Title
Population genetic simulation study of power in association testing across genetic architectures and study designs

Permalink
https://escholarship.org/uc/item/07v2f9x3

Author
Tong, Dominic Ming Hay

Publication Date
2019

Peer reviewed|Thesis/dissertation
Population genetic simulation study of power in association testing across genetic architectures and study designs

by
Dominic Tong

DISSEPTION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in
Bioengineering

in the
GRADUATE DIVISION

of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:
Ryan Hernandez
Chair

Ian Holmes

Jeffrey D Wall

Committee Members
To Canada, my home and native land.
ACKNOWLEDGMENTS

To Ryan Hernandez, my Ph.D. advisor: Thank you for taking a fourth rotation student on at the last minute and providing an academic home these past four years. I have learned much about population genetics, statistical genetics, wrangling with code and computational clusters. More importantly, I have learned how to be a scientist, a thinker, and a better citizen of the world. Thank you for showing me how to be a leader, a mentor, and a friend, all at the same time.

To the Hernandez lab at UCSF, and then McGill for a week: Thank you for the welcoming environment that we created all these years together. To Dr. Raul Torres and Dr. Nicolas Strauli, thank you for the guidance and support as the elder graduate students. To Kevin Hartman and Melissa Spear, also soon-to-be PhD’s, thank you for being friends and colleagues with whom much could be shared over a meal or drinks. We have shared a lab for my past four years and I will be eternally grateful for those memories.

To the Bioengineering Association of Students (BEAST): Thank you for the second home and the opportunity to lead and guide the next generation of Bioengineering PhD students. Service is one of my highest values, and leading Retreat Committee 2017 and being Co-President in 2018 have been my honour. I hope I have improved the life of at least one other student in our program through my tenure.

To my friends in the program: Cameron Nemeth, for being the greatest friend and roommate one could ask for over the past five years, and congratulations, at the time of writing, on graduating; Jennifer Hu, for steadfast support along the way; Katie Cabral, for all the Gloomhaven; Jinny Sun, for all the lunches; Kristen Cotner, for pushing me to be adventurous; Suraj Makhija, for those discussions about the past and the future; Aniket Tolpadi, for being a
hockey fan (there really aren’t that many of those in the Bay Area, it seems, also Go Canucks Go!); and Jasmine Hughes, for being a fantastic guide as a fellow Canadian and data scientist.

To the people that aided me academically along the way: Mr. Loyie, for the belief in my abilities and moving me up to grade 9; Mr. Edwin Lai, for showing me that service to alleviate the suffering of others is a high ideal to strive for; Mr. Robert Ward, for the cranky old appreciation of art and literature; Mr. D. Barber, for showing me the brilliance and beauty of physics; Dr. Andre Marziali, Dr. Jon Nakane, Bernhard Zender, and the UBC Engineering Physics program, for being the leaders and troubleshooters and advice-givers and pushing all of us Fizzers to be better students, engineers, and people; Dr. Peter Zandstra, formerly of the University of Toronto and now at my alma mater, for taking on an Engineering Physicist in his stem cell bioengineering laboratory and teaching me all the biology I know; and Professors Jeffrey D. Wall and Ian Holmes, for their support and mentorship as members of my thesis committee and as scientific colleagues.

Last, to my parents: Mom and Dad, thank you for being there every single step of the way and pushing me to becoming a better man. I love you.

We say Tuum est at my alma mater – it’s yours, it’s up to you. It truly has been.
ACKNOWLEDGMENT OF PREVIOUSLY PUBLISHED MATERIALS AND RESEARCH CONTRIBUTIONS

The text of Chapters 2, 3, and 4 are adapted from Tong and Hernandez, 2019. Ryan Hernandez supervised the research that forms the basis of all three chapters. All analyses were designed by a close collaboration between Dominic Tong and Ryan Hernandez. The figures were created by Dominic Tong. The manuscript was written by Dominic Tong and edited by Ryan Hernandez.

REFERENCES

Tong, DMH., Hernandez, RD. Population genetic simulation study of power in association testing across genetic architectures and study designs. bioRxiv 632786; doi: https://doi.org/10.1101/632786
ABSTRACT

Population genetic simulation study of power in association testing across genetic architectures and study designs

Dominic Tong

The role of rare variants in complex disease is hotly debated, but the design of genetic association studies to statistically associate rare variants is not well understood. Here, we simulate rare variant association studies across different case/control panel sampling strategies, sequencing methods, and genetic architecture models based on evolutionary forces to determine the statistical performance of RVATs widely in use. We find that the highest statistical power of RVATs is achieved by sampling case/control individuals from the extremes of an underlying quantitative trait distribution. We also demonstrate that the use of genotyping arrays, in conjunction with imputation from a whole genome sequenced (WGS) reference panel, recovers the vast majority (90%) of the power that could be achieved by sequencing the case/control panel using current tools. Finally, we show that the statistical performance of RVATs decreases as rare variants become more important in the trait architecture. Our work shows that RVATs are not yet well-powered enough to make generalizable conclusions about the role of rare variants in complex trait architectures.
# TABLE OF CONTENTS

Chapter 1: Introduction ......................................................................................................................... 1  REFERENCEs............................................................................................................................................... 6

Chapter 2: Designing a simulation pipeline to evaluate the statistical power of genetic association tests under different genetic architectures and study designs ......................................................................................... 14
  Overview ................................................................................................................................................... 15
  Simulating genomic sequence data ........................................................................................................... 15
  Simulating genotype data ......................................................................................................................... 16
  Simulating quantitative phenotypes ......................................................................................................... 17
  Selecting sampling strategies for association tests ................................................................................... 19
  Imputing genotyped data .......................................................................................................................... 20
  Running tests of association on simulated data .......................................................................................... 20
  Calculating cumulative genetic variance .................................................................................................. 21
  Performing Haseman-Elston regression to estimate heritability ................................................................. 22
  ReferenceS.................................................................................................................................................. 23

Chapter 3: The statistical power of genetic association tests under different genetic architectures .......................................................................................................................... 31
  Rare variants explain a majority of heritability only under restrictive scenarios ....................................... 32
  Haseman-Elston regression to identify observable heritability in simulation ............................................. 33
  Statistical power varies dramatically across different study designs, genetic architectures, and polygenicity, but not across RVATs ................................................................................................................. 34
  As rare variants explain more genetic variance of the trait, RVAT power decreases ................................... 37
Chapter 4: The statistical power of genetic association tests under different study designs

Using extreme cases/controls as sampling strategy improves statistical power of SKAT-O ... 41
RVATs perform nearly as well on imputed data as they do on sequence data ....................... 42
RVATs under a GWAS peak .................................................................................................. 43
Window of discovery around causal loci ............................................................................. 47

Chapter 5: Discussion ............................................................................................................. 49

REFERENCES .......................................................................................................................... 57

Appendix A: Supplemental Material to Chapter 3 ................................................................. 65
Appendix B: Supplemental Material to Chapter 4 ................................................................. 76
LIST OF FIGURES

Figure 2.1 The software pipeline built to study the statistical power of association testing in a simulation framework. .................................................................................................................................................. 15

Figure 3.1 The cumulative proportion of the genetic variance explained by variants under minor allele frequency x .................................................................................................................................................. 33

Figure 3.2 Haseman-Elston regression on simulated case/control data shows the inferred cumulative heritability explained by variants under a given MAF ........................................................................................................... 34

Figure 3.3 A global overview of the statistical power in 50,000 African individuals of a burden test (KBAC), a variance-component test (SKAT), and a combined test (SKAT-O) for all parameters .................................................................................................................................................. 36

Figure 3.4 A global overview of the statistical power in 50,000 European individuals of a burden test (KBAC), a variance-component test (SKAT), and a combined test (SKAT-O) for all parameters .................................................................................................................................................. 37

Figure 3.5 The statistical power of SKAT-O in African populations ................................................................................................................................. 39

Figure 4.1 Increase in SKAT-O power as a function of sample size in an African population. .................................................................................................................................................. 42

Figure 4.2 The mean power of SKAT-O across different genetic architectures using imputed data, compared to using sequence data .................................................................................................................................................. 43

Figure 4.3 The statistical power of GWAS given the results of SKAT-O in an African population .................................................................................................................................................. 45

Figure 4.4 SKAT-O power using sequencing data, given the results of SKAT-O applied to imputed data in a simulated African population .................................................................................................................................................. 46

Figure 4.5 The window of discovery around causal loci in an African population using SKAT-O .................................................................................................................................................. 48
Figure A.1 The statistical power of SKAT-O in European populations .......................... 65
Figure A.2 The statistical power of KBAC in African populations .............................. 66
Figure A.3 The statistical power of KBAC in European populations ....................... 67
Figure A.4 The statistical power of SKAT in African populations ............................ 68
Figure A.5 The statistical power of SKAT in European populations ....................... 69
Figure A.6 The statistical power of SKAT-O in African populations ....................... 70
Figure A.7 The statistical power of SKAT-O in European populations .................... 71
Figure A.8 The statistical power of KBAC in African populations ....................... 72
Figure A.9 The statistical power of KBAC in European populations .................... 73
Figure A.10 The statistical power of SKAT in African populations ....................... 74
Figure A.11 The statistical power of SKAT in European populations .................... 75
Figure B.1 Increase in SKAT-O power as a function of sample size in an European  
population. ....................................................................................................................... 76
Figure B.2 Increase in KBAC power as a function of sample size in an African (left) and 
European (right) population ......................................................................................... 76
Figure B.3 Increase in SKAT power as a function of sample size in an African (left) and 
European (right) population ......................................................................................... 77
Figure B.4 The mean power of SKAT-O in an European population across different genetic 
architectures using imputed data, compared to using sequence data .......................... 77
Figure B.5 The mean power of KBAC in an African (left) and European (right) population 
across different genetic architectures using imputed data, compared to using sequence data. ... 78
Figure B.6 The mean power of SKAT in an African (left) and European (right) population 
across different genetic architectures using imputed data, compared to using sequence data. ... 78
Figure B.7 The statistical power of GWAS given the results of SKAT-O in an European population ................................................................. 79
Figure B.8 The statistical power of GWAS given the results of KBAC in an African population .................................................................................. 80
Figure B.9 The statistical power of GWAS given the results of KBAC in an European population .................................................................................. 81
Figure B.10 The statistical power of GWAS given the results of SKAT in an African population .................................................................................. 82
Figure B.11 The statistical power of GWAS given the results of SKAT in an European population .................................................................................. 83
Figure B.12 SKAT-O power using sequencing data, given the results of SKAT-O applied to imputed data in an European population ........................................................................ 84
Figure B.13 KBAC power using sequencing data, given the results of KBAC applied to imputed data in an African population ........................................................................ 85
Figure B.14 KBAC power using sequencing data, given the results of KBAC applied to imputed data in an European population ........................................................................ 86
Figure B.15 SKAT power using sequencing data, given the results of SKAT applied to imputed data in an African population ........................................................................ 87
Figure B.16 SKAT power using sequencing data, given the results of SKAT applied to imputed data in an European population ........................................................................ 88
Figure B.17 The window of discovery around causal loci in an European population using SKAT-O ........................................................................ 89
Figure B.18 The window of discovery around causal loci in an African population using KBAC ................................................................. 90

Figure B.19 The window of discovery around causal loci in an European population using KBAC ............................................................... 91

Figure B.20 The window of discovery around causal loci in an African population using SKAT ................................................................. 92

Figure B.21 The window of discovery around causal loci in an European population using SKAT ................................................................. 93
LIST OF TABLES

Table 2.1 Genetic architectures examined in this study .............................................................. 19
Table 2.2 Study design parameters in this study.......................................................................... 20
Table 2.3 Genetic architecture parameters under the Uricchio model and the genetic variance explained by variants under MAF=1%. ................................................................. 21
Chapter 1: Introduction
The advent of massively-parallel high-throughput ‘-omics’ technologies, combined with the Moore’s Law increase in computational power, has enabled a revolution in the way discoveries in biology are made. In the past, Gregor Mendel bred peas in his monastery’s experimental garden to observe and formulate the laws of Mendelian inheritance; Charles Darwin collected and observed finches in the Galapagos Islands before eventually coming to the theory of evolution in London; Fisher, Wright, and Haldane all gathered observations by hand to make their immense contributions to the field of population and statistical genetics. In 2019, we are able to sequence hundreds of thousands of whole human genomes, able to make edits to genomes in vitro and observe their effects, and empirically demonstrate the theories from the grand trio of Fisher, Wright, and Haldane, while adding wrinkles of our own. It astounds me that I present statistical genetics research, while, as Newton said, standing on the shoulders of giants.

In this thesis, I focus on understanding the link between genetics and phenotype by thoroughly characterizing our existing statistical association methods within a simulation framework. It is clear that the use of genome sequencing and “big data” approaches has the potential to revolutionize our understanding of the genetic architecture underlying phenotypes and vastly improve the discovery of drugs and treatment of individuals within the healthcare system, termed “precision medicine”. This potential has yielded some fruit, in screening for breast cancer (Mavaddat et al., 2010), in warfarin dosing (Dean, 2012), in identifying a significant risk factor for cardiovascular disease (Cohen et al., 2006), but the statistical tests underlying these detected common variant associations are under-powered to detect rarer causal variants. To measure the impact of the genetic variants associated, we use the total heritability of the trait, a statistic used in the fields of breeding and genetics that estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals.
in that population. In many cases, this heritability is much greater than the heritability explained by the associated genetic variants using current statistical tests.

One proposed source of this so-called “missing heritability” are rare variants, which are hotly debated but have been implicated as a non-negligible source of genetic variance in prostate cancer (Mancuso et al., 2016), gene expression (Hernandez et al., 2017), height and BMI (Wainschtein et al., 2019). Unfortunately, power to detect rare variant associations is low in single-marker statistical tests at the genome-wide scale. Researchers have proposed many rare variant association tests (RVATs), statistical methods to pool rare variants within a putatively causal locus and test for association with the phenotype. These RVATs are broadly classified into three categories: burden tests (Liu & Leal, 2010), variance-component tests (Neale et al., 2011; Wu et al., 2011), and combined tests (Lee et al., 2012; Sun, Zheng, & Hsu, 2013). Though each test is published with its own validation simulations, these simulations are generally not comparable, and have their own flaws. (Moutsianas et al., 2015) systematically characterized the performance of commonly used gene-based rare variant association tests under a range of genetic architectures, sample sizes, variant effect sizes, and significance thresholds, and found that MiST, SKAT-O, and KBAC have the highest mean power across simulated data, but that these tests had overall low power even in the cases of loci with relatively large effect sizes.

It is well-known in the population genetics literature that population expansions and contractions (i.e. demography) can dramatically affect genome-wide patterns of genetic variation in a population (Auton et al., 2009; Bhaskar, Wang, & Song, 2015; Gravel et al., 2011; Uricchio, Zaitlen, Ye, Witte, & Hernandez, 2016), and that the action of natural selection can amplify or inhibit the frequency of functional alleles (Boyko et al., 2008; Adam Eyre-Walker, Woolfit, & Phelps, 2006; Lohmueller et al., 2011). Together, these evolutionary forces shape the genetic
architecture of complex traits (Lohmueller, 2014; Uricchio et al., 2016), and are critical components to understand in the pursuit of identifying the genetic basis for the bevy of human phenotypes understudy. Inferred demographic models of non-African human populations show a serial bottleneck model as populations migrated in waves across the globe, followed by explosive exponential growth since the dawn of agriculture. Moreover, studies of selection have found that most amino acid changes in proteins and changes in conserved non-coding loci are weakly deleterious (Boyko et al., 2008; Torgerson et al., 2009). Together, growth and selection has resulted in a preponderance of ultra-rare mutations (MAF<0.1%), which contribute a plurality of heritability in gene expression (Hernandez et al., 2017), BMI (Wainschtein et al., 2019), and possibly other traits. Accounting for demographic and selective effects on the frequency spectrum of causal variation is therefore crucial in characterizing the statistical power of RVATs. However, while previous evaluations of RVAT power have attempted to mimic the frequency spectrum of observed variants, they typically use phenotype models (or genetic architectures) that do not directly account for evolutionary forces like demography and natural selection and are often biologically unrealistic [e.g. effect sizes that are simple functions of the minor allele frequency (Wu et al., 2011)], limited to specific relative risks (Wray & Goddard, 2010), or lack pleiotropy (Moutsianas et al., 2015).

Another vital component of designing genetic association studies is the method of acquiring genetic data. Although the gold standard for capturing rare variation remains deep whole genome sequencing (WGS), the $1000 per genome cost still means performing WGS on any sizeable group of individuals remains prohibitively expensive for all but the largest consortia. Genotyping arrays make acquiring genetic data for a large number of individuals significantly less expensive, but lack coverage of rare variation. With larger WGS reference panels like the
Haplotype Reference Consortium (HRC; McCarthy et al., 2016), large numbers of genotyped samples can be imputed to gain some insight into rare variation. With such large reference panels, imputation coverage of genetic variation down to MAF≥0.1% is near perfect in European individuals (Quick et al., 2019). As more diverse reference panels become available [e.g. TOPMed (Taliun et al., 2019)], imputation in non-European and admixed populations will also improve, particularly for rare variants. Capturing these rare variants using genotyping arrays and imputation is more cost-effective and can lead to many more individuals in a study. However, imputation is limited by the variants that are carried by the individuals in the reference panel, and by the accuracy of the algorithm being used. Imputation accuracy falls off at lower minor allele frequencies (MAF), but the use of large WGS reference panels reduces the threshold of acceptable imputation quality ($r^2>0.3$) to ~0.004-0.006% (Taliun et al., 2019) in European and African populations. Despite these limitations, imputation has been used to identify rare variant associations in acute macular degeneration (Helgason et al., 2013), lipid levels in type 2 diabetes patients (Marvel et al., 2017), systemic lupus erythematosus (Martínez-Bueno & Alarcón-Riquelme, 2019), among others. It is possible that additional rare variant association signals can be found in imputed data as imputation quality improves, but it is unclear what the statistical properties of RVATs in this setting are.

In this thesis, I evaluate the statistical power of rare variant association tests in a simulation study under different genetic architectures, methods of acquiring genetic data, and methods of selecting individuals to be a part of the case-control cohort. I demonstrate how statistical power of RVATs is dependent on genetic architecture as well as sampling strategy for the case/control cohort. In particular, I find that sampling the extremes of a quantitative phenotype has the highest RVAT power, but counterintuitively, power erodes quickly for all sampling strategies as the amount of genetic variance explained by rare variants increases.
REFERENCES


https://doi.org/10.1038/nmeth.1785


https://doi.org/10.1534/genetics.106.057570


architecture of diseases and complex traits. *Nature Genetics.*

https://doi.org/10.1038/s41588-018-0148-2


https://doi.org/10.1534/g3.111.001198


https://doi.org/10.1371/journal.pgen.1004722


https://doi.org/10.1016/j.ajhg.2012.06.007


https://doi.org/10.1371/journal.pgen.1001156


Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLOS Genetics, 11*(4), e1005165. https://doi.org/10.1371/journal.pgen.1005165


https://doi.org/10.1101/548321


https://doi.org/10.1038/ng.2896


https://doi.org/10.1038/s41397-018-0042-4


Chapter 2: Designing a simulation pipeline to evaluate the statistical power of genetic association tests under different genetic architectures and study designs
Overview

To study the statistical power of genetic association tests under a variety of genetic architectures and study designs, we constructed a simulation pipeline from existing software and wrote code to fill in necessary gaps. This process mimics the process of running an empirical genetic association study: to gather genetic and phenotypic data, to impute, and to run association tests. The advantage of a simulation is that we have access to the ground truth and can objectively evaluate these statistical association tests across a variety of parameters. A brief overview of this process is shown in Figure 2.1, and the details follow below.

**Figure 2.1 The software pipeline built to study the statistical power of association testing in a simulation framework.** The branches are for evaluating the power of association test on imputation data and on whole genome sequencing (WGS) data.

**Simulating genomic sequence data**

To simulate genomic sequence data, we considered options from each of the classes of genetic simulators: haplotype resamplers, forward simulators, and backward or coalescent simulators. HAPGEN2 (Su, Marchini, & Donnelly, 2011), a haplotype resampler, was initially selected to match the method in (Moutsianas et al., 2015); however, our lab has previously shown that HAPGEN2 would simulate too few rare variants (Uricchio, Torres, Witte, & Hernandez, 2015) and efforts to scale up the mutation rate ran into memory and coding issues. I moved on to using SFS_CODE (R. D. Hernandez, 2008), a forward simulator that efficiently mimics the course of evolution given a human demographic model and evolutionary parameters like recombination rate,
and mutation rate. However, these simulations were not feasible on the scales required to investigate modern genetic association tests. Thankfully, Kelleher and colleagues developed an efficient coalescent simulator in 2016 (Kelleher, Etheridge, & McVean, 2016), where they simplified the storage and traversal of trees key to doing these simulations a hundredfold. This advance enabled me to simulate the genetic sequences required under a realistic demographic model and realistic evolutionary parameters to comprehensively evaluate modern statistical association tests.

We simulate neutral genetic sequence data under a coalescent model using msprime (Kelleher et al., 2016) with a European and African demographic history (Tennessen et al., 2012). Under this demographic model, the European population experienced a series of bottlenecks as they moved out of Africa and into Europe. These bottlenecks were followed by super exponential growth in the European population and recent exponential growth in the African population, along with bi-directional migration. Using this neutral demographic model, we generate a 5Mb region with a mutation rate of 1e-8 and with genetic map arbitrarily chosen to mimic chr22:17000000-22000000 in hg19.

**Simulating genotype data**

Some analyses are based on genotype array data. To simulate a genotyping array, we downsample the simulated neutral sequence data above to match the allele frequency spectrum and the average distance between variants of the Illumina OmniExpress2.5 genotyping chip, used in the GoT2D study (Fuchsberger et al., 2016). Though other genotyping arrays exist, we take this Illumina chip to be the standard, and this choice is unlikely to matter in simulated genetic data regardless. We do not simulate errors in the genotyping process here, so results from this
simulation will overestimate the overall effectiveness of genotyping then imputing as a method of collecting genetic data.

**Simulating quantitative phenotypes**

To simulate phenotypes needed for a genetic association study, we need a phenotype model. We considered several models and software packages to do so. Some models were overly simplistic, like the Wu et al, 2011 model which simply assumes a logarithmic relationship between effect size and minor allele frequency \( z = \log(x) \), where \( z \) is the effect size and \( x \) is the minor allele frequency. Other models relied on relative risks of a certain genetic variant [Wray and Goddard 2010], for which there is no real theoretical basis, and is hard to arbitrarily assign. The model Moutsianas et al, 2015 used is a step in the right direction, using the selection coefficient on a particular genetic variant as a basis for computing the effect size, but this model, based on (A. Eyre-Walker, 2010) lacks pleiotropy. We therefore use the Uricchio model, which computes an effect size for a genetic variant based on that variant’s selection coefficient, and accounting for pleiotropic effects.

We transform our simulated neutral genetic data into quantitative phenotypes using a three-step procedure, following Uricchio et al (Uricchio et al., 2016). First, we simulate functional variants using the forward simulator SFS_CODE (Hernandez, 2008) under the same demographic model as above, but with purifying selection. Specifically, we generate 2000 independent loci of length 100kbp (for a total of 200Mb) with 100,000 individuals, where new mutations receive a fitness effect drawn from a gamma distribution [as inferred for non-synonymous sites (Boyko et al., 2008)]. This procedure generates a large table of functional variants, with corresponding derived allele counts and fitness effects.
The second step is to project the allele frequencies of our list of functional variants down to the desired sample size (using a binomial model), and transform fitness effects to phenotypic effect sizes using the Uricchio et al. model (Uricchio et al., 2016). This model parameterized the correlation between fitness effects and phenotypic effect sizes (through $\rho$) and the functional relationship between fitness effects and phenotypic effect sizes (through $\tau$ and $\delta$). In particular, a causal variant with fitness effect $s$ will have effect size $z_s$ as follows:

$$z_s = \begin{cases} 
\delta s^\tau & \text{with probability } \rho \\
\delta s^\tau & \text{otherwise}
\end{cases}$$

Under this model, with probability $\rho$, the effect size $z_s$ of a site is a direct function of the site’s fitness effect ($s$), otherwise the effect size $z_s$ is a function of a randomly sampled fitness effect ($s_r$) drawn from the entire list of functional variants generated by the first step above. In this model, $\delta$ is +1 or -1 with equal probability to enable trait-increasing and trait-decreasing effects.

The third step for generating quantitative phenotypes is to identify the desired number of causal loci in our 5mb simulated sequence. For each variant within the causal loci, we sample a random variant from our list of functional variants generated in step two with the exact same allele frequency, and assign derived alleles at this causal site the effect size of the sampled functional variant. The quantitative phenotype of each individual ($Y_i$, for the $i$th individual) is then generated under an additive model by summing the effect sizes of all causal alleles that they carry:

$$Y_i = \sum_j X_{ij}z_j + \epsilon$$
Where \( z_j \) is the effect size of causal variant \( j \), \( X_{ij} \) is the number of causal alleles carried by individual \( i \) at site \( j \), and \( \epsilon \) is a Normal random variable with mean 0 and variance \( \sigma^2_{\text{environment}} \) (which ensures the desired level of heritability of the trait). See Table 1 for the specific values of \( \rho, \tau \), and heritability that are evaluated in this study.

### Table 2.1 Genetic architectures examined in this study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of causal loci</td>
<td>10, 100</td>
</tr>
<tr>
<td>Heritability (( h^2 ))</td>
<td>0.2, 0.8</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.5, 0.8, 0.9, 1.0</td>
</tr>
<tr>
<td>( \tau )</td>
<td>0.5, 1.0</td>
</tr>
</tbody>
</table>

### Selecting sampling strategies for association tests

The quantitative phenotypes can be dichotomized to simulate three different sampling strategies: random, 50/50, and extremes. In the extreme sampling strategy, we sample the desired number of individuals from the top and bottom of the quantitative phenotype distribution. For the random and 50/50 sampling strategy, we first define the individuals with quantitative phenotypes in the top \( P\% \) to be our population of cases (where \( P \) represents the prevalence of our trait of interest, set at 25\%), and the remaining individuals to be our population of controls. We then sample cases and controls from their respective populations. For the random sampling strategy, we sample cases in proportion to the prevalence of the trait, while for the 50/50 sampling strategy we sample equal numbers of cases and controls. The random sampling strategy is used as a worst-case scenario to establish the worst possible power under that sampling strategy.
Table 2.2 Study design parameters in this study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Random, 50/50, Extremes</td>
</tr>
<tr>
<td>Number of case/control individuals</td>
<td>5000, 10000</td>
</tr>
<tr>
<td>Number of reference panel individuals for imputing</td>
<td>10000, 20000</td>
</tr>
</tbody>
</table>

Imputing genotyped data

In some analyses, we evaluate the effectiveness of genotype imputation. Such analyses require two data sets: the phenotype sample (e.g. case/control or continuous phenotype), and an imputation reference panel. The dichotomized phenotype individuals are generated as above, with their genetic data down-sampled to mimic a genotype array platform. We then sample an additional set of individuals from the total population to form the imputation panel. The down-sampled genotype data is then pre-phased using SHAPEIT2 (Delaneau, Marchini, & Zagury, 2012) and imputed using IMPUTE4 (Bycroft et al., 2017). We considered other imputation algorithms, particularly BEAGLE (Browning, Zhou, & Browning, 2018) and minimac (Das et al., 2016), but these proved problematic with memory and software constraints. IMPUTE4, being a newly developed algorithm at the time, far outperformed BEAGLE and minimac, and so was selected for this study.

Running tests of association on simulated data

We ran rare variant association tests (RVATs) using the rvtests software (Zhan, Hu, Li, Abecasis, & Liu, 2016). We focus on SKAT (Wu et al., 2011), SKAT-O (Lee et al., 2012), and KBAC (Liu & Leal, 2010), which were found to be most powerful in detecting disease-associated variation in a previous study (Moutsianas et al., 2015). We applied each RVAT to non-overlapping analysis blocks of 10kbp across the simulated region, and computed power and false-positive rates...
for each test as the proportion of simulations with p-values below $2.5e$-$6$. We ran logistic regression on each variant above MAF=1% to determine associations with the phenotype using PLINK. The detection threshold was set at $5e$-$8$. To compare GWAS to RVAT power, we evaluate if there is a variant under the GWAS p-value threshold within the 10kb analysis block. If there is such a variant, we deem the GWAS to have found that analysis block to be causal for comparisons with RVAT.

**Calculating cumulative genetic variance**

We follow (Uricchio et al., 2016) in calculating $V_x$, the genetic variance due to variants at or below allele frequency $x$, which is given by:

$$V_x = 0.5 \int_{y=0}^{x} E(z^2 | y) f(y)(1 - y)(y) dy$$

Where $f(y)$ is the site frequency spectra (SFS), i.e. the proportion of sampled alleles at frequency $y$, and $E(z^2 | y)$ is the mean-squared effect size of variants at frequency $y$. We pool 20 simulations of 300kbp in 50k African individuals using msprime to obtain an accurate measure of the SFS and the expected effect size of variants at frequency $x$. To normalize across genetic architectures, we divide by $V_1$, which is the total additive genetic variance. The $V_{0.01}/V_1$ values (denoted as just $V'_{0.01}$ below) are used to denote the degree to which rare variants (variants with MAF ≤ 1%) matter under a particular pair of parameters under the Uricchio genetic architectures.

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\rho$</th>
<th>$V'_{0.01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.23191008598492735</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.35820017698651052</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.40029687398703817</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.44239357098756588</td>
</tr>
</tbody>
</table>
Performing Haseman-Elston regression to estimate heritability

We run Haseman-Elston regression to estimate heritability on our simulated traits. Haseman-Elston regression is a linear regression of the normalized covariance of the phenotype against the normalized covariance of the genotype (in the form of a genetic relatedness matrix (GRM)); the narrow-sense heritability of that phenotype is the slope of the regression line. We simulate the phenotype using the Uricchio model, above. We use PLINK v1.90beta (Chang et al., 2015) to calculate GRMs for different frequency bins in our samples with the frequency bin thresholds as follows: \([1.e-05, 2.e-05, 5.e-05, 1.e-04, 2.e-04, 5.e-04, 0.001, 0.01, 0.02, 0.1, 1]\). We then use GCTA v1.92.0 (Yang, Lee, Goddard, & Visscher, 2011) to run the Haseman-Elston regression.

<table>
<thead>
<tr>
<th>(\tau)</th>
<th>(\rho)</th>
<th>(V'_{0.01})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>0.50718348211637121</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.79549982205815639</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.89160526870541801</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.98771071535267974</td>
</tr>
</tbody>
</table>

PLINK: /netapp/home/dmctong/programs/plink/plink --vcf \${VCF} --memory \${N_MEM} --double-id --make-grm-bin --out \${OUT_PREFIX} --mac \${MIN_MAC} --max-mac \${MAX_MAC}

GCTA: /netapp/home/dmctong/programs/gcta_1.92.0beta3/gcta64 --HEreg --mgrm \${GCTA_MULTIGRM} --pheno \${GCTA_PHENO} --out \${PREF_OUT}
REFERENCES

Auton, A., Bryc, K., Boyko, A. R., Lohmueller, K. E., Novembre, J., Reynolds, A., … 


architecture of diseases and complex traits. *Nature Genetics*.

https://doi.org/10.1038/s41588-018-0148-2


https://doi.org/10.1534/g3.111.001198


https://doi.org/10.1371/journal.pgen.1004722


https://doi.org/10.1016/j.ajhg.2012.06.007


https://doi.org/10.1371/journal.pgen.1001156

26


Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLOS Genetics, 11*(4), e1005165. https://doi.org/10.1371/journal.pgen.1005165


Chapter 3: The statistical power of genetic association tests under different genetic architectures
**Rare variants explain a majority of heritability only under restrictive scenarios**

To determine whether there is genetic variance explained by rare variants, we calculated the expected genetic variance analytically under different ($\rho, \tau$) combinations of the Uricchio model being studied here (see Methods and Table 1). In Figure 1, we show that the proportion of genetic variance explained as a function of MAF. We focus on the genetic variance explained by variants with MAF < 1% ($V'_{0.01}$), which varies dramatically between 99% when $\rho=1$, $\tau=1$, to less than 1% when $\rho=0$, $\tau=0.5$. We note that when $\tau=1$ and $\rho \neq 0$, rare variants constitute a substantial fraction of the genetic variance ($V_{0.01} > 40\%$), and the majority of the rare variant contribution is explained by singletons in this simulated sample of 50,000 individuals. In contrast, when $\tau=0.5$ and $\rho \neq 0$, $V_{0.01}$ ranges from ~20%-60% but singletons are expected to make a more subtle contribution to the genetic architecture of the trait.
Figure 3.1 The cumulative proportion of the genetic variance explained by variants under minor allele frequency x (Vx/V1) for a sample of 5000 individuals drawn from an African population demographic model under different values of ρ and τ in the Uricchio model. Left: τ = 0.5; right: τ = 1. Dotted lines indicate the proportion of genetic variance explained by alleles under 1% MAF (referred to as V_{0.01}).

Haseman-Elston regression to identify observable heritability in simulation

Since the range of the phenotype model parameters (Table 2.1) are expected to produce a broad range of genetic architectures, we used heritability analyses (based on Haseman-Elston regression) to see if there was sufficient evidence in the data to differentiate the various models. In Figure 3.2, we show that in a simulated case/control panel of 10000 individuals across a 5Mb region, there is a difference in heritability explained by rare variants under the different genetic architectures, particularly when in the high heritability condition (h² = 0.8), and when the phenotypic effect sizes are directly correlated to the selection coefficient of that allele (ρ= 1). There is also an increase in heritability in the bin where MAF is between 0.5% and 5%, indicating there is heritability in causal variants expected to be captured by RVATs.

Possible reasons that these simulated plots do not match the analytical expectation shown in Figure 3.1 are: we do not simulate a sufficiently large genetic region, and hence do not generate sufficient numbers of ultra-rare variants to produce substantial rare variant heritability; or that stochasticity in the Haseman-Elston estimator is dominating the signal with 300 simulations; or another yet uncharacterized effect. However, even with only a 10,000 individual sample size, there can be a significant amount of heritability explained by rare variants under certain genetic architectures, and we can capture a lot of it using heritability estimation methods. The next step is to identify the causal loci driving the pattern, but this cannot not be done with such heritability analyses.
Figure 3.2 Haseman-Elston regression on simulated case/control data shows the inferred cumulative heritability explained by variants under a given MAF. True heritabilities in the columns (0.2 and 0.8); tau in the rows (0.5 and 1). 300 replicates of a 5Mb region with 10,000 samples simulated (5,000 cases and 5,000 controls), with 100 causal regions of size 10kb each.

Statistical power varies dramatically across different study designs, genetic architectures, and polygenicity, but not across RVATs

Figure 3.3 shows that rare variants can contribute substantial heritability to a trait under certain genetic architectures. Now we ask if we can detect the loci that harbor the causal rare variants using existing RVATs. To quantify the effects of genetic architecture and study design on the statistical power of RVATs, we focus on KBAC, SKAT, and SKAT-O, which represent each
of the three major categories of RVATs and have been shown to be among the most powerful (Moutsianas et al., 2015). For the 5Mb region we simulated (see Methods), we raster over parameters in genetic architecture (heritability, number of causal loci, and the relationship between selection and phenotypic effect sizes; Table 1) and in study design (sequencing vs genotype imputation and selection of individuals in the case/control vs extreme phenotype panels). In Figure 3.2, we show the global overview of statistical power across all simulations in African populations; in Figure 3.3, we show the overview for European populations. In African populations, we find that the statistical power of all three RVATs is similar regardless of simulated parameters, but tend to be highest with SKAT and SKAT-O \[p_{MWU}(SKAT, SKAT-O)=0.795; p_{MWU}(KBAC, SKAT-O)=0.002304\]. As expected, power is higher when the causal signal is more concentrated (e.g. when heritability is high or the effect sizes are large due to few causal loci). However, in European populations, the power of KBAC is much lower across all parameters than in SKAT or SKAT-O \[p_{MWU}(SKAT, SKAT-O)=0.730; p_{MWU}(KBAC, SKAT-O)=1.883e-6\], suggesting that burden tests are less effective given the explosive growth in European populations in the Tennessen demographic model. In the main figures of this thesis, we will focus on SKAT-O and give the results of SKAT and KBAC in the appendices.
Figure 3.3 A global overview of the statistical power in 50,000 African individuals of a burden test (KBAC), a variance-component test (SKAT), and a combined test (SKAT-O) for all parameters shown in Table 2.1 simulated under an Out-of-Africa demographic model. Each point represents a genetic architecture tested with 10 independent simulations under the RVAT indicated; lines connect the same simulated parameters across RVATs to show that, generally speaking, the rank of statistical power is preserved across RVATs.
Figure 3.4 A global overview of the statistical power in 50,000 European individuals of a burden test (KBAC), a variance-component test (SKAT), and a combined test (SKAT-O) for all parameters shown in Table 2.1 simulated under an Out-of-Africa demographic model. Each point represents a genetic architecture tested with 10 independent simulations under the RVAT indicated; lines connect the same simulated parameters across RVATs to show that, generally speaking, the rank of statistical power is preserved across RVATs.

As rare variants explain more genetic variance of the trait, RVAT power decreases

We then ask how RVAT power changes as a function of genetic architecture. In Figure 3.3, we show that as $V_{0.01}$ increases (i.e., as rare variants explain increasing amounts of the genetic variance of the trait), the power of SKAT-O decreases. This pattern holds across all sampling strategies and for different levels of polygenicity (Figure S3). These results show patterns that will repeat in future sections: the extremes study design demonstrates the best overall power, followed
by 50/50 and then random. Further, a more concentrated signal (higher heritability and/or lower number of causal loci, see Supplemental Figures) improves power. We found that as the functional form relating effect size to selection coefficient changes from $\tau=0.5$ to $\tau=1$, power increases slightly again, suggesting that $V'_{0.01}$ may be an overly simplistic characterization of the genetic architecture. Finally, applying SKAT-O to imputed data (bottom facet) reproduces all of the patterns we see when RVATs are applied to sequencing data (top facet), albeit with slightly worse power. Across tests, we see what Figures 3.3 and 3.4 indicated – that SKAT and KBAC perform worse than SKAT-O in general across all genetic architectures (see Figures B.1. through B.10.). We also note that RVATs are less effective in the European population, reflecting the increased number of rare variants as a result of explosive recent population growth in the Tennessen demographic model.
Figure 3.5 The statistical power of SKAT-O in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Chapter 4: The statistical power of genetic association tests under different study designs
Using extreme cases/controls as sampling strategy improves statistical power of SKAT-O

The number of individuals sequenced as part of a study is a key design parameter of that study. To understand how increasing the number of individuals improves the statistical power of SKAT-O, we simulate across genetic architectures and study designs to find the increase in power per individual using SKAT-O from 2,500 to 20,000 individuals (Figure 4). In the extremes study design, where half of the individuals in the panel are selected from the extreme cases and half of the individuals are selected from the extreme controls (from a total population of 50,000 individuals), we find that mean power gain is zero. Increasing the number of individuals in this design means more individuals are drawn from closer to the mean of the distribution, so power is already maximized with a smaller sample of 2,500 individuals (and may actually decrease under some scenarios!). In the random and 50/50 study designs, increasing the size of the case/control panel increases the number of relevant individuals, and so mean power gain is approximately 2e-5 per individual added. This increase is highly dependent on the genetic architecture underlying the trait of interest. These increases are consistent across populations and across RVATs studied, though the increase in power per individual added is lower using KBAC (see Figures C.1., C.2., and C.3.).
Figure 4.1 Increase in SKAT-O power as a function of sample size in an African population. SKAT-O power increases when increasing the sample size in non-extreme sampling strategies. Each point represents the slope from increasing the number of individuals in the case/control panel under a simulated genetic architecture.

**RVATs perform nearly as well on imputed data as they do on sequence data**

Most genetic association studies have started with genotyping arrays to collect genomic data, followed by imputation against WGS reference panels to maximize discovery potential with single variant analyses. As WGS cost falls, more studies will conduct large-scale WGS, but here we ask if there is a potential opportunity to discover rare variant associations with imputed data. In Figure 5, we compare the mean power of SKAT-O when applied to genotyped-then-imputed samples to the mean power of SKAT-O applied to sequencing data from the same samples. We find that the decrease in power is minimal. Indeed, we find a robust linear relationship between RVAT power with sequencing vs imputed data, suggesting that for all scenarios evaluated here, imputation loses 10% power, on average, compared to sequencing data. This trend holds across European and African populations and across tests, except in the case of KBAC with European populations (Figure C.5.). This suggests that burden tests are less capable
of handling recent explosive population growth leading to an introduction of more ultra-rare variants within a population.

Figure 4.2 The mean power of SKAT-O across different genetic architectures using imputed data, compared to using sequence data. Each point represents a different simulated genetic architecture where we vary the number of causal bins (10 or 100), heritability (0.2 or 0.8), sampling strategy, \((\rho, \tau)\) for the underlying phenotype distribution, and the number of simulated case/control individuals in the study. The red line indicates the best linear fit with a slope is 0.92, interpreted as imputed data produces 92% of the power expected from WGS data.

RVATs under a GWAS peak

The general process of discovering genetic associations typically begins with genotyping and imputing a sample of individuals, followed by GWAS. The (typically unknown) genetic architecture of the trait determines the likelihood that a common variant will be detected with GWAS, and whether a rare variant association signal should be expected. Rastering over
parameters of our phenotype model, a genome-wide significant single marker association (GWAS) was identified at 44.4% of causal loci. Figure 4.3 shows the power of SKAT-O using sequencing or imputed data conditional on seeing (circles) or not seeing (x’s) a statistically significant GWAS hit at a causal locus. We find that under all phenotype model parameters and sampling strategies evaluated, when a GWAS hit is identified, SKAT-O has at least 70% power to detect a rare variant signal with sequence data (and slightly less power with imputed data), except when $V_{0.01}'$ is almost 1. If no GWAS peak is identified, there is considerably less power to identify a rare variant signal (and power further erodes as the genetic variance explained by rare variants increases). In a simulated European population, the power of RVATs is decreased due to increased numbers of rare variants and that patterns is observed whether GWAS detected a causal loci or not. Across tests, SKAT-O power is highest regardless of whether the causal loci was detected by GWAS or not.
Figure 4.3 The statistical power of GWAS given the results of SKAT-O in an African population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of SKAT-O; the colours show the underlying number of causal loci and heritability of the trait.

We then mimic the process of first doing locus discovery on a sample of imputed individuals followed by sequencing for different sampling strategies. In Figure 4.4, we show that sequencing data has at least 75% power to replicate causal loci identified with imputed data (regardless of the genetic architecture and case-control sampling strategy). However, when no association is found with imputed data, power to identify causal loci with sequencing data is highly dependent on the case-control sampling strategy, and the overall heritability and genetic architecture of the trait (with power generally decreasing as $V_{0.01}$ increases). In an European
population, the test power is lower across the board. Across tests, SKAT-O remains the highest powered test, although SKAT remains fairly close in performance.

Figure 4.4 SKAT-O power using sequencing data, given the results of SKAT-O applied to imputed data in a simulated African population. The shape indicates whether SKAT-O applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Window of discovery around causal loci

In Figure 4.5, we plot the probability of SKAT-O detecting an association signal as a function of the distance from a causal locus. To benchmark the width of this discovery window, we use the full-width half-maximum statistic, which is the distance at which the probability of a significant association crosses below 50% of its maximum value (i.e. falls below 50% of the power estimated at the causal locus). Consistent with previous results, the full-width half-maximum is largest when there is a large amount of heritability concentrated in few causal loci and under the extremes study design. The larger points in Figure 8 represent this window of discovery, which is, on average, 34.3kb (sd 18.4kb) in the random study design, 42.8kb (sd 19.3kb) in the 50/50 design, and 64.3kb (sd 34.2kb) in the extremes design. We observe similar trends within populations and across tests as observed before – the window of discovery is larger for the simulated African population than for the European population, and SKAT-O gives the largest window of discovery across tests, reflecting its ability to detect causal regions and neighboring loci.
Figure 4.5 The window of discovery around causal loci in an African population using SKAT-O, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and $V_{0.01}$ thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Chapter 5: Discussion
Genome-wide association studies (GWAS) so far have produced thousands of SNP associations for hundreds of traits (Buniello et al., 2019). However, in these GWAS, the associated SNPs do not recapitulate the estimated heritability of the trait, leading to the problem of “missing heritability”. Though there are many proposed sources of this missing heritability, one popular hypothesis is that this missing heritability resides in rare variants. This has led to the development of rare variant association tests and massive investment in large whole genome sequencing studies. With these tests and this data becoming more and more prevalent, we look at how to optimize the design of a rare variant association study to maximize power.

It is clear that RVATs can be very powerful for detecting associations under simple genetic architectures [like when the effect size is proportional to $\log_{10}(\text{MAF})$ as proposed by (Wu et al., 2011)]. Such phenotype models do not take into account evolutionary forces like natural selection and demography, and it is well appreciated that genetic architectures are sensitive to these non-equilibrium evolutionary forces (Gazave, Chang, Clark, & Keinan, 2013; Simons, Turchin, Pritchard, & Sella, 2014). Uricchio et al presented a phenotype model that accounts for selection and pleiotropy and showed that existing RVATs struggle at realistic variance explained in genes across different human demographic histories (Uricchio et al., 2016). The Uricchio model captures modularity through the parameter $\rho$ and the relationship between selection and effect size through $\tau$, which enables a thorough exploration of different genetic architectures a trait could have (Figure 1).

We showed analytically that there is a significant amount of genetic variance explained in rare variants across different ($\rho, \tau$) parameterizations under the Uricchio model (Figure 1), particularly when $\tau$ is equal to 1. These results are not surprising, as it has been shown that a substantial amount of heritability derives from rare variants in real traits like gene expression
(Hernandez et al., 2017), height and BMI (Wainschtein et al., 2019). Taken together, the significant amount of heritability explained by rare variants under different parameterizations of the Uricchio model shows that RVATs have the potential to associate much of the causal variation underlying a complex trait.

Many existing rare variant association tests were thoroughly characterized by (Moutsianas et al., 2015). We chose the most powerful representatives of the three classes of RVATs to use in our study: a variance-component test (SKAT), a burden method (KBAC), and a combined method (SKAT-O). Across all genetic architectures and study designs, we found that SKAT-O is the best performer, so we used SKAT-O in all further analyses on RVAT power in a case/control association study.

To run a case/control association study, the first step is to determine which individuals to select for your study, and how to acquire their genetic data. We simulated three different sampling strategies: randomly sampling cases and controls proportional to the trait prevalence; sampling half of your study size from cases and half from your controls; and sampling individuals from the extreme tails of a quantitative distribution [or a proxy underlying the trait such as bronchodilator response (Spear et al., 2018), for example]. Our results show that choosing from the tails of an underlying quantitative distribution produces the best power. This means for any case/control association study, spending some time to find the extreme tails of an underlying quantitative distribution for a trait will likely produce the best possible RVAT power.

We considered two ways of acquiring genetic data: using a genotyping array followed by imputation against a large reference panel, and direct sequencing of your study sample. Although a $1,000 whole genome is now possible, over the sample sizes required for an effective rare variant association study, the cost is prohibitive except for the largest consortia. Using genotyping arrays
then imputing is still much less expensive than WGS (Quick et al., 2019), which could enable more than 5x more genotyped samples than WGS samples. Practically speaking, this multiplier could be higher, as the cost to enroll a new individual and acquire their WGS and phenotype data is much higher than genotyping.

Applying SKAT-O to imputed data is expected to have lower power for several reasons. First, imputation accuracy decreases as MAF decreases (Howie, Marchini, Stephens, & Chakravarti, 2011; Quick et al., 2019), meaning fewer rare variants will be accurately imputed and correctly identified in the study sample. Second, imputation accuracy is highest when the study sample population and the reference panel population match, and this is not guaranteed to be the case, particularly when the study sample is from a minority population or an admixed population. Third, a majority of rare variants carried by the imputed samples are unlikely to be carried by the reference panel.

Comparing SKAT-O power across genetic architectures and study designs, we show that genotyping then imputing is about 90% as powerful as WGS using the same number of individuals. We note here again that genotyping error is not simulated, and so this result is an upper bound on SKAT-O power using imputed data. This implies that using genotyping then imputing with a larger sample size could produce as much if not more power than a smaller WGS sample. For most current rare variant association studies, our results suggest that using genotyping then imputing is the most cost-effective way to proceed. We also looked at the increase in SKAT-O power using WGS after running a genotyping and imputation study; there is a boost in SKAT-O power when using WGS data following imputed data, but the trade-off between cost and power is something to be considered on an individual study basis.
The next step in characterizing RVAT power is to consider the genetic architecture of the trait of interest. Though complex trait architectures are not thoroughly understood, we used the Uricchio model to simulate different architectures and label these architectures using the amount of cumulative genetic variance explained by all variants under 1% minor allele frequency ($V'_{0.01}$). We show that SKAT-O power decreases as $V_{0.01}$ increases, meaning SKAT-O performance is worst when rare alleles make the largest contributions to trait variance. Although counterintuitive, as one would expect RVATs are best tuned for the scenarios where rare variants matter most in the genetic architecture, our result mirrors the findings of Uricchio et al, 2016. One explanation is that as $V_{0.01}$ goes up, the proportion of $V_{0.01}$ due to singletons and other ultra-rare alleles increases as well, and statistically associating these ultra-rare alleles is difficult in the RVAT frameworks we evaluated here. We also note that the explosive exponential growth of the Tennessen demographic model used to simulate genetic data leads to an excess of ultra-rare alleles compared to the neutral expectation, such that both cases and controls harbor many ultra-rare variants (thereby confounding RVAT power). The converse may be true in a population that has not experienced recent explosive population growth – with fewer relative ultra-rare alleles, RVAT power may increase in these populations.

With the decrease in power as rare variants mattered more, we wondered whether nearby regions in rare variant-dominated architectures would provide additional information. We looked at how the probability of SKAT-O detecting a causal region decreases as a function of distance from a causal region. The results suggest that in an unbiased window-based approach to scanning the genome with SKAT-O, positive hits that are not in causal regions may be useful in helping identify true causal regions, although again only in genetic architectures where rare variants do not contribute the majority of genetic variance. Interestingly, the power ranking of study designs is
inverse of the ranking of precision, meaning that with higher power comes a larger window of discovery.

We also looked at the statistical properties of a common analytical path from GWAS to RVATs, and from imputed data to sequence data. We found that GWAS and SKAT-O are generally concordant, with causal regions identified by GWAS being identified by SKAT-O, while a smaller proportion (~15%) of causal regions are identified by SKAT-O and not by GWAS. We see little downside in testing for causal regions using SKAT-O following GWAS, with the ability to pick up additional causal regions on the same data. We caution that this effect declines significantly as rare variants explain more of the genetic variance.

Finally, the number of loci contributing to a trait (or its polygenicity) may be another important component of the trait’s architecture. It is not surprising that we found that for a fixed heritability of a trait, RVAT discovery power is higher when there are fewer true causal loci (as effect sizes are concentrated into fewer variants). However, it is possible that the polygenicity of a trait could be constraining the possible range of genetic architectures.

We observed the same patterns for all parameters across populations and across tests. In general, the European population experienced a recent explosive population growth, leading to the introduction of larger numbers of ultra-rare variants (singletons, mostly) compared to the African population. This introduction of ultra-rare variants reduces the power of every statistical test used, particularly the burden test KBAC. Given that KBAC is the worst performing test across the genetic architectures and study designs, we suggest using SKAT-O as the default standard RVAT for future studies.
This study has a few limitations. It is based on simulated data that matches inferred human evolutionary history (including selection, and demographic history) but these models and simulations are incomplete representations of nature. We do not explore the effects of gene size, mutation rate, haplotype length, or degree of linkage disequilibrium between causal regions. We do not consider the differences between coding and non-coding regions, which have different selection coefficient distributions and potentially different contributions to the genetic architectures for a trait. Future work should consider a phenotype model where the function of a region is taken into account, as ENCODE (The ENCODE Project Consortium, 2012) and other consortia are rapidly adding more dimensions to genomic data. One major shortcoming is that we analyze only African and European populations in this study. With significant growth in admixed populations already happening - the US Census in 2014-15 predicts that the US will be a “majority-minority” country by 2050 (Projections of the Size and Composition of the U.S. Population: 2014 to 2060, 2014), meaning significant growth in African-American and Latino populations - it will be important to study association testing power in admixed populations. We also believe that incorporating functional annotations, evolutionary forces, and admixture into rare variant association tests would significantly improve statistical power.

This study shows that RVATs have little power in regimes where rare variants truly matter. We suggest three future directions to investigate the role of rare variants within genetic architectures and improve upon RVAT power. First, Haseman-Elston regression on simulated genetic architectures revealed that simulating a 5Mb locus generates too few rare variants to get a significant effect. At maximum, we simulate causal regions totaling 1Mb, which is insufficient to have ultra-rare variants make a large impact on heritability. Under some simulated genetic architectures, these ultra-rare variants are expected to contribute the majority of heritability. This
is reflected in real complex traits, where rare variants contribute a majority of heritability in gene expression (Ryan D. Hernandez et al., 2017). One future direction is therefore to discover the causal region size necessary for ultra-rare variants to make a significant contribution to heritability under different genetic architectures, which can help guide design of genetic association studies when the size of causal region is better understood.

Another obvious avenue is to incorporate functional annotations in improving the power of genetic association studies. With the ENCODE project (The ENCODE Project Consortium, 2012) annotating much of the genome, we now have the functional annotations to pair with genetic data gathered in any study. Though this has been pursued in various guises to date (Hormozdiari et al., 2018; Kichaev et al., 2014; Pickrell, 2014), the results have been lackluster and further improvements in both data and method will improve RVAT power.

Last, we propose simulating theoretical models that explain patterns of heritability within complex traits like the omnigenic model (Boyle, Li, & Pritchard, 2017) and various polygenic models [like (Uricchio et al., 2016) and others] to investigate the properties of RVATs under these theoretical models and suggest mechanisms for improving current statistical methods. This work would also serve as a reference for future genetic association studies as a comparison for their empirical results to determine the closest model underlying the complex trait under study.
REFERENCES


architecture of diseases and complex traits. *Nature Genetics.*

https://doi.org/10.1038/s41588-018-0148-2


https://doi.org/10.1534/g3.111.001198


https://doi.org/10.1371/journal.pgen.1004722


https://doi.org/10.1016/j.ajhg.2012.06.007


https://doi.org/10.1371/journal.pgen.1001156


Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLOS Genetics*, 11(4), e1005165. https://doi.org/10.1371/journal.pgen.1005165


Appendix A: Supplemental Material to Chapter 3

Figure A.1 The statistical power of SKAT-O in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.2 The statistical power of KBAC in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.3 The statistical power of KBAC in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.4 The statistical power of SKAT in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.5 The statistical power of SKAT in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual European population.
Figure A.6 The statistical power of SKAT-O in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.7 The statistical power of SKAT-O in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual European population.
Figure A.8 The statistical power of KBAC in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.9 The statistical power of KBAC in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual European population.
Figure A.10 The statistical power of SKAT in African populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual African population.
Figure A.11 The statistical power of SKAT in European populations across different sampling strategies (columns) and across different sequencing methods (rows), as a function of the proportion of genetic variance explained by that genetic architecture at MAF=1%. Each point represents 20 independent simulations of 10 or 100 causal loci of 10kb each across a 5Mb simulated region for a given genetic architecture for a 50,000 individual European population.
Appendix B: Supplemental Material to Chapter 4

Figure B.1 Increase in SKAT-O power as a function of sample size in an European population. SKAT-O power increases when increasing the sample size in non-extreme sampling strategies. Each point represents the slope from increasing the number of individuals in the case/control panel under a simulated genetic architecture.

Figure B.2 Increase in KBAC power as a function of sample size in an African (left) and European (right) population. KBAC power increases when increasing the sample size in non-extreme sampling strategies. Each point represents the slope from increasing the number of individuals in the case/control panel under a simulated genetic architecture.
Figure B.3 Increase in SKAT power as a function of sample size in an African (left) and European (right) population. SKAT power increases when increasing the sample size in non-extreme sampling strategies. Each point represents the slope from increasing the number of individuals in the case/control panel under a simulated genetic architecture.

Figure B.4 The mean power of SKAT-O in an European population across different genetic architectures using imputed data, compared to using sequence data. Each point represents a different simulated genetic architecture where we vary the number of causal bins (10 or 100), heritability (0.2 or 0.8), sampling strategy, (ρ, τ) for the underlying phenotype distribution, and the number of simulated case/control individuals in the study. The red best fit line has slope 0.89.
Figure B.5 The mean power of KBAC in an African (left) and European (right) population across different genetic architectures using imputed data, compared to using sequence data. Each point represents a different simulated genetic architecture where we vary the number of causal bins (10 or 100), heritability (0.2 or 0.8), sampling strategy, $(\rho, \tau)$ for the underlying phenotype distribution, and the number of simulated case/control individuals in the study. The best fit line in red has a slope of 0.89 in Africans and 0.79 in Europeans.

Figure B.6 The mean power of SKAT in an African (left) and European (right) population across different genetic architectures using imputed data, compared to using sequence data. Each point represents a different simulated genetic architecture where we vary the number of causal bins (10 or 100), heritability (0.2 or 0.8), sampling strategy, $(\rho, \tau)$ for the underlying phenotype distribution, and the number of simulated case/control individuals in the study. The best fit line in red has a slope of 0.92 in Africans and 0.89 in Europeans.
Figure B.7 The statistical power of GWAS given the results of SKAT-O in an European population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of SKAT-O; the colours show the underlying number of causal loci and heritability of the trait.
Figure B.8 The statistical power of GWAS given the results of KBAC in an African population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of KBAC; the colours show the underlying number of causal loci and heritability of the trait.
Figure B.9 The statistical power of GWAS given the results of KBAC in an European population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of KBAC; the colours show the underlying number of causal loci and heritability of the trait.
Figure B.10 The statistical power of GWAS given the results of SKAT in an African population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of SKAT; the colours show the underlying number of causal loci and heritability of the trait.
Figure B.11 The statistical power of GWAS given the results of SKAT in an European population, across different sequencing methods (rows) and across different sampling strategies (columns), as a function of the cumulative genetic variance explained by variants under 1% minor allele frequency. The shape shows the prediction of SKAT; the colours show the underlying number of causal loci and heritability of the trait.
Figure B.12 SKAT-O power using sequencing data, given the results of SKAT-O applied to imputed data in an European population. The shape indicates whether SKAT-O applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Figure B.13 KBAC power using sequencing data, given the results of KBAC applied to imputed data in an African population. The shape indicates whether KBAC applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Figure B.14 KBAC power using sequencing data, given the results of KBAC applied to imputed data in an European population. The shape indicates whether KBAC applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Figure B.15 SKAT power using sequencing data, given the results of SKAT applied to imputed data in an African population. The shape indicates whether SKAT applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Figure B.16 SKAT power using sequencing data, given the results of SKAT applied to imputed data in an European population. The shape indicates whether SKAT applied to imputed data correctly identified the causal locus (circles) or missed it (x). The colours show the underlying causal number of loci and heritability of the trait.
Figure B.17 The window of discovery around causal loci in an European population using SKAT-O, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and V_{0.01} thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Figure B.18 The window of discovery around causal loci in an African population using KBAC, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and $V_{0.01}$ thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Figure B.19 The window of discovery around causal loci in an European population using KBAC, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and $V_{0.01}$ thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Figure B.20 The window of discovery around causal loci in an African population using SKAT, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and $V_{0.01}$ thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Figure B.21 The window of discovery around causal loci in an European population using SKAT, shown as the fraction of simulations that result in a statistically significant RVAT p-value as a function of distance from the nearest causal locus. Different sampling strategies are shown in columns, and $V_{0.01}$ thresholds are shown in rows. Error bars are binomial standard errors of the mean. Bigger points represent full-width half-maximum points.
Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

_____________________________________   ________________
Author Signature        Date

June 8, 2019

Date