations also show that single-variant OLS based inference can be fully recovered from the ATR based inference given the summary statistic R^2 for each variant. For multi-variant analyses, the entire R_{GC}^2 matrix is required.

For single-variant association tests, previous papers show by computer simulation that ATR is less powerful than the (theoretically justified) two-sided t and Wald tests when the variant is correlated with the covariates (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012; Sofer et al., 2019). We extend previous results by deriving the exact distribution of the ATR analogs for single-variant Wald, likelihood ratio, score, and F tests, and the asymptotic distributions for gene-based burden and score tests, and assessing size and power at significance levels appropriate for GWAS.

For single-variant tests, we show that the loss of power of the ATR method is completely characterized by the coefficient of determination (R^2) obtained by regressing the variant onto the covariates, with the power loss increasing with increasing R^2. Further, we show that loss of power increases as the p-value cutoff used to declare significance becomes more stringent. Characterizing power loss for the ATR analogs of gene-based tests is more complex. For gene-based score tests, the power loss depends on both the (true) strength of association between each variant and the outcome, and the correlation between each variant and the covariates. Power loss is greater when the subset of variants driving the association is also the subset that is driving the canonical correlation between variants and covariates. For the ATR analogs of the multiple linear regression omnibus test of association, we show that the *maximum* possible power loss is completely characterized by the canonical correlations between the variants and covariates with maximum power loss increasing with increasing values of any of the canonical correlations. When there is only a single non-zero canonical correlation, the maximum power loss is similar to the single-variant case.

At the significance threshold of $\alpha = 5\times10^{-8}$ typically used in GWAS, an R^2 of 0.1 results in power decreasing from 80% (for the two-sided t test) to about 71% for the single-variant ATR test. Thus, we recommend that ATR based methods only be used when the R^2 for the majority of variants is expected be substantially less than 0.1. We re-iterate that sets of covariates not associated with the variant do not result in loss of power due to using ATR; in fact, they increase power if they explain some of the trait variance (Robinson & Jewell, 1991). Covariates that are associated with the trait but not genetic variants in a population based sample may be associated with genetic variants in studies that sample participants non-randomly (Munafo et al., 2018; Greenland et al., 1999); for example, two variables that both cause a disease but are independent in a population will be associated in a case-control sample (Monsees et al., 2009).

In GWAS, the most commonly included covariates that are likely to be correlated with a large number of variants are indicators of genetic ancestry (e.g. principal components). The distribution of correlation depends on the degree of population structure in the sample and the mean R^2 across variants is (approximately) the sample F_{st}. For intra-continental samples, typically $F_{st} < 0.05$ but for inter-continental samples it can be > 0.1 [The 1000 Genomes Project Consortium, 2015]. As a further example, we calculated R^2 between ~750,000 genotyped variants and the first 2, 5, and 10 genetic principal components for ~409,000 participants with white-British ancestry in the UK Biobank (details of SNP QC and PCA generation in Bycroft et al., 2018) and found all R^2 values were < 0.05. In the analysis including the remaining 78,000 non-white participants (total sample size ~487,000), 6% of variants showed $R^2 > 0.05$ and 2.5% showed $R^2 > 0.10$ (the results were approximately similar with 2, 5, and 10 PCs).

Other commonly included covariates that may be correlated with variants are intermediate traits lying in between the gene and primary trait in the causal pathway, and indicators of sample processing or batch effects. For intermediate traits that are genetically complex, values of R^2 will typically be much smaller than 0.1. The situation with batch effects is less clear, especially for sequencing data which are sensitive to both sample processing and genotype calling methods. Finally, variants which are known to be associated with the trait may also be included as covariates, especially in fine mapping analyses or while searching for multiple independent signals within the same locus. Here, we recommend against using ATR based methods since there is potentially a large power loss for variants in even moderate linkage disequilibrium with the associated variant.

In multiple-variant tests such as burden and omnibus tests (like the F-test or SKAT), we note that least-squares effect size estimator for any particular variant may be biased either *towards or away* from the null for ATR. Thus, although ATR based tests are valid for the omnibus hypothesis that none of the variants are associated, an ATR based test for the conditional effect of a variant given the remaining variants may not be valid. This is of particular importance for *post-hoc* testing when the omnibus test is rejected and the analyst wishes to identify the subset of variants driving the association. We recommend against using ATR for such purposes.

When the distribution of the trait differs substantially from the normal distribution, ATR based methods are commonly used in conjunction with applying the inverse normal transform to the adjusted trait. Sofer et al. (2019) show that testing for association between the transformed adjusted trait and unadjusted variants may lead to increased Type 1 error and instead recommend using adjusted variants. McCaw et al. (2019) implement an omnibus test with this strategy.

Finally, we have assumed throughout that the multiple linear model (M1) is appropriate to answer the research question at hand and that β truly measures the effect of interest. This necessitates including certain covariates (e.g. confounders), *excluding* others (e.g. colliders; see Greenland et al., 1999) and accounting for sample-selection effects (Munafo et al., 2018). For example, Aschard et al. (2015) show that simply adjusting for heritable covariates may lead to biased estimates of the direct (unmediated) effect of the variant on the trait and may lead to increased Type 1 error. We note that when OLS analysis of the full regression model results in increased Type 1 error, ATR will also be unable to fully control Type 1 error (although, the magnitude of Type 1 error will be lower with increasing R^2). Thus, ATR is invalid whenever OLS analysis of the full regression model is invalid.

In summary, we derive distributions of the ATR analogs of commonly used association test statistics. We show that ATR based methods are conservative when variants are correlated with covariates. We quantify the power loss and recommend that

ATR based methods be used only when the squared correlation between variants and covariates can be confidently bounded to be substantially smaller than 0.1. We note that for commonly included covariates like age, gender and known or inferred ancestry, this is typically true and ATR based methods will likely result in negligible power loss. However, we reiterate that ATR is an ad-hoc methodology. Thus, we recommend that analysts carefully choose covariates based on a plausible causal model (accounting for sample-selection effects) and employ estimation/hypothesis-testing methods that are theoretically justified for those models.

APPENDIX

All notation in the Appendix is as defined in the main text.

ATR estimator for $\boldsymbol{\beta}$

The OLS estimator for β is given by $\hat{\beta} = [G'(I_n - H_C)G]^{-1}G'Y_r$ where $Y_r = (I_n - H_C)Y$ is the residual vector obtained from regressing Y onto C, and $H_C = C(C'C)^{-1}C'$. Note that $Var(\hat{\beta}) = \sigma^2[□'(I_n - H_C)G]^{-1}$. Since ATR simply replaces Y and C with Y_r and J_n, we have

$$\begin{aligned}
\hat{\beta}_{ATR} &= [G'(I_n - H_1)G]^{-1}G'(I_n - H_1)Y_r \\
&= [G'(I_n - H_1)G]^{-1}G'Y_r \\
&= [G'(I_n - H_1)G]^{-1}[□'(I_n - H_C)G]\hat{\beta} \\
&= (I_m - [G'(I_n - H_1)G]^{-1}[G'(H_C - H_1)G])\hat{\beta} \\
&\overset{\text{def}}{=} (I_m - R_{GC}^2)\hat{\beta}
\end{aligned}$$

The second equality holds because $(I_n - H_1)Y_r = Y_r - \overline{Y_r}$ (where $\overline{Y_r}$ is the sample mean of Y_r) and $\overline{Y_r} = 0$. The third equality holds because $G'Y_r = [G'(I_n - H_C)G]\hat{\beta}$ which follows from the expression for $\hat{\beta}$. The fourth equality follows with straightforward algebra. Note that the eigenvalues of $R_{GC}^2 \overset{\text{def}}{=} [G'(I_n - \square_1)G]^{-1}[G'(H_C - H_1)G]$ are the canonical correlations between G and C. Thus, when each variant is uncorrelated with all the covariates, all the eigenvalues of R_{GC}^2 are 0 and $\hat{\beta}_{ATR} = \hat{\beta}$.

When the model contains only one variant ($m = 1$), we have $[G'(I - H_1)G]^{-1}[G'(I - H_C)G] = 1 - R^2$ where R^2 is the coefficient of determination obtained by regressing the variant onto the covariates.

Relationship between the score test statistic and its ATR analog

The score test-statistic for testing $H_0: \beta = 0$ is given by

$$S = \frac{1}{\tilde{\sigma}^2}\hat{\beta}'G'(I_n - H_C)G\hat{\beta}$$

where $\tilde{\sigma}^2 = \frac{1}{n}Y'(I_n - H_C)Y = \frac{1}{n}Y_r'Y_r$ is the maximum likelihood estimator (MLE) for σ^2 under the null (Vandaele 1981).

Note that $\tilde{\sigma}_{ATR}^2 = \frac{1}{n}Y_r(I_n - H_1)Y_r = \frac{1}{n}Y_r'Y_r = \tilde{\sigma}^2$ since $(I_n - H_1)Y_r = Y_r - \bar{Y}_r = Y_r$. Thus, we have

$$\begin{aligned}
S_{ATR} &= \frac{1}{\tilde{\sigma}^2}\hat{\beta}_{ATR}'G'(I_n - H_1)G\hat{\beta}_{ATR} \\
&= \frac{1}{\tilde{\sigma}^2}\,\hat{\beta}'[G'(I_n - H_C)G][G'(I_n - H_1)G]^{-1}\,[G'(I_n - H_C)G]\hat{\beta} \\
&= \frac{1}{\tilde{\sigma}^2}\,\hat{\beta}'[G'(I_n - H_C)G][I_m - R_{GC}^2]\hat{\beta} \\
&= S - \frac{1}{\tilde{\sigma}^2}\,\hat{\beta}'[G'(I_n - H_C)G]R_{GC}^2\hat{\beta}
\end{aligned}$$

Equivalently, we have

$$\frac{S_{ATR}}{S} = \frac{\hat{\beta}'[G'(I_n - H_C)G][I_m - R_{GC}^2]\hat{\beta}}{\hat{\beta}'G'(I_n - H_C)G\hat{\beta}}$$

Recall that, for all vectors x such that $x'Bx = c$ (for any constant $c > 0$) the generalized Rayleigh quotient $Q = \frac{x'Ax}{x'Bx}$ is bounded below and above by the minimum and maximum eigenvalues of $B^{-1}A$. Thus, we have

$$(1 - R^2_{max})S \leq S_{ATR} \leq (1 - R^2_{min})S$$

where R^2_{min} and R^2_{max} are the smallest and largest eigenvalues of $R^2_{\square C}$. The lower (upper) bound is attained when $\hat{\beta}$ is parallel to the eigenvector corresponding to maximum (minimum) eigenvalue of R^2_{GC}. When each variant is orthogonal to each of the covariates we have $R^2_{min} = R^2_{max} = 0$ and $S_{ATR} = S$.

When the model contains only one variant, the above relationship simplifies to the deterministic relationship $S_{ATR} = (1 - R^2)S$ (with R^2 as defined previously). For $m > 1$, the relationship is not deterministic (that is, S_{ATR} can take multiple values for any given value of S) unless all the variants are collinear. We can use the relationships between the score, Wald, likelihood-ratio, and F test statistics (Vandaele 1981) to derive exact expressions for the relationships between each of these tests and their ATR analogs for single variant models. We state these relationships in the main text (but omit the straightforward algebra).

Asymptotic distribution of S_{ATR}

Asymptotically, S_{ATR} converges in distribution to the distribution of the quadratic form $\hat{\beta}'A\hat{\beta}$ with $A = \sigma^{-2}[G'(I_n - H_C)G][G'(I_n - H_1)G]^{-1}\,[G'(I_n - H_C)G]$. With suitable regularity conditions, asymptotically $\hat{\beta} \sim N(\mu, V)$ with $V = \sigma^2[G(I_n - H_C)G]^{-1}$. Baldessari (1967) derived the distribution of quadratic forms in multivariate normal variables. Since A is symmetric and V positive definite, there exists an invertible matrix M such that

$M'V^{-1}M = I_m$ and $M'AM = \Lambda$ with Λ an $m\times m$ diagonal matrix. Thus, we have that $I_m - R_{GC}^2 = VA = M\Lambda M^{-1}$; that is, the columns of M are the eigenvectors of $I_m - R_{GC}^2$ (and R_{GC}^2) and the i^{th} element of the diagonal of Λ is $1 - l_i^2$ with l_i^2 the eigenvalue of R_{GC}^2 corresponding to the i^{th} column of M. Let $R_j^2, 1 \leq j \leq p$ denote the $p \leq m$ distinct eigenvalues of R_{GC}^2 with $R_1^2 > \cdots > R_p^2$. Let B_j be the $m\times m$ diagonal matrix which has elements 1 where Λ has elements $1 - R_j^2$ and 0 otherwise. Then, from Baldessari (1967, Theorem 1) and some trivial algebra, S_{ATR} follows the same distribution as $\sum_{j=1}^{p}(1 - R_j^2)Z_j$, where $Z_j \sim \chi_{\nu_j}^2(\lambda_j^2)$ (that is, a non-central chi-squared distribution with ν_j degrees of freedom and non-centrality parameter λ_j^2), $\lambda_j^2 = (M^{-1}\beta)'B_jM^{-1}\beta$ and ν_j is the geometric multiplicity of R_j^2.

Recall that, asymptotically, $S \sim \chi_m^2(\delta^2)$ with $\delta^2 = \beta'V^{-1}\beta = (M^{-1}\beta)'M^{-1}\beta$. Thus, we have $\sum_{j=1}^{p}\lambda_j^2 = \delta^2$. When β lies in the space spanned by the eigenvector(s) of R_{GC}^2 corresponding to the (distinct) eigenvalue R_k^2, $1 \leq k \leq p$, we have $\lambda_k^2 = \delta^2$ and $\lambda_i^2 = 0, i \neq k$. Consider the set □ of vectors β that yield the same power for the score test (that is, all vectors β for which $\beta'V^{-1}\beta = \delta^2$ for a given δ^2). Unlike S, the power of S_{ATR} may differ when β takes different values in this set. We use a result derived by Matthew and Nordstöm (1997) to find values in Δ that lead to minimum power for S_{ATR}:

Theorem 3 (Matthew and Nordstöm, 1997). *Let X_i and Y_i be distributed, respectively, as $\chi_{\nu_i}^2(\delta_i^2)$ and $\chi_{\nu_i}^2(\mu_i^2)$, $i = 1, \ldots, n$, with $X_1, \ldots, X_n$ independent and $Y_1, \ldots, Y_n$ independent. Then*

$$\sum_{i=i}^{n} \lambda_i X_i \leq_D \sum_{i=1}^{n} \lambda_i Y_i$$

holds for all nonnegative λ_i's satisfying $\lambda_1 \geq \cdots \geq \lambda_n$ if and only if

$$\sum_{i=1}^{k} \delta_i^2 \leq \sum_{i=1}^{k} \mu_i^2 \text{ for all } k = 1, \ldots, n.$$

In the above theorem, $X \leq_D Y$ denotes that the random variable Y stochastically dominates X. From the above theorem and preceding details of the distribution of S_{ATR}, it follows that distribution followed by S_{ATR} when β lies in the space spanned by the eigenvectors of R_{GC}^2 corresponding to the maximum eigenvalue $R_1^2 = R_{max}^2$ is dominated by the distribution followed by S_{ATR} when β takes any other value in Δ.

REFERENCES

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., & Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, *96*(2), 329-339.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Cortes, A. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203-209.

Che, R., Motsinger-Reif, A. A., & Brown, C. C. (2012). Loss of power in two-stage residual-outcome regression analysis in genetic association studies. *Genetic epidemiology*, *36*(8), 890-894.

Demissie, S., & Cupples, L. A. (2011). Bias due to two-stage residual-outcome regression analysis in genetic association studies. *Genetic epidemiology*, *35*(7), 592-596.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, 387-401.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37-48.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., ... & Kubo, M. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature genetics*, *50*(3), 390-400.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, *95*(1), 5-23.

Lovell, M. C. (2008). A simple proof of the FWL theorem. *The Journal of Economic Education*, *39*(1), 88-91.

Mathew, T., & Nordström, K. (1997). Inequalities for the probability content of a rotated ellipse and related stochastic domination results. *The Annals of Applied Probability*, *7*(4), 1106-1117.

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., & Lin, X. (2019). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*.

de Micheaux, P. L., & de Micheaux, M. P. L. (2017). Package 'CompQuadForm'. *CRAN Repository*.

Monsees, G. M., Tamimi, R. M., & Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*(8), 717-728.

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: when selection bias can substantially influence observed associations. *International journal of epidemiology*, *47*(1), 226-235.

Niarchou, M., Byrne, E. M., Trzaskowski, M., Sidorenko, J., Kemper, K. E., McGrath, J. J., ... & Wray, N. R. (2020). Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits. *Translational Psychiatry*, *10*(1), 1-11.

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., ... & Workalemahu, T. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet*, *9*(6), e1003500.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, 227-240.

Sofer, T., Zheng, X., Gogarten, S. M., Laurie, C. A., Grinde, K., Shaffer, J. R., ... & Lange, L. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic epidemiology*, *43*(3), 263-275.

Styrkarsdottir, U., Stefansson, O. A., Gunnarsdottir, K., Thorleifsson, G., Lund, S. H., Stefansdottir, L., ... & Ivarsdottir, E. V. (2019). GWAS of bone size yields twelve loci that also affect height, BMD, osteoarthritis or fractures. *Nature communications*, *10*(1), 1-13.

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R., Steinberg, J., Walter, K., ... & McCarthy, S. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *The American Journal of Human Genetics*, *100*(6), 865-884.

UK10K consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82-90.

Vandaele, W. (1981). Wald, likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters*, *8*(4), 361-365.

Xing, G., Lin, C. Y., & Xing, C. (2011). A comparison of approaches to control for confounding factors by regression models. *Human heredity*, *72*(3), 194-205.

Chapter 3 A comparison of predicted and observed imputation quality for rare and low-frequency rare variants

INTRODUCTION

The advent of high-throughput DNA sequencing has made it possible to assay nearly the complete spectrum of human genetic variation. However, genome sequencing remains expensive for large studies. Genome-wide genotyping with dense genotype arrays followed by imputation based on sequenced reference panels is a cost-effective alternative to sequencing. This strategy is used routinely in genome-wide association studies (GWAS) to improve genome coverage, facilitate meta-analyses and increase power. Availability of larger reference panels is improving imputation accuracy for low-frequency $(0.5\% \leq MAF < 5\%)$ and rare variants $(MAF < 0.5\%)$ and has the potential to reduce the need for sequencing.

Genotype imputation uses information from a densely typed reference panel (e.g. HapMap, 1000 Genomes, Haplotype Reference Consortium, CAAPA, or TOPMed) to predict the genotypes at markers that are genotyped/sequenced in the reference panel but not in the study participants. Commonly used imputation algorithms such as Beagle 4.1 (Browning and Browning, 2007), Impute2 (Howie et al., 2009), and Minimac3 (Das et al., 2016) calculate posterior probabilities for each possible genotype (at sites where genotypes are unobserved/missing) conditional on observed genotypes of nearby variants. Under the assumption of an additive genetic model, single-variant association analyses typically use imputed genotype dosages (predicted expected allele counts for the unobserved genotypes) as a substitute for the (unknown) true genotypes since, in contrast to best-guess genotypes (genotypes with maximum posterior probability), dosages account for genotype uncertainty (Zheng et al., 2011).

In this paper, we measure variant level imputation quality as the squared correlation between dosages and true genotypes since this is the most commonly used measure. We refer to this measure as the ‘realized imputation quality’ or the ‘realized RSQ’. Since true genotypes are unobserved in practice for imputed variants, the realized imputation quality is also unknown. Beagle, IMPUTE, and Minimac provide imputation quality scores (called allelic-R2, INFO, and MACH R2, respectively) for each imputed variant. Each of these quality scores summarizes the uncertainty in the genotype posterior probability distributions with apparently different metrics. Each score can be calculated from the output of any imputation program that provides imputed genotype posterior probabilities as part of its output. We note that these scores only reflect the genotype prediction uncertainty, and cannot reflect bias due to miscalibration which occurs when the imputed genotype posterior probabilities do not match the true posterior probabilities. However, realized imputation quality reflects both the genotype prediction uncertainty and miscalibration bias.

Many studies use quality scores to filter out (presumed) poorly imputed variants from all downstream analyses. For example, common variants with MACH R2 < 0.3 or IMPUTE INFO < 0.4 are typically filtered out. All three estimated quality scores are highly correlated with the realized RSQ between dosages and genotypes and yield high sensitivity and specificity when used as classifiers of well (realized RSQ > 0.5) and poorly imputed common variants. Recent studies (Deelen et al., 2014; Pistis et al.,

2015) have shown that the classification accuracy of two of these scores (MACH R2 and INFO) decreases with decreasing minor allele frequency and have suggested that more stringent cutoffs are required for filtering low-frequency and rare variants. However, none of these studies calculated the three scores based on the results of the same imputation algorithms, making it impossible to attribute differences in the behavior of the quality scores to the scores themselves or to differences in the underlying imputation algorithms and the corresponding genotype posterior probabilities. Further, no study has assessed whether the accuracy of quality scores improves when using large reference panels like the Haplotype Reference Consortium (HRC) panel.

The goal of this study is to assess the reliability of quality scores as indicators of observed imputation quality for low-frequency and rare variants and to assess how this reliability depends on the choice of imputation algorithm, quality score and reference panel.

We performed whole-genome imputation (using Beagle 4.1, IMPUTE2 and Minimac 3) on 8,378 Finnish participants in the METSIM study (Stančáková et al., 2009) that were genotyped on the Illumina OmniExpress GWAS array (used as the imputation backbone). We also submitted these data to the Michigan Imputation Server (Das et al., 2016) for imputation using the HRC panel (McCarthy et al., 2016). We used genotypes on the same participants from the separately genotyped Illumina ExomeChip array as the “gold-standard” or true genotypes; the exome array includes large numbers of common, low-frequency and rare variants.

We calculated all three quality scores from the output from each of the three imputation algorithms. As expected, we found that the classification accuracy of quality scores decreases as a function of minor allele frequency regardless of which imputation algorithm and quality score are used. The reduction in classification accuracy is attributable to the poorer realized imputation quality for low-frequency and rare variants. The classification accuracy for low-frequency and rare variants is better when using the HRC reference panel (relative to the classification accuracy when using 1KG) due to substantially better realized imputation accuracy afforded by the larger reference panel size. We find that that an INFO/MACH RSQ cutoff of 0.3 works as well for classification with HRC based imputation as does a cutoff of 0.6 for imputation using the 1KG panel.

METHODS

Description of sample and genotype data

METSIM is a population based study of 10,197 men aged 45 to 73 years randomly selected from the population register of Kuopio, Finland (Stančáková et al., 2009). We included in our analysis the 8,378 METSIM participants with less than second-degree relatedness to any other participant. Study participants were genotyped with both the Illumina HumanOmniexpress-12v1_C array (OmniExpress) and the Illumina HumanExome-12v1_A array (ExomeArray). The ExomeArray contains a large number of variants that are classified as low-frequency and rare in Europeans. Approximately 600,000 phased variants that passed quality control (QC) for the OmniExpress were used as the imputation backbone; details of QC and phasing can be found in Teslovich et al. (2018). We treated ExomeArray genotypes passing QC as "gold-standard" (true) genotypes. Details of ExomeArray QC can be found in Huyghe et al. (2013).

Imputation procedures

We used the 1KG Phase 3 phased reference panel VCF file available from the minimac3 website (https://genome.sph.umich.edu/wiki/Minimac3) as the reference panel input for Beagle, minimac3 and IMPUTE2. We transformed the OmniExpress variant data using the *conform-gt* script available from the Beagle website (https://faculty.washington.edu/browning/conform-gt.html) to make strand and reference alleles consistent with the minimac3 reference panel VCF. We used the same transformed file as the backbone input to all three algorithms and ran all algorithms with default options. In addition, we submitted our data to the Michigan Imputation Server (https://imputationserver.sph.umich.edu/) for imputation using the HRC reference panel (which contains approximately 65,000 haplotypes).The Michigan Imputation Server used the minimac3 algorithm to perform imputation.

Definitions of quality scores

Three genotype imputation quality scores are commonly used. Allelic-RSQ is defined as the squared correlation between imputed genotype dosages and genotypes with maximum posterior probability (Marchini and Howie, 2010). MACH-RSQ is defined as the ratio between the variance of (haplotype) dosages (based on allelic posterior probabilities rather than genotype posterior probabilities) and the variance of a Bernoulli random variable with the same mean as the dosages (). The INFO quality score is defined as the relative information about the population allele-frequency contained in the imputed data (Marchini and Howie, 2010).

Formulae for the scores are as follows. Following the notation in Marchini and Howie (2010), for a given variant let $\mathrm{e_i}$ denote the imputed dosage for individual i, and $\mathrm{z_i}$ the imputed genotype with maximum posterior probability. Let $\mathrm{p_{ij}}$ $(\mathrm{j} \in \{0,1,2\})$ denote the imputed posterior probability that the alternate allele count for individual i is j. Finally, let h_{i1} & h_{i2} denote the imputed posterior probability of observing the alternate allele for two chromosomes of individual i. Define $\mathrm{f_i = p_{i1} + 4p_{i2}}$, and $\hat{\theta} = \sum_\mathrm{i} \mathrm{e_i} / 2\mathrm{N}$ (the estimated allele frequency), where N is the sample size.

The MACH-RSQ quality score is calculated as

$$\frac{\sum_\mathrm{i} \frac{h_{i1}^2 + h_{i2}^2}{2N} - \hat{\theta}^2}{\hat{\theta}(1-\hat{\theta})}$$

The allelic-RSQ score is calculated as

$$\frac{\left[\sum_\mathrm{i} \mathrm{z_i e_i} - \left(\frac{1}{\mathrm{N}}\right)\left(\sum_\mathrm{i} \mathrm{z_i} \sum_\mathrm{i} \mathrm{e_i}\right)\right]^2}{\left[\sum_\mathrm{i} \mathrm{f_i} - \left(\frac{1}{\mathrm{N}}\right)\sum_\mathrm{i} \mathrm{e_i}\right]^2 \left[\sum_\mathrm{i} \mathrm{z_i^2} - \left(\frac{1}{\mathrm{N}}\right)\sum_\mathrm{i} \mathrm{z_i}\right]^2}$$

The INFO score is calculated as

$$1 - \frac{\sum_\mathrm{i}(\mathrm{f_i} - \mathrm{e_i^2})}{2\mathrm{N}\hat{\theta}(1-\hat{\theta})}$$

We note that the definition for MACH-RSQ used here differs from the one in Li et al. (2010) and Marchini and Howie (2010) but matches the formula currently employed by the minimac software (Sayantan Das, personal communication). Note that each quality score can be computed given only genotype posterior probabilities (since $h_{i1}^2 + h_{i2}^2 = e_i^2 - 2p_{i2}$). Thus, although Beagle, IMPUTE and minimac output different quality scores (allelic-RSQ, INFO and MACH-RSQ respectively), each quality score can be computed on imputation results using any of the three algorithms, since each algorithm provides an option to output genotype posterior probabilities.

To distinguish between trends in the results attributable to differences between scores as opposed differences between algorithms, we calculated each score on the output from each algorithm (to yield a total of nine score-algorithm pairs). It is easy to show that MACH-RSQ (as defined above) and INFO are numerically identical when calculated on the same imputed genotype posterior probabilities.

RESULTS

Imputation with 1KG reference panel

43,073 imputed variants were also present on the ExomeArray. Of these, 21,790 are rare, 10,707 low-frequency and 10,576 common. The realized RSQ decreased with decreasing minor-allele frequency and had a bimodal distribution for low-frequency and rare variants (Figure 3-1 shows results for minimac; results for Beagle and IMPUTE are similar).

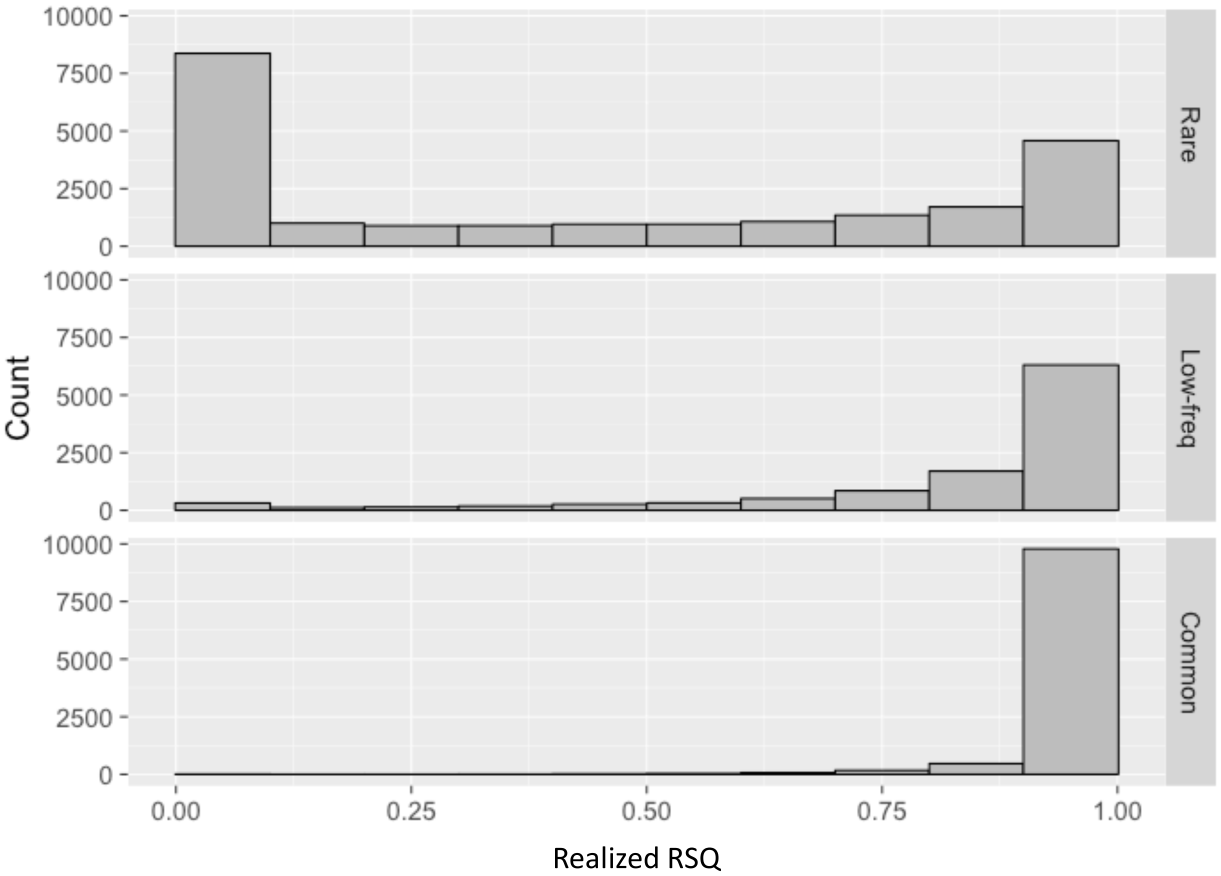

Figure 3-1 Histogram of realized RSQ stratified by allele-frequency group; rows show data for rare (top), low-frequency and common variants (bottom). Results are for imputation with minimac3 using 1KG reference panel.

We found that allelic-RSQ is a poor discriminator of realized RSQ for low-frequency and rare variants with realized RSQ typically being lower than allelic-RSQ except when allelic-RSQ was close to zero. Since allelic-RSQ works poorly for rare variants, we restrict attention to the INFO quality score for the remainder of this paper. The distribution of realized RSQ given INFO had larger variances and longer tails for rare and low-frequency variants (Figure 3-2) as compared to common variants, especially for lower INFO values, implying greater uncertainty about realized RSQ for a given INFO value. Using an INFO threshold of 0.3 (commonly employed for common variants) to identify variants with realized RSQ > 0.5 yields positive predictive values

(PPVs) of 99.7%, 93% and 74% for common, low-frequency and rare variants for imputation with minimac3; the corresponding sensitivities are 99.7%, 99% and 85%. These values are similar for Beagle except for rare variants where it yields PPV and sensitivity of 71% and 92% respectively. Using a more stringent INFO threshold of 0.6 yields PPVs 99.9%, 97% and 87% for minimac3; the corresponding sensitivities are 99.7%, 94% and 60%.

IMPUTE2 tends to systematically predict larger than observed values for realized RSQ and has slightly lower accuracy of classification. The PPVs for INFO thresholds of 0.3 and 0.6 are 99.5%, 89%, 49% and 99.5%, 92%, 65%.

In summary, both the PPV and sensitivity of INFO for classifying variants with realizes RSQ > 0.5 decreases with decreasing minor allele frequency regardless of the algorithm used to perform imputation. Increasing the threshold from 0.3 to 0.6 results in better PPVs at the cost of sensitivity.

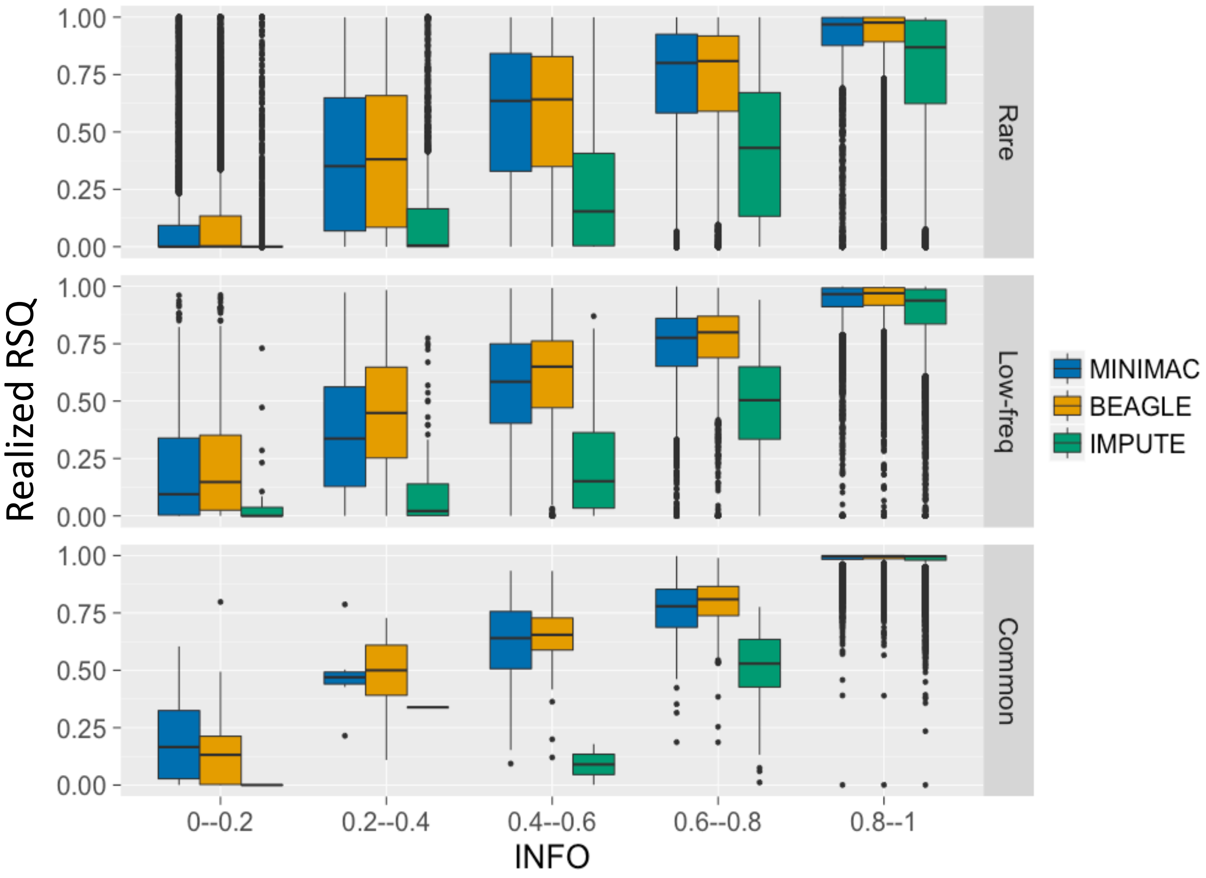

Figure 3-2 Boxplots of realized RSQ within INFO bins of width 0.2. Rows show data for rare (top), low-frequency and common variants (bottom). Colors represent imputation algorithms. Data are for imputation using the 1KG reference panel.

Imputation with HRC reference panel

HRC imputation yielded 65,382 imputed variants (39,799 rare, 13,317 low-frequency, and 12,266 common) also present on the ExomeArray. The mean imputation quality was substantially better than imputation based on 1KG, especially for low-frequency and rare variants (Figure 3-3). Realized RSQ was greater than 0.5 for 68% of rare and 99% of low-frequency variants (compared to 44% and 90% for imputation with 1KG using minimac3).

The conditional variance of realized RSQ given INFO was slightly larger for HRC compared to 1KG. Also, for rare variants, the HRC panel imputation INFO tended to slightly underestimate the realized RSQ when the INFO values were low. An INFO threshold of 0.3 yielded PPVs of 99.8%, 99% and 87% for common, low-frequency and rare variants respectively with sensitivities of 99.9%, 99.9% and 93% respectively. Thus, for the HRC panel an INFO threshold of 0.3 worked as well as a threshold of 0.6 for the 1KG panel with respect to the PPV (and better in terms of sensitivity).

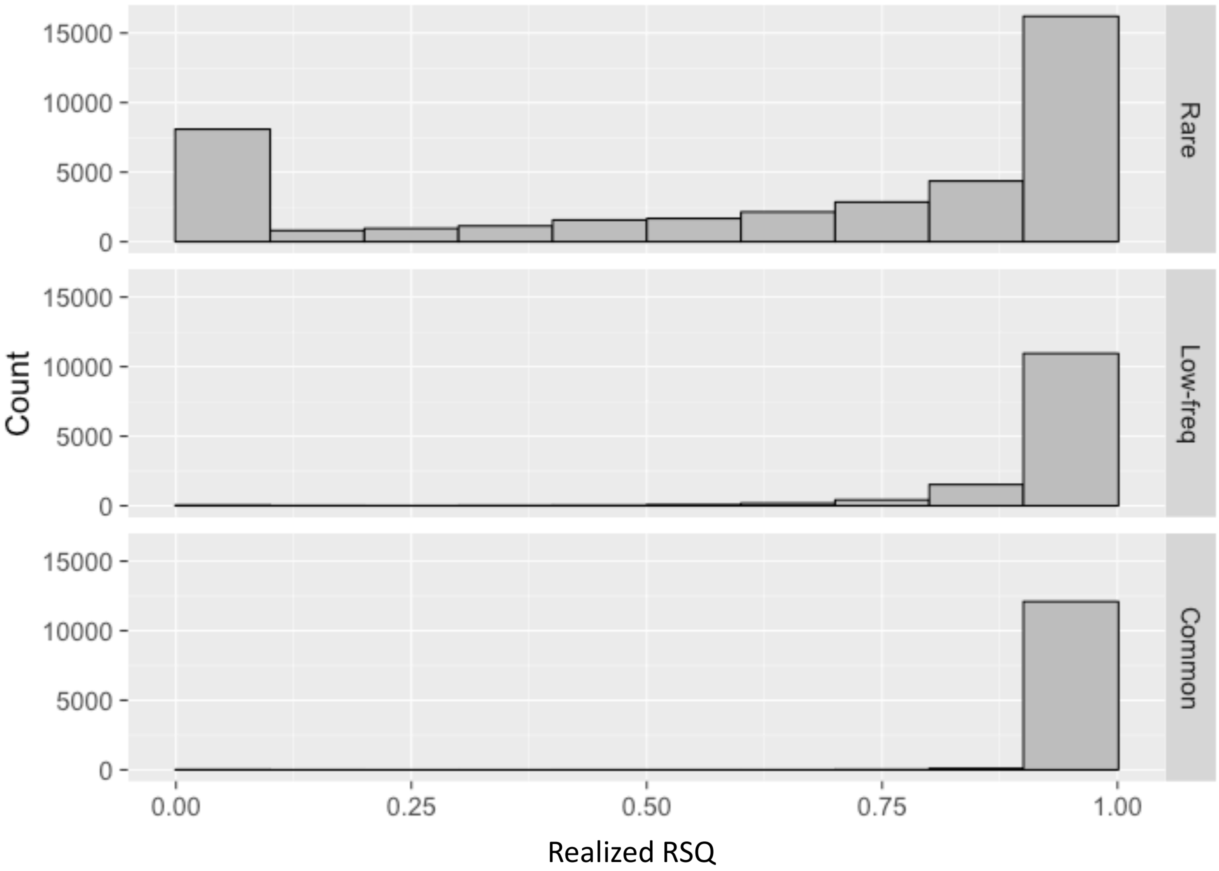

Figure 3-3 Histogram of realized RSQ stratified by allele-frequency group; rows show data for rare (top), low-frequency and common variants (bottom). Data are for imputation with minimac3 using HRC reference panel.

DISCUSSION

We assessed the impact of imputation algorithm, reference panel size, allele frequency of imputed variant and quality score definition on the utility of imputation quality scores as an indicator of realized quality of imputed variants. We reiterate that imputation quality scores are only informative about the precision of imputed genotypes insofar as the inferred posterior distribution of genotypes is well calibrated.

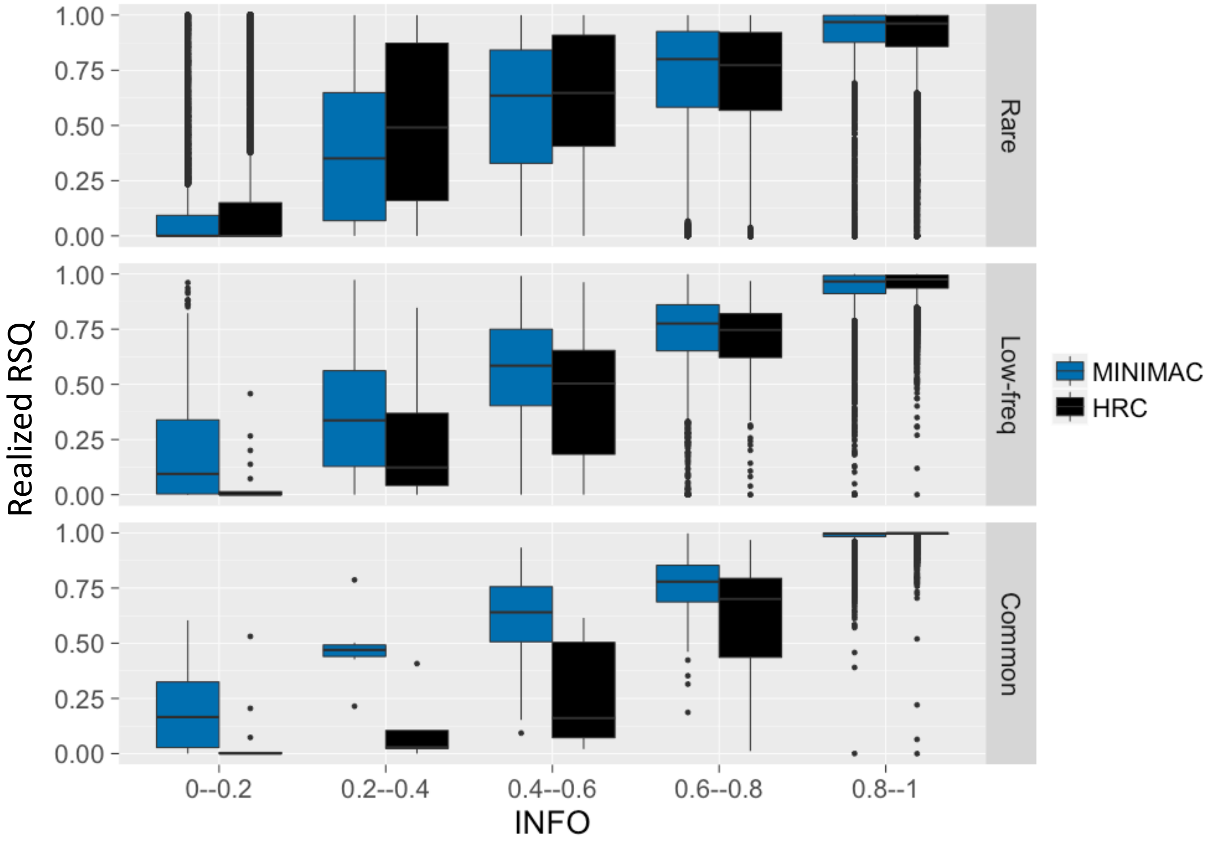

Figure 3-4 Boxplots of realized RSQ within INFO bins of width 0.2. Rows show data for rare (top), low-frequency and common variants (bottom). Colors represent reference panel. Blue represents imputation with minimac3 using 1KG reference panel. Black represents imputation with minimac3 using HRC reference panel.

We showed that INFO and MACH-RSQ (as currently implemented in minimac3) are numerically equal when calculated on the same imputed data. Further, we showed that allelic-RSQ performs poorly for low-frequency and rare variants as compared to INFO when calculated on the same data. Thus, we recommend implementation of INFO as a standard imputation quality score for any algorithm that calculates genotype posterior probabilities as part of the imputation procedure.

The IMPUTE2 imputations have both slightly lower realized imputation quality and poorer concordance between quality scores and realized quality as compared to minimac3 and Beagle. This may be because we used default parameter settings to run each algorithm or due to inherent differences between the algorithms. For IMPUTE2, the k_hap parameter controls the number of reference haplotypes used to perform imputation, with larger values resulting in greater imputation accuracy but increased run-time. We re-ran imputation with IMPUTE2 on chromosome 1 changing the default k_hap setting from 500 to 1000. We observed a slight improvement in realized quality and also for classification; the PPV for rare variant classification increased from 49% (k_hap=500) to 54% (k_hap=1000). Increasing k_hap further could result in greater improvement at the cost of run-time.

The ability of imputation quality scores to reflect realized imputation quality is reduced for low-frequency and rare variants for two reasons. First, the conditional variance of observed quality for a given imputation quality score is slightly higher for rarer variants (especially for variants with INFO < 0.8; see Figures 3-2 and 3-4). Second, the ability to classify well (realized RSQ > 0.5) and poorly imputed variants is reduced for rarer variants. The classification ability is influenced by the distribution of realized quality. Almost all common variants have realized RSQ > 0.8 (even for the smaller 1KG panel); the proportion of poorly imputed variants grows with decreasing allele frequency. Thus, a stringent INFO threshold of 0.6-0.7 should be used to filter out poorly imputed rare and low-frequency variants when using the 1KG reference panel. In contrast, for data imputed with the HRC panel, the currently used INFO threshold of 0.3 works well for low-frequency variants and adequately for rare variants. This result is driven primarily by the fact that realized RSQ values are higher for low-frequency and rare variants for data imputed with the HRC reference panel. Thus, using the HRC reference panel improves the proportion of correctly classified high quality variants (as compared to the 1KG reference panel) due to both, the higher average observed imputation quality and the lower acceptable threshold of quality score used for classification.

REFERENCES

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084-1097.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., ... & Schlessinger, D. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, *48*(10), 1284-1287.

Deelen, P., Menelaou, A., Van Leeuwen, E. M., Kanterakis, A., Van Dijk, F., Medina-Gomez, C., ... & Kreiner-Møller, E. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European Journal of Human Genetics*, *22*(11), 1321-1326.

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

Huyghe, J. R., Jackson, A. U., Fogarty, M. P., Buchkovich, M. L., Stančáková, A., Stringham, H. M., ... & Chines, P. S. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics*, *45*(2), 197-201.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics, 11(7), 499.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... & Luo, Y. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, *48*(10), 1279-1283.

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., ... & Brennan, C. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. European Journal of Human Genetics, 23(7), 975.

Stančáková, A., Javorský, M., Kuulasmaa, T., Haffner, S. M., Kuusisto, J., & Laakso, M. (2009). Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes, 58(5), 1212-1221.

Teslovich, T. M., Kim, D. S., Yin, X., Stančáková, A., Jackson, A. U., Wielscher, M., ... & Davis, J. P. (2018). Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study. *Human molecular genetics*, *27*(9), 1664-1674.

Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic epidemiology*, *35*(2), 102-110.

Chapter 4 Increasing statistical power of disease association studies by including external control

INTRODUCTION

The two-sample comparison framework of case-control studies has proven remarkably successful for uncovering genetic association with disease and genome-wide association studies (GWAS) have successfully identified thousands of disease-variant associations over the last decade. However, the percentage of heritability explained by currently identified loci is low for many diseases, especially polygenic diseases which tend to have complex genetic and environmental etiology.

Recently, databases like the UK Biobank (Bycroft et al., 2018), Genome Aggregation Database (gnomAD) (Karczewski et al., 2020), and Exome Aggregation Consortium (ExAC) (Lek et al., 2016) with high-quality genetic information on large numbers of individuals have emerged. These data are freely available to the scientific community and may be added as additional *external* data into an existing *internal* case-control dataset to increase sample size. This approach has been recently used for single variant association testing (NINDS Stroke Genetics Network and International Stroke Genetics Consortium, 2016) and gene-based burden tests (Guo et al., 2018; Hendricks et al., 2018) where, in each study, the external individuals were included as controls.

Although including external data has the potential to increase power, there are multiple challenges to using this approach. Differences in genetic ancestry between external and internal individuals may lead to bias due to confounding if disease prevalence varies between populations. Selection bias (Hernan et al., 2014) may occur due to different strategies of participant recruitment or differences in willingness of individuals to participate (Aigner et al., 2018) between internal and external studies. Bias may also occur due to differential measurement errors between the internal and external samples, especially differential genotyping/sequencing quality due to differences in sample collection, preparation and analysis (Guo et al., 2018; Hendricks et al., 2018). Lastly, the external sample may be missing information about case-control status for the particular disease of interest.

In this paper, we focus on the consequences of missing case-control status in the external sample and assume that other challenges do not occur, or have been successfully dealt with prior to association analysis (for example, through stringent pre-analysis quality control). Missing case-control status may be a common occurrence in large, public datasets since they are typically designed as a general resource rather than being focused on particular phenotypes. When the proportion of cases in the external sample is confidently known to be low, a naïve analysis strategy is to treat each external individual as a control. This strategy enables analysis of the combined (internal and external) data using existing methods and software (for example, logistic regression).

Even though case-control status may be missing at the individual level, the proportion of individuals in the external sample that are cases may be accurately know (for example, when the external sample is designed to be representative of a population and disease prevalence of that population is accurately known). Lancaster and Imbens

(1994) propose a method-of-moments estimation procedure for studies with known/validated cases and individuals with unknown case-control status (which they call contaminated controls). Ward et al. (2009) present an EM algorithm to estimate parameters in the same setting (a design that is called “presence-only design” in ecology). Both of these methods cannot incorporate internal (validated) controls. Thornton & McPeek (2010) present a score test based on estimating equations that is primarily designed to deal with missing genotypes in related individuals but also accounts for missing case-control status in a subset of individuals by incorporating a presumptive case proportion (or prevalence) parameter; this method cannot incorporate covariates.

We propose a method (and corresponding association test) that accounts for known external sample case proportion by modeling the external sample as a mixture of cases and controls. Analysis treating external individuals as controls occurs as a special case and is obtained by setting external sample case proportion to zero. The proposed method allows inclusion of internal controls in addition to internal cases and can also incorporate covariates.

We derive a closed-form expression for the score test non-centrality parameter to assess analytically the efficiency of the proposed method. We show that incorrect specification of the external sample case proportion leads to biased effect size (odds-ratio) estimates under the alternative but not under the null; in particular, treating external individuals as controls leads to underestimation of odds-ratios. Incorrect specification of the external sample case proportion does not lead to increased Type 1 error but may lead to decreased power when the external sample case proportion is an appreciable fraction of proportion of cases in the internal sample. In particular, when treating the external individuals as controls, including external data leads to decreased power (relative to analyzing only internal participants) when the external sample case proportion is more than half the internal sample case proportion. In such a situation, accurate specification of the external sample case proportion may still allow use of the external data to increase power.

METHODS

Model

We model the data as consisting of three groups: internal (that is, truly affected) cases, internal (truly unaffected) controls and the external sample. For ease of exposition, we assume no misclassification of case/control status in the internal sample (but, as we describe in the discussion, our model is easily extended to relax this assumption). We assume that external individuals (with missing case/control status) are recruited from the same pools/populations of cases and controls from which internal cases and internal controls are recruited (that is, we assume that selection bias is absent); thus, the external sample differs from the internal sample only due to missing case/control status. Further, we assume that each external individual is a case with probability π or control with probability $1 - \pi$. When the external sample is a simple random sample from the same population from which the internal cases and controls are recruited, π is simply the prevalence (Π) of the disease in that population. This model can be conceptualized as a partially missing data model, with case/control status missing in the external sample. Alternatively, when external individuals are assumed to be true controls (but may contain some non-zero proportion of cases), the model can be conceptualized as a misclassification model with misclassification present only in external sample.

We model the dependence of the probability of case/control status on the vector of predictors $\mathbf{x}$ via logistic regression: $P(\text{case}|\mathbf{x}) = e^{\theta(\mathbf{x})} \,/\, (1 + e^{\theta(\mathbf{x})})$ where $\theta(\mathbf{x}) = \alpha + \mathbf{x}'\boldsymbol{\beta}$. In the Appendix, we derive the prospective likelihood for data with n_0 internal controls, n_1 internal cases, and n_e external individuals as

$$P(\text{internal case}|\mathbf{x}) = (1-\gamma_1)\frac{e^{\delta(\mathbf{x})}}{1+e^{\delta(\mathbf{x})}},$$

$$P(\text{internal control}|\mathbf{x}) = (1-\gamma_0)\frac{1}{1+e^{\delta(\mathbf{x})}},$$

$$P(\text{external}|\mathbf{x}) = \gamma_0\frac{1}{1+e^{\delta(\mathbf{x})}} + \gamma_1\frac{e^{\delta(\mathbf{x})}}{1+e^{\delta(\mathbf{x})}},$$

where $\gamma_0 = \frac{(1-\pi)n_e}{n_0+(1-\pi)n_e}$, $\gamma_1 = \frac{\pi n_e}{n_1+\pi n_e}$, and $\delta(\mathbf{x}) = \alpha_D + \mathbf{x}'\boldsymbol{\beta}$. The prospective model intercept $\alpha_D = \alpha + \log\left(\frac{[n_1+\pi n_e][1-\Pi]}{[n_0+(1-\pi)n_e]\Pi}\right)$ accounts for the proportion of total cases (both known internal cases and unknown external cases) in the dataset as dictated by the design parameters n_0, n_1, n_e and the external sample case proportion π. We treat π as a fixed (or known) quantity. When $\pi = 0$ (or $\pi = 1$) the external individuals are treated as controls (or cases) and the usual (prospective) logistic-regression likelihood is recovered.

Straightforward calculus shows that the score with respect to $\boldsymbol{\beta}$ is $\sum_{i=1}^{n} \mathbf{x}_i(Y_i^* - \mu_i)$. Here $n = n_0 + n_1 + n_e$ is the total sample size, $\mu_i = e^{\delta(\mathbf{x}_i)}/(1 + e^{\delta(\mathbf{x}_i)})$ is the (prospective model) conditional probability of being a case given $\mathbf{x}_i$, and $Y_i^* = 0$ for internal controls, $Y_i^* = 1$ for internal cases, and $Y_i^* = e^{\eta(\mathbf{x}_i)}/(1 + e^{\eta(\mathbf{x}_i)})$ for external individuals, with $\eta(\mathbf{x}_i) = \log\left(\frac{\pi[n_0+(1-\pi)n_e]}{[1-\pi][n_1+\pi n_e]}\right) + \delta(\mathbf{x}_i) = \alpha + \log\left(\frac{\pi(1-\Pi)}{(1-\pi)\Pi}\right) + \mathbf{x}_i'\boldsymbol{\beta}$. For external individuals, Y_i^* can be interpreted as the conditional probability of being a case, given $\mathbf{x}_i$**,** in a population that is constructed by including cases with probability π and controls with probability $1 - \pi$. When the external sample is a simple random sample from the same population from which cases and controls are recruited, $Y_i^* = e^{\theta(\mathbf{x}_i)}/(1 + e^{\theta(\mathbf{x}_i)})$; that is, Y_i^* is simply the conditional probability of being a case (given $\mathbf{x}_i$) in the population from which all individuals are drawn. The score statistic is computed by replacing the parameter values in the expression for the score by their maximum-likelihood estimates (MLEs) under the null. Here, the $\hat{Y}_i^*$ (with the hat indicating MLEs) values computed for the external individuals can be interpreted as mean imputations (under the null model) for the unknown case-control status.

Score test-statistic for single-variant tests

To derive a closed-form non-centrality parameter, we focus on score tests for single-variant association analysis with no covariates. That is, we consider an association model with $\theta(G) = \alpha + \beta G$, where G is a (scalar) random variable representing allele count or imputation dosages for a bi-allelic genetic variant. Our interest is in testing the null hypothesis $\beta = 0$ (that is, the disease and variant are not associated). We assume an additive model, so the test described below is an extension of the trend-test for binary data.

Since the exact external sample case proportion (π) will rarely be known in practice, we assume that a presumptive value π^* is used for the analysis. Setting $\pi^* = 0$

is equivalent to treating the external individuals as controls; thus, the family of tests defined by the range of permissible π^* values ($0 \leq \pi^* \leq 1$) represents a generalization of this naïve strategy.

In the absence of any covariates, the null model includes only an intercept. In the Appendix, we show that $\hat{\mu}_i = \hat{\mu} = (n_1 + \pi^* n_e)/n$ and, for external individuals, $\hat{Y}_i^* = \pi^*$. The score statistic can then be expressed as

$$S_{\pi^*} = n_1(1 - \hat{\mu})\bar{G}_1 - n_0\hat{\mu}\bar{G}_0 + n_e(n_0 + n_1)(\pi^* - P_{CC})\bar{G}_e$$

where $\bar{G}_0, \bar{G}_1$, and $\bar{G}_e$ are the mean dosages in the internal controls, the internal cases, and the external sample, respectively, and $P_{CC} = n_1/(n_0 + n_1)$ is the proportion of cases in the internal case-control sample. Note that when $\pi^* = \pi_{CC}$ the score statistic depends only on the internal sample. The score statistic can be represented equivalently as $S_{\pi^*} = S_0 + \pi^* S_{E|CC}$, where $S_{E|CC} = n_e(1 - P_e)\bar{G}_e - P_e(n_0\bar{G}_0 + n_1\bar{G}_1)$ and $P_e = n_e/N$. Note that $S_{E|CC}$ is the score statistic obtained for a logistic-regression model that treats external individuals as cases and internal individuals as controls. Thus, the score statistic S_{π^*} is a linear combination of two score statistics; one that compares internal cases to the combined group of internal controls and external individuals (S_0) and one that compares the external sample to the internal sample ($S_{E|CC}$).

The model-based estimator of the variance of the score statistic ($\hat{V}_{\pi^*}$) is provided in the Appendix. The score test-statistic is given by $T_{\pi^*} = S_{\pi^*}^2/\hat{V}_{\pi^*}$. For $\pi^* = 0$, the test-statistic is well defined when $n_1 > 0$ *and* at least one of n_0, n_e are non-zero. In contrast, for $\pi^* > 0$, the test-statistic is well defined if *any* two of n_0, n_1, and n_e are non-zero; indeed, when $\pi^* > 0$, we can in principle build a valid test by comparing internal controls to the external sample in the absence of internal cases, although this approach will likely have low power. Finally, we note that when $n_e = 0$ or $\pi^* = 0$, the resulting test is simply the usual trend-test for binary data.

Asymptotic non-centrality parameter and relative efficiency

From standard likelihood-theory, T_{π^*} asymptotically follows a chi-squared distribution with one degree of freedom. In the Appendix we show that the expectation of the score statistic $\mathbb{E}S_{\pi^*} = nC_{\pi,\pi^*}\mathbb{E}(\bar{G}_1 - \bar{G}_0)$ where $C_{\pi,\pi^*} = (1-\pi^*)(1-\pi)P_1(1-P_1) + \pi\pi^* P_0(1-P_0) + (\pi + \pi^* - 2\pi\pi^*)P_0P_1$ and $P_i = n_i/n$ for $i = 1,2$. Under the null hypothesis, $\mathbb{E}S_{\pi^*} = 0$ regardless of the value of π^*, since $\mathbb{E}(\bar{G}_1 - \bar{G}_0) = 0$. Further, we show that under the null and local alternatives (that is, $\beta = O(\sqrt{n^{-1}})$), the estimated variance $\hat{V}_{\pi^*}$ converges to $V_{\pi^*} = nK_{\pi^*}V_g$ where V_G is the variance of G (in the population) and $K_{\pi^*} = (1-\pi^*)^2 P_1(1-P_1) + {\pi^*}^2 P_0(1-P_0) + 2\pi^*(1-\pi^*)P_0P_1$. Straightforward algebra shows that V_{π^*} is the exact (finite sample) variance of the score statistic under the null.

The asymptotic non-centrality parameter is $\lambda_{\pi^*} = \mathbb{E}^2 S_{\pi^*}/V_{\pi^*} = [C^2_{\pi,\pi^*}/K_{\pi^*}]\Delta^2$ where $\Delta^2 = \lim_{n\to\infty} n\,\mathbb{E}(\bar{G}_1 - \bar{G}_0)/V_g$. Under the null, $\lambda_{\pi^*} = 0$ for $0 \le \pi^* \le 1$; that is, the score test is valid (controls Type 1 error) regardless of the specific value π^* used for the analysis. Further, it is easy to verify that, for a fixed design that includes external individuals (that is, fixed P_0, P_1, P_e with $P_e > 0$) with fixed effect size, the non-centrality parameter is maximized when $\pi^* = \pi$, that is, when the presumptive external sample case proportion is equal to the true external sample case proportion. Thus, for a fixed design, the test based on the test-statistic T_{π} is most powerful amongst the class of tests based on T_{π^*} with $0 \leq \square^* \leq 1$.

In this paper, we compare two tests by assessing the relative efficiency of one with respect to the other. The relative efficiency of test T_x with respect to a reference test T_y is defined as λ_x/λ_y, the ratio of their non-centrality parameters. In contrast to a comparison of power, the (asymptotic) relative efficiency does not depend on either the effect size β or the variance of the genetic variant V_G. The relative efficiency can be interpreted as a ratio of (effective) sample sizes. A relative efficiency greater (less than) than one indicates that T_x is more (less) powerful than T_y.

RESULTS

Asymptotic bias

Denote by β_{π,π^*} the value to which the MLE asymptotically converges when assuming π^* is the external sample case proportion and the true external sample case proportion is π. We say that the MLE is asymptotically unbiased when $\beta_{\pi,\pi^*} = \beta$, where β is the true single-variant effect size. From asymptotic likelihood-theory, we know that $\beta_{\pi,\pi}$ is consistent (since the MLE for a correctly specified model is consistent under the usual regularity conditions). Using the results in the Methods and the fact that the Wald and score tests are asymptotically equivalent under the null and under local alternatives, in the Appendix we show that $\beta_{\pi,\pi^*}/\beta \approx b_{\pi,\pi^*}$ where $b_{\pi,\pi^*} = (C_{\pi,\pi^*}K_{\pi})/(C_{\pi,\pi}K_{\pi^*})$. When $\beta \neq 0$, b_{π,π^*} is a measure of the relative asymptotic bias; values greater than one indicate bias away from the null and values less than one indicate bias towards the null. Note that when $\beta = 0$, β_{π,π^*} is also equal to 0 for all π and π^*; that is, under the null, misspecification of π^* does not result in bias.

Figure 4-1 shows how the relative bias b_{π,π^*} varies with π^* for various values of π for a design with equal number of internal cases and internal controls ($P_{CC} = 0.5$) and equal external

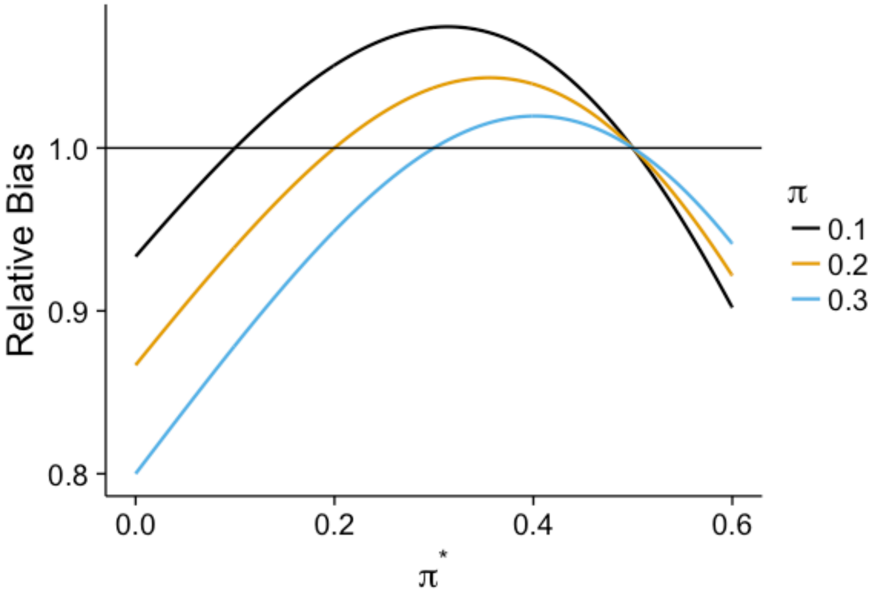

Figure 4-1 Dependence of relative bias b_{π,π^*} (Y-axis) on the presumptive external sample case proportion (π^*, X-axis) used for analysis for three different values of true external sample case proportion (π). Black, orange, and blue curves depict π values of 0.1, 0.2, and 0.3 respectively. The values in this plot were calculated for a study in which the external sample comprises half of the total sample ($P_e = 0.5$) and the number of internal cases is equal to the number of internal controls ($P_1 = P_0$). The relative bias is defined as $b_{\pi,\pi^*} = \beta_{\pi^*}/\beta$ where β is the true effect size and β_{π^*} is (asymptotically) the expectation of the effect size estimator for analysis with presumptive external sample case proportion set to π^*.

and internal sample sizes ($P_e = 0.5$). Note that all displayed values of π are less than P_{CC}, a situation that we believe will be typical in practice. The figure illustrates some general properties. First, $b_{\pi,\pi^*} < 1$ when $\pi^* < \pi$ or $\pi^* > P_{CC}$; that is, underspecifying π^* or setting π^* to be greater than internal sample case proportion results in under estimation of β (on average). Second, $b_{\pi,\pi^*} > 1$ when $\pi < \pi^* < P_{CC}$; that is, setting π^* to any value between the external and internal sample case proportions results in overestimation of β (on average). When $\pi^* = 0$ (analysis assuming external individuals are true controls), the under-estimation becomes worse (that is, $b_{\pi,0}$ decreases) as π increases.

Bias increases (b_{π,π^*} moves away from one) as the proportion of external individuals increases. When $\pi^* = 0$, we have $\lim_{P_e \to 1} b_{\pi,0} = 1 - \pi$. Thus, for analysis with $\pi^* = 0$, we have $1 - \pi < b_{\pi,0} \leq 1$ for all designs, so that for analysis treating external individuals as controls, the underestimation can be no worse than π%.

Conditions under which including the external sample increases power

Although treating the external individuals as controls results in bias towards the null when $\pi > 0$, the increase in sample size due to including the external sample may still result in increased power for association tests (compared to analyzing internal samples only) if the external sample case proportion is small. Figure 4-2 shows the relative efficiency of tests applied to designs obtained by adding varying number of external individuals to $n_0 = 5000$ internal controls and $n_1 = 5000$ internal cases; the reference design contains no external individuals ($n_e = 0$). We consider analysis with $\pi^* = 0$ (solid lines) and $\pi^* = \pi$ (dashed lines) when true external sample case proportions (π) are 0, 0.1, 0.2, and 0.3. Figure 4-2 shows that, when π is sufficiently low, adding external individuals increases efficiency even when external individuals are incorrectly assumed to be controls, although efficiency relative to a design with equal n_e decreases with increasing π. However, if π is large, treating external individuals as controls may actually lead to decreased efficiency relative to analysis that avoids the external sample entirely. In contrast, analysis with $\pi^* = \pi$ yields increased power even when π is large.

To provide useful guidance for how accurately π^* must be specified so that including the external sample increases efficiency and power, we focus on situations where $\pi \le P_{CC}$, the situation that we expect to be typical in practice. In the Appendix, we derive the exact range of π^* values that result in increased power for a given design (n_0, n_1, n_e) relative to the internal-only design $(n_0, n_1, 0)$. While the range is a complex function of n_0, n_1, n_e and π, it is uniformly true that including external controls increases power whenever $\max(2\pi - P_{CC}, 0) < \pi^* < P_{CC}$; in other words, π^* needs to be accurately specified to within a (one-sided) margin of length $P_{CC} - \pi$ from the true value π for increased power with addition of the external sample. In particular, treating external individuals as controls yields increased power compared to not

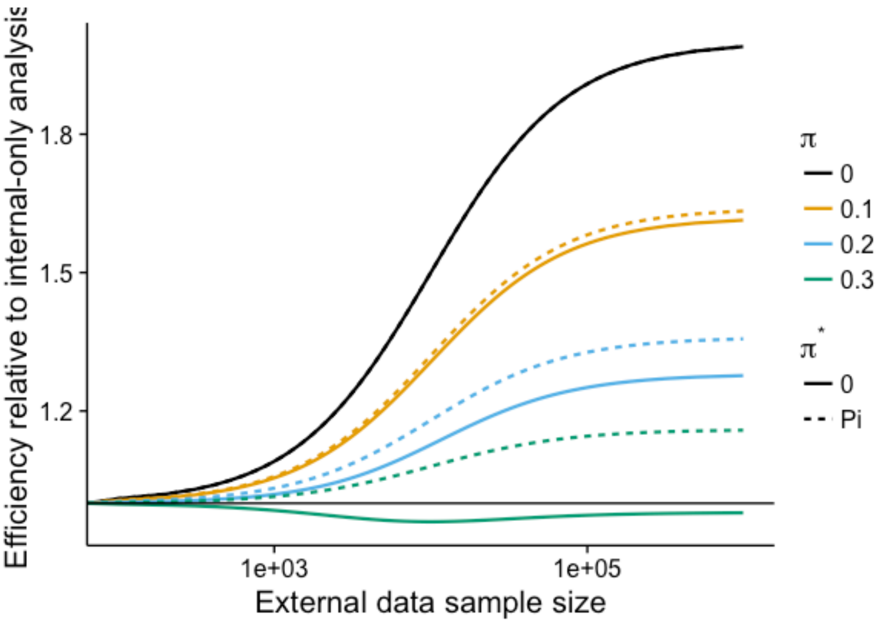

Figure 4-2 Relative efficiency (Y-axis) of analysis including the external sample compared to internal-only analysis as a function of external sample size (X-axis, log

scale). Internal sample consist of 5000 cases and 5000 controls. Colors depict different external sample case proportions (π). Dashed lines show relative efficiency for analysis with $\pi^* = \pi$. Solid lines represent analysis assuming external individuals are true controls ($\pi^* = 0$). Black line depicts relative efficiency of 1.

including the external sample when $\pi < P_{CC}/2$; that is, the internal case proportion should be at least twice the external case proportion.

Note that, the range $\max(2\pi - P_{CC}, 0) < \pi^* < P_{CC}$ includes π when $\pi < P_{CC}$. Thus, the test based on T_π (setting $\pi^* = \pi$) results in increased power with the external sample for all values $\pi < P_{CC}$. However, as noted above, when $\pi^* = P_{CC}$ the test-statistic T_π does not depend on the external sample; thus, when $\pi = P_{CC}$ the optimal testing procedure discards the external sample entirely. Here, any test T_{π^*} with $\pi^* \neq P_{CC}$ will result in decreased power upon adding external controls. Indeed, for a given external sample case proportion, the range of π^* values for which including the external sample increases power becomes smaller as π gets closer to P_{CC} and the set is empty when $\pi = P_{CC}$.

Power gain due to accurate specification of π^*

As noted earlier, analysis with $\pi^* = \pi$ yields the most powerful test amongst all possible π^* values. Figure 4-2 shows that the power gained due to accurate specification of π^* (relative to treating external individuals as controls, $\pi^* = 0$) is modest for small π but can be more appreciable when π approaches or exceeds $P_{CC}/2$. Figure 4-3 further illustrates this pattern as P_{CC} varies for fixed $P_e = 0.5$ and a fixed external sample case proportion $\pi = 0.25$ that we expect to be at the upper end of possible values in typical circumstances. Here, correctly specifying $\pi^* = 0.25$ results in an efficiency gain of 8%, 4% and 1% when the internal sample case proportions are 0.5, 0.75 and 0.95 respectively. Thus, when internal sample contains a large proportion of cases, treating external individuals as controls is not substantially less powerful than the optimal testing procedure. Indeed, it is easy to verify that when the internal sample consists of only cases ($P_{CC} = 1$), the non-centrality parameter λ_{π^*} does not vary with π^*; that is, when $P_{CC} = 1$, tests based on different π^* values all have the same power. However, estimation bias does depend on π^*. In the Appendix we show that the (conservative) range of π^* values that yield tests more powerful than tests setting $\pi^* = 0$ is $0 < \pi^* < 2\pi/(1+\pi)$. Thus, as π becomes smaller, π^* needs to be specified more accurately for T_{π^*} to be more powerful than the naïve test T_0.

DISCUSSION

We provide an analytical evaluation of a general testing strategy that incorporates knowledge about the proportion of cases in an external sample with missing case-control status. The method is a generalization of the naïve strategy that treats all external individuals as controls. We show that inaccurate specification of the external sample case proportion leads to biased estimation of effect sizes (except under the null). We provide an approximation for the

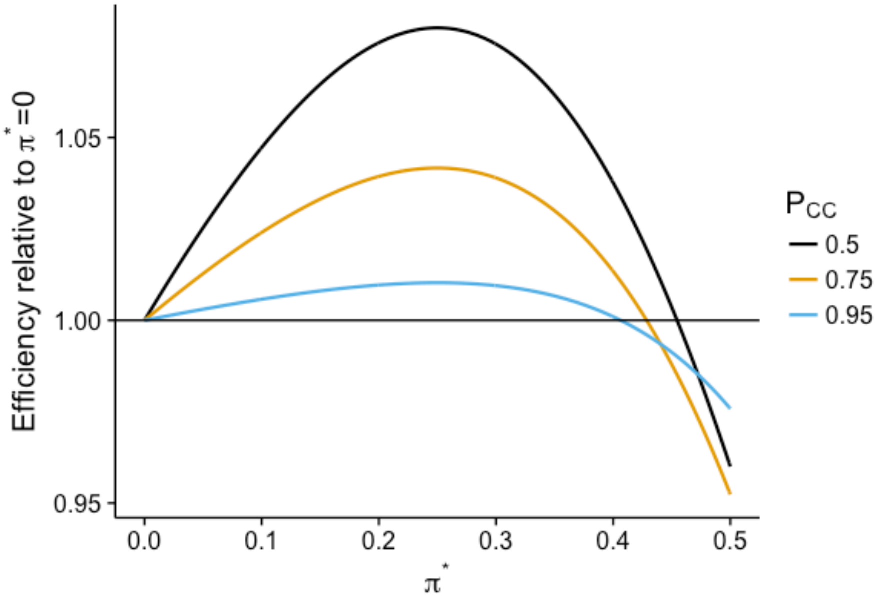

Figure 4-3 Relative efficiency (Y-axis) of tests with varying π^* (X-axis) compared to the test that treats external individuals as controls ($\pi^* = 0$) for designs with equal internal and external sample sizes ($P_e = 0.5$). The true external sample case proportion is $\pi = 0.25$. Colors depict different internal sample case proportions ($P_{CC} = n_1/(n_1 + n_0)$).

bias and show that treating external individuals as controls leads to underestimation of the effect size.

Inaccurate specification of the external sample case proportion does not increase Type 1 error, but may lead to decreased power relative to analysis that avoids the external sample. We show that the accuracy with which the external sample case proportion needs to be specified to prevent power loss increases as the true proportion approaches the internal sample case proportion. In particular, for analyses that treat external individuals as controls, the internal case proportion needs to be larger than twice the external case proportion to prevent loss of power upon including external individuals. We note this condition is likely to be satisfied in practice since most case-control studies tend to be close to balanced. When this condition is not satisfied, the external sample can still be included (using the proposed method to model π) to increase power if the external sample case proportion is known to within a range of $P_{CC} - \pi$ around either side of π. Our results show that, when the internal sample is balanced between cases and controls, treating external individuals as controls is a powerful strategy for including external individuals with unknown case-control status even when external sample case proportions are as large as 20%.

Quantification of loss of efficiency and bias due to misclassification of case-control status across all individuals has been addressed in both the statistics (Neuhaus, 1999) and genetic epidemiology (Edwards et al., 2005) literature. We extend these results to the situation where only a subset of observations are missing/misclassified for case-control status. Note that our likelihood expression for external individuals matches that provided in Neuhaus (1999) under a prospective design. In Neuhaus (1999), γ_0 is the probability that a control is incorrectly classified as a case and γ_1 is the probability that a case is incorrectly classified as a control. Neuhaus (1999) assumed that the constants γ_0 and γ_1 are known in a prospective study setting; we clarify how they relate to π, n_0, n_1 and n_e in a retrospective study with external controls. Further, we note that the likelihood presented here can be easily modified to incorporate misclassification in the internal sample by considering internal cases and internal controls as mixtures.

The proposed method is based on and is an extension of other methods that deal with missing case-control status in a subset of samples. Lancaster and Imbens (1994) propose a method-of-moments estimation procedure for studies with known/validated cases and individuals with unknown case-control status (which they call contaminated controls). Ward et al. (2009) present an EM algorithm to estimate parameters in the same setting (a design that is called “presence-only design” in ecology). We extend the model proposed by them to include known (internal) controls. Availability of the information matrix reveals that the proposed likelihood can be maximized by the Newton-Raphson or iteratively re-weighted least-squares algorithm. Further, we derive the score test that is computationally more efficient than the Wald or likelihood ratio tests in a GWAS setting. Thornton & McPeek (2010) present a score test based on estimating equations that is primarily designed to deal with missing genotypes in related individuals but also accounts for missing case-control status in a subset of individuals by incorporating a presumptive case proportion (or prevalence) parameter; their method cannot incorporate covariates. Our proposed score test is equivalent to their test statistic in the absence of covariates, missing genotypes and relatedness.

We reiterate that our model and power/Type 1 error calculations assume that, apart from missing case-control status, no other systematic differences exist between internal and external samples. This assumption is unlikely to hold in practice given expected differences in ancestry, sample recruitment (selection bias), biological sample collection and assay techniques (differential measurement error) between internal and external samples. The SiGN study (NINDS Stroke Genetics Network and International Stroke Genetics Consortium, 2016) with external controls performed multiple rounds of variant quality control and genomic control corrections before genomic inflation factors reduced to acceptable levels. Guo et al. (2018) recommend multiple rounds of stringent quality control based on sequencing read depths, sequencing quality scores, variant pathogenicity filters and ancestry for burden tests with external controls. Hendricks et al. (2018) present a burden testing method that uses burdens of benign (presumed null) variants to control for test statistic inflation. Lee et al. (2017) propose a method (iECAT) that controls for systematic differences between internal and external samples by down-weighting external samples according the allele-frequency differences between internal and external controls. We note that appropriately controlling for systematic differences between internal and external samples will likely reduce power; thus, our power calculations are therefore likely optimistic.

In summary, we present a method to account for known case proportion in the external sample and present associated power calculations. We show that, in the absence of other challenges and for internal designs with a suitably large proportion of cases, making use of external data with missing case-control status can be a useful strategy even when external individuals are incorrectly treated as controls.

APPENDIX

Derivation of the prospective likelihood

Let Y be a binary random variable identifying cases ($Y = 1$) and controls ($Y = 0$) in a population of interest. Let $\mathbf{x}$ denote a vector of predictors and $f(\mathbf{x})$ denote the probability density function (PDF) for $\mathbf{x}$ in this population. We assume that $P(Y = 1|\mathbf{x}) = e^{\theta(\mathbf{x})} / (1 + e^{\theta(\mathbf{x})})$ where $\theta(\mathbf{x}) = \alpha + \mathbf{x}'\boldsymbol{\beta}$. Let $P(Y = 1) = \Pi$; that is, Π is the prevalence of the disease in the population of interest. Denote by $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ the conditional PDFs of $\mathbf{x}$ in controls and cases, respectively. Then, by Bayes Theorem, $f_0(\mathbf{x}) = \frac{f(\mathbf{x})}{(1-\Pi)(1+e^{\theta(\mathbf{x})})}$ and $f_1(\mathbf{x}) = \frac{f(\mathbf{x})e^{\theta(\mathbf{x})}}{\Pi(1+e^{\theta(\mathbf{x})})}$.

We assume that the internal sample is constructed by recruiting n_0 individuals (internal controls) by simple random sampling from the subset of controls in the population and n_1 individuals (internal cases) by simple random sampling from the subset of cases. We assume that each individual in the external sample is, with probability π, a randomly sampled case or, with probability $1-\pi$, a randomly sampled control. Case-control status is not recorded for any individual in the external sample. Then, $f_e(\mathbf{x}) = (1-\pi)f_0(\mathbf{x}) + \pi f_1(\mathbf{x})$ is the PDF of $\mathbf{x}$ for an individual in the external sample, where $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ are the PDFs of $\mathbf{x}$ for internal controls and internal cases, respectively. The study sample is the collection of n_0 internal controls, n_1 internal cases and n_e external individuals with total sample size $n = n_0 + n_1 + n_e$.

Let Z be a categorical variable taking values “internal control”, “internal case”, and “external”. Define $P_0 = n_0/n$, $P_1 = n_1/n$, and $P_e = n_e/n$. Consider an alternative sampling scheme in which a total of n independent observations $(Z, \mathbf{x})$ are obtained in the following manner: Z is recorded as “internal control” and $\mathbf{x}$ is sampled from $f_0(\mathbf{x})$ with probability P_0, or Z is recorded as “internal case” and $\mathbf{x}$ is sampled from $f_1(\mathbf{x})$ with probability P_1, or Z is recorded as “external” and $\mathbf{x}$ is sampled from $f_e(\mathbf{x})$ with probability P_e. Define $D(\mathbf{x}) = P_0 f_0(\mathbf{x}) + P_1 f_1(\mathbf{x}) + P_e f_e(\mathbf{x})$. For this sampling scheme, $P(Z = \text{"internal control"}|\mathbf{x}) = f_0(\mathbf{x})/D(\mathbf{x})$, $P(Z = \text{"internal case"}|\mathbf{x}) = f_1(\mathbf{x})/D(\mathbf{x})$ and $P(Z = \text{"external"}|\mathbf{x}) = f_e(\mathbf{x})/D(\mathbf{x})$. Straightforward algebra yields the expressions presented in the Model section of the main text.

The function $L(\alpha_D, \boldsymbol{\beta}) = \prod_{i=1}^{n} P(Z_i|\mathbf{x}_i)$ is called the prospective likelihood. It is well known that maximizing the prospective likelihood yields a consistent estimator ($\widehat{\boldsymbol{\beta}}$) for $\boldsymbol{\beta}$ and the usual variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ is asymptotically correct (Prentice & Pyke, 1979; Scott & Wild, 1997). However, we note that since we treat π as a known parameter, consistent estimation is dependent on correctly specifying π. That is, when π is incorrectly specified, $\widehat{\boldsymbol{\beta}}$ may not be asymptotically unbiased.

Let $\mathbf{X}$ be the design matrix (with the first column being the intercept) and $\Lambda' = (\alpha_D, \boldsymbol{\beta}')$ be the vector of parameters for the prospective likelihood. Straightforward calculations show that the score corresponding to the prospective likelihood is $\mathbf{S} = \frac{\partial \log L}{\partial \Lambda} = (\mathbf{Y}^* - \boldsymbol{\mu})'\mathbf{X}$ where $\mathbf{Y}^*$ and $\boldsymbol{\mu}$ are the vectors $(Y_1^*, \dots, Y_n^*)'$ and $(\mu_1, \dots, \mu_n)'$ with Y_i^* and μ_i as defined in the main text. Similar calculations show that the prospective likelihood Fisher-Information matrix is $i_\Lambda = \mathbf{X}'\mathbf{V}_\mu\mathbf{X}$ where $\mathbf{V}_\mu$ is a diagonal matrix with i, i^{th} entry $\mu_i(1-\mu_i)$ if the i^{th} individual is an internal case or internal control and $\mu_i(1-\mu_i) - Y_i^*(1-Y_i^*)$ if the i^{th} individual is external.

Score test for single variant analysis with no covariates

For a single variant model with no covariates, the null model contains only an intercept. Thus, under the null model, $\mu_i = \mu = e^{\alpha_D}/(1+e^{\alpha_D})$ for $i = 1, \dots, n$. Further, under the null we have $\alpha = \log\left(\frac{\Pi}{1-\Pi}\right)$ and, for external individuals $\eta(G_i) = \alpha + \log\left(\frac{\pi^*(1-\Pi)}{(1-\pi^*)\Pi}\right) = \log\left(\frac{\pi^*}{1-\pi^*}\right)$ so that, for external individuals, $Y_i^* = \pi^*$ under the null. The null MLE for μ is obtained by solving the score equation $\sum_i (Y_i^* - \mu) = 0$ which yields $\hat{\mu} = (n_1 + \pi^*\, n_e)/n$. The score statistic for single variant association is given by $S_{\pi^*} = \sum_i G_i(Y_i^* - \hat{\mu})$. Straightforward algebra yields the two expressions provided in the Methods section.

Define $p = \hat{\mu}(1-\hat{\mu})/K_{\pi^*}$, $\bar{G} = \sum_i^n G_i/n$, $\overline{G^2} = \sum_i^n G_i^2/n$ and $\overline{G_e^2} = \sum_{i \in external} G_i^2/n$. For the single variant model with no covariates, the estimated information matrix is $\hat{\imath}_\Lambda = nK_{\pi^*}\bar{\imath}$ where $\bar{\imath}$ is a 2×2 matrix with entries $\bar{\imath}_{1,1} = 1$, $\bar{\imath}_{1,2} = \bar{\imath}_{2,1} = p\bar{G} - (1-p)P_e\bar{G}_e$

and $\bar{\iota}_{2,2} = p\overline{G^2} - (1-p)P_e\overline{G_e^2}$. Define $\hat{V}_{\hat{\beta}} = \bar{\iota}_{2,2}^{-1}/(nK_{\pi^*})$. The usual estimator for the variance of the score statistic under the null is $\hat{V}_{\pi^*} = \hat{V}_{\hat{\beta}}^{-1}$.

Asymptotic variance of the score statistic and $\hat{\boldsymbol{\beta}}$

Under a model with no covariates, the genotypes for internal controls, internal cases, and external individuals are three independent, identically distributed samples from the distributions defined by the densities $f_0(G), f_1(G),$ and $f_e(G)$, respectively; under the null, $f_0 = f_1 = f_e$. To calculate the expectation of the score, we make use of the fact that $\mathbb{E}\bar{G}_e = (1-\pi)\mathbb{E}\bar{G}_0 + \pi\mathbb{E}\bar{G}_1$; straightforward algebra yields the expression for expectation provided in the Methods section.

Asymptotically, by the law of large numbers, the matrix $I = \hat{\iota}_{\Lambda}/n$ converges to the matrix $K_{\pi^*}\mathbb{E}\bar{\iota}$. Noting that $\mathbb{E}\overline{G_e^2} = (1-\pi)\mathbb{E}\overline{G_0^2} + \pi\mathbb{E}\overline{G_1^2}$, straightforward algebra yields $\mathbb{E}\bar{\iota}_{1,2} = (1-Q)\mathbb{E}\bar{G}_0 + Q\mathbb{E}\bar{G}_1$, and $\mathbb{E}\bar{\iota}_{2,2} = (1-Q)\mathbb{E}\overline{G_0^2} + Q\mathbb{E}\overline{G_1^2}$, where $Q = \{\hat{\mu}(1-\hat{\mu})(n_1 + \pi n_e)/n - \pi P_e \pi^*(1-\pi^*)\}$.

Let V_0 and V_1 denote the variances of G amongst internal controls and internal cases, respectively. Noting that $\mathbb{E}\overline{G_i^2} = V_i + \mathbb{E}^2\bar{G}_i$ for $i = 0,1$, it is straightforward to show that $(\mathbb{E}\bar{\iota})_{2,2}^{-1} = \{(1-Q)V_0 + QV_1 + Q(1-Q)\mathbb{E}^2(\bar{G}_1 - \bar{G}_0)\}^{-1}$. Under the null (and approximately, for local alternatives with small effect size), $V_0 = V_1 = V_G$ and $\bar{G}_0 = \bar{G}_1$, so that $(\mathbb{E}\bar{\iota})_{2,2}^{-1} = V_G^{-1}$. The asymptotic variance of $\hat{\beta}_{\pi,\pi^*}$ is, thus, $V_{\hat{\beta}_{\pi,\pi^*}} = (nK_{\pi^*}V_G)^{-1}$ and the asymptotic variance of the score statistic is $V_{\pi^*} = nK_{\pi^*}V_G$.

Asymptotic bias

We make use of the well-known result that the score and Wald tests are asymptotically equivalent under the null and local alternatives (Cox and Hinkley, 1974). Thus, for large sample sizes and small effect sizes, the non-centrality parameters for the two tests (applied to the same data) are approximately equal: $\lambda_{\pi^*}^{Wald} \approx \lambda_{\pi^*}$. Note that $\lambda_{\pi^*}^{Wald} = \beta_{\pi,\pi^*}^2 / V_{\widehat{\beta}_{\pi,\pi^*}}$. Further, we have $\lambda_{\pi^*}^{Wald} / \lambda_{\pi}^{Wald} \approx \lambda_{\pi^*} / \lambda_{\pi}$ which yields $\beta_{\pi,\pi^*} / \beta_{\pi,\pi} \approx \sqrt{\lambda_{\pi^*} V_{\pi^*} / \lambda_{\pi} V_{\pi}}$. Note that $\beta_{\pi,\pi} = \beta$, since the MLE is consistent when the model is correctly specified. Straightforward algebra yields the result presented in the asymptotic bias section of the main text.

Ranges for efficient π^* values

For a given reference non-centrality parameter value λ_{REF}, we wish to find all values π^* such that $\lambda_{\pi^*} > \lambda_{REF}$. Define $a = \{P_e[\pi(1-P_e) - P_1]\}^2 - \lambda_{REF} P_e(1-P_e)/\Delta^2$, $b = 2P_1P_e\{[\pi(1-P_e) - P_1](1 - P_1 - \pi P_e) + \lambda_{REF}/\Delta^2\}$ and $c = [P_1(1 - P_1 - \pi P_e)]^2 - P_1(1-P_1)\lambda_{REF}/\Delta^2$. The equation $\lambda_{\pi^*} = \lambda_{REF}$ implies the quadratic constraint $a\pi^* + b\pi + c = 0$. The two roots of this equation define the limits of range of efficient π^* values.

REFERENCES

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Cortes, A. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203-209.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Edwards, B. J., Haynes, C., Levenstien, M. A., Finch, S. J., & Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC genetics*, *6*(1), 18.

Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., & Lippincott, M. F. (2018). Burden testing of rare variants identified through exome sequencing via publicly available control data. *The American Journal of Human Genetics*, *103*(4), 522-534.

Hendricks, A. E., Billups, S. C., Pike, H. N., Farooqi, I. S., Zeggini, E., Santorico, S. A., ... & Dupuis, J. (2018). ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. *PLoS genetics*, *14*(10), e1007591.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... & Gauthier, L. D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443.

Lancaster, T., & Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, *71*(1-2), 145-160.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285-291.

N. S. G. Network, Pulit, S. L., McArdle, P. F., Wong, Q., Malik, R., Gwinn, K., ... & Arnett, D. K. (2016). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *The Lancet Neurology*, *15*(2), 174-184.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, *86*(4), 843-855.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403-411.

Scott, A. J., & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, *84*(1), 57-71.

Thornton, T., & McPeek, M. S. (2010). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics*, *86*(2), 172-184.

Ward, G., Hastie, T., Barry, S., Elith, J., & Leathwick, J. R. (2009). Presence‐only data and the EM algorithm. *Biometrics*, *65*(2), 554-563.

Chapter 5 Conclusion

This thesis characterized bias, precision and power for three statistical techniques used in GWAS. In Chapter 2, we compare adjusted-trait regression (ATR) to traditional techniques of analysis for multiple linear regression. We show, through analytical calculations, that ATR is biased towards the null when the genetic variants included in the association test are correlated with covariates. In this situation, ATR is less powerful than traditional tests that appropriately account for this correlation and the loss of power increases as stringency for controlling Type 1 error increases (equivalently, as the size of the test decreases). We show that, for single variant tests, the loss of power is completely determined by the coefficient of determination of the regression of variant onto covariates. For the multi-variant omnibus test, we show that the maximum possible loss of power is completely determined by the canonical correlation between the set of variants and covariates.

ATR is an ad-hoc methodology that is not justified by a formal statistical framework. Despite previous work that outlines its shortcomings (Demissie & Cupples, 2011; Xing et al., 2011; Che et al., 2012) through simulations and approximations, multiple studies in the last decade continue to use this method. We show, via exact power calculations, that at the Type 1 error thresholds commonly used in genome-wide analyses, even relatively small correlations between variants and covariates can lead to substantial power loss. Although the correlation between genetic variants and the most commonly included confounder (genetic ancestry) is expected to be low on average, correlations may be large for ancestry informative variants. These variants may be enriched for causal associations due to population specific selection. We hope our results encourage the statistical genetics community to use theoretically justified methods in favor of ad-hoc ones.

We note that our results for multi-variant (gene-based) tests other than the burden-test are limited to comparisons between the traditional omnibus test and its ATR analog. Many published papers rely on the SKAT framework (Lee et al., 2014) to perform gene-based testing. We were unable to derive closed-form results to compare SKAT and its ATR analog. I intend to carry out simulations to assess whether our results hold, at least approximately, for this comparison.

In Chapter 3, we assessed three commonly used imputation quality scores (allelic-RSQ, MACH-RSQ and INFO) as predictors of realized quality for low-frequency and rare variants. In addition, we assessed whether choice of imputation algorithm and reference panel size affected the relationship between imputation quality scores and realized quality. To achieve this, we performed genome-wide imputation on 8378 participants from the METSIM study using three different imputation algorithms (Beagle 4.2, IMPUTE 2 and minimac3) with the 1000 Genomes Phase 3 reference panel (1KG reference panel). We also imputed genotypes into the same set of individuals using the much larger Haplotype Reference Consortium (HRC) reference panel with minimac3.

We show that MACH-RSQ and INFO are identical when calculated on the same dataset. We observed that allelic-RSQ is a poorer predictor of realized quality compared to MACH-RSQ/INFO. We observed that average realized quality decreases as MAF decreases and the average difference between imputation quality scores and realized quality increases as MAF decreases; this necessitates utilizing relatively more stringent thresholds to classify low-frequency and rare variants based on realized quality as compared to thresholds appropriate for common variants. Since average realized quality for low-frequency and rare variants is substantially higher for HRC based imputation compared to 1KG based imputation, less stringent thresholds (e.g. MACH-RSQ/INFO = 0.3) for HRC based imputation yield similar classification operating characteristics to higher thresholds (e.g. MACH-RSQ/INFO=0.6) for 1KG based imputation.

Imputation quality scores are an important measure of how much information imputed genotypes carry. Our work clarifies that two commonly used quality scores (MACH-RSQ and INFO) are, in fact, numerically equivalent when evaluated on the same dataset; the different thresholds used for classifying well/poorly imputed common variants for MACH-RSQ and INFO in the extant literature are, thus, attributable to differences in the imputation algorithms that use each score (Minimac for MACH-RSQ and IMPUTE for INFO). Further, our work clarifies that as realized imputation quality increases due to availability of larger reference panels, stringency of thresholds used to classify well/poorly imputed variants can be reduced due increased realized qualities for low-frequency and rare variants.

We noted that our study relies on variants genotyped on the Illumina Exomechip for gold-standard genotypes. It is likely that the distribution of MAF for variants included on the Exomechip differs from the distribution of MAF of all imputable variants (that is, variants present in the reference panel) due to the variant selection strategy used to include variants on the Exomechip. We are currently performing the same analysis presented in Chapter 3 on a subset of approximately 3000 METSIM participants that have been sequenced at high depth. This analysis will offer a less biased view across the entire genome. Further, we will assess the calibration of genotype imputed posterior probabilities by performing posterior predictive checks to compare predicted and realized quality scores and MAFs.

In Chapter 4, we proposed a method to account for known/presumed proportion of cases in external samples that are missing case-control status. We presented analytical power calculations that acknowledge that the exact external sample case proportion is unknown in practice. Our power calculations include the important and easily implementable strategy that assumes the external sample contains only controls. We provided clear guidelines about when inclusion of a mixed sample is beneficial in terms of power. Specifically, for analyses assuming external participants are controls, we showed that including external samples increases power (relative to avoiding them) when the internal sample case proportion (or proportion of known cases to total known case-control samples) is greater than twice the external sample case proportion. We showed that this condition can be weakened (that is, the internal sample case proportion can be closer to the external sample case proportion) based on how accurately the external sample case proportion is known. We showed that the strategy of treating external participants as controls is close to the best possible analysis (under the proposed framework) in a wide variety of situations.

We note two important limitations of the proposed method. First, the method does not accommodate related individuals. This could be achieved by extending the

CERAMIC framework (Zhong et al., 2016) that implements a score test based on estimating equations. The estimating equations can be modified to reflect the constraints introduced into the score equations derived from the proposed model. CERAMIC properly accounts for relatedness and, additionally, improves power by incorporating information about affected relatives on whom genotypes are missing. Second, I propose to make the method suitable for rare variants by implementing calculation of p-values based on the saddle-point approximation (Dey et al., 2017). P-values for the logistic regression score test are known to be poorly calibrated for variants with low minor allele counts under designs with a large imbalance between cases and controls (Ma et al., 2013). The saddle-point approximation ameliorates this deficiency by utilizing a second-order approximation to the distribution of the score test.

Our work shows that missing case-control status/misclassification in the external sample is not a major concern (in terms of decreased efficiency) when the proportion of cases in the internal sample is low. Thus, we advocate that future development of methods and designs utilizing external data focus on addressing challenges (like differential genotyping error, selection bias etc.) that lead to increased Type 1 error.

In summary, this thesis discusses bias, precision and power for three statistical methods employed in GWAS. As datasets become even larger and complex, understanding (and controlling) the sources and mechanisms of bias will become crucial. Indeed, as sample sizes increase, biases of fixed magnitude exert larger negative effects relative to sampling variability (Meng, 2018). I look forward to this and other challenges!

REFERENCES

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics*, *101*(1), 37-49.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, *95*(1), 5-23.

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., & GoT2D Investigators. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*, *37*(6), 539-550.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685-726.

Zhong, S., Jiang, D., & McPeek, M. S. (2016). CERAMIC: Case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS genetics*, *12*(10), e100632