Significance and Background

Most common diseases (those with prevalence > 0.1%, including for example schizophrenia, autism, and diabetes to name a few) are complex. In contrast to Mendelian diseases, the risk of developing complex diseases is affected by variants at many loci across the genome, each of which typically has low penetrance. While the genetic basis of many Mendelian diseases have been successfully mapped during the 20^{th} century, the systematic mapping of variants underlying complex diseases became possible only over the last decade with the maturation of the genome wide association study approach.[1,2,?] GWAS and related variance partitioning methods have largely validated estimates of heritability based on the twin studies (although, in some cases the latter were slight overestimates), and thus shown that the heritabilities of common complex diseases are generally high. They have also shown that the bulk of the variance in risk can be attributed to a very large number of common alleles (generally each with small effects),[3,4,5,6,4] while sequencing and exome studies indicate a potential role for rare variants with larger effect as well.[7,8,9] While the significant associations for even the most successful disease GWAS number in the hundreds, statistical analyses of GWAS results indicate that many thousands of variants affect their susceptibility (refs) .

However, we still remain unclear about why complex diseases are as common as they are, and what determines their genetic architecture, i.e. the observed relationships between variants' effect size and frequency. In contrast to Mendelian diseases, where we have a fairly good quantitative understanding of how the mutational target size, recessivity, selection and demography combine to determine disease prevalence and allele frequencies, we lack a satisfactory quantitative framework describing how these factors and others determine the prevalence and architecture of complex diseases.

We do have, nonetheless, a qualitative understanding of some of the population genetics process that affect genetic variation for complex diseases. Perhaps most straightforward is that genetic variation for disease risk reflects a balance between mutation and selection.[10] This suggests that a given genetic disease is common if it has a large mutational target sizes; that variation in prevalence among diseases with similar fitness costs reflects differences in target sizes; and that variation in genetic architectures arises from differences in the distribution of mutational effect sizes. Similarly, differences in the fitness costs among diseases should lead to differences in the strength of selection on variants (that have similar effects on disease risk), which would also cause variation in both architecture and prevalence.

However, selection acting on mutations due to their effect on a given disease is only part of the story, as ample evidence suggests that variants affecting a given disease often have pleiotropic effects on other traits that are themselves under selection.[11,12] The genetic architecture of a given disease may then reflect both direct selection due to the disease as well as indirect selection due to these pleiotropic relationshipsto other diseases or quantitatie traits.[13,14,15,16,17,18] The effects of pleiotropy on disease prevalence and architecture are far from straightforward and await a more systematic treatment (see below).

Lastly, recent changes to the environment have also likely affected the prevalence and architecture of complex diseases. For example, the "thrifty gene" hypothesis[19,20] posits that diseases such as type 2 diabetes and hypertension exhibited a drastic increase in prevalence due to profound recent changes in diet and lifestyle, which caused a mismatch between the ancestral human environment and present conditions. A recent study documenting rapid changes in the prevalence of type 2 diabetes and obesity in response to food shortage and economic crisis in Cuba in the 1990s,[21] illustrates this effect. It has been further suggested that such profound changes may disrupt developmental canalization, leading to wholesale changes to the distribution of genetic effects on disease risk.[22,23] While some of these ideas likely apply to some diseases (and they are also not mutually exclusive), they generally lack quantitative predictions that allow for them to be rigorously tested and for their effects to be quantified.

Indeed, there has been surprisingly little work aimed at relating population genetic processes with the results emerging from GWAS. In this vein, two studies been especially influential. Pritchard[24] considered a model in which a mutation's effect on disease risk is completely uncoupled from its fitness cost, i.e. an extreme pleiotropic limit. While this paper played a central role in grounding the debate surrounding the common disease-common variant hypothesis in population genetic theory,[25] many complex diseases are known to affect fitness (refs) and the magnitude of variants' effects on different traits are often correlated,[26,11,12] suggesting that it is unrealistic to assume that a mutation's effect on disease risk is independent of its effect on fitness. A second influential study by Eyre-Walker[27] posited that all mutations are deleterious, but allows for their effect size, α, and selection

coefficient, S, to be related by the functional form: $\alpha = \delta S^{\tau}(1+\epsilon)$, where δ is a randomly chosen sign (i.e. 1 or -1), ϵ is a normally distributed noise term, and τ is a "coupling parameter" meant to capture the effect of pleiotropy. When τ equals 1 a mutation's selection coefficient is tightly coupled to its effect on disease, and when it equals zero the model converges to Pritchard's pleiotropic limit.[24] Eyre-Walker's model has also been used as a basis for inferences based on GWAS results for type 2 diabetes[28,29] and prostate cancer[?] (see Aim 2 below for more background on the Eyre-Walker model in this context). However, the relationship between effect size and fitness in this model has little justification beyond mathematical convenience (and fact possesses some odd features, such as fitenss equivalence of mutations with opposing effects on disease). Moreover, more recent work illustrates that assuming a different relationship profoundly affects the predictions about genetic architecture.[30] This suggests that understanding the relationship between GWAS findings and population genetics processes will require a more principled way of relating the effects of mutations on disease phenotypes with their effect on fitness.

Here I propose to develop generative models for the way that genetic architecture and disease prevalence are affected by evolutionary parameters such as mutation, selection, pleiotropy, and recent changes to the environment and to demography. I will then develop and apply statistical approaches to test these models and estimate their parameters based on data from GWAS. These models and inferences will substantially advance our understanding about the causes, prevalence and architecture of complex diseases, and will provide us with the first reliable estimates of their underlying evolutionary parameters.

Approach

Preliminary Results All our models build on the canonical quantitative genetics description of the relationship between genotype and complex disease risk. In this model, each individual's risk (R) of developing disease is a monotonic function (ℓ) of an underlying (and generally unobserved) disease liability trait (Z), such that

$$R = \ell(Z), \qquad Z = \sum_i \alpha_i g_i + \epsilon \tag{1}$$

where α_i and g_i are the liability scale effect size and genotype respectively at site i, and ϵ is a normally distributed deviate which captures stochastic variation in risk among genotypes with the same mean liability, (i.e. the "environment" of classical quantitative genetics). The canonical models used to relate genotype to disease risk in human genetics, including Wright's threshold model,[31,32,?,33] the logistic model typically employed in GWAS,[1] and Risch's multiplicative model[34] are all included in this description and correspond to different choices for the function ℓ. Previous work[35,36] indicates that the exact choice of ℓ is unlikely to be particularly important, and as such I will take ℓ to be the probit link function of Wright's threshold model for the remainder of this proposal, though I will explore other choices to understand what impact they might have on our results.

Our simplest model and departure point considers the impact of mutation and selection on a single disease in a constant environment. We assume the standard Wright-Fisher model of a diploid panmictic population of size N, including infinite sites mutation, free recombination, Mendelian segregation and viability selection. Liability increasing and decreasing mutations arise at genome wide rates μ^+ and μ^-, with effect size distributions $f^+(\alpha)$ and $f^-(\alpha)$ respectively. Individuals with the disease have their fitness reduced by a factor $1-S$.

The genetic architecture of the disease under this simple model is shaped by mutation-selection-drift balance, and can be related to standard results from quantitative genetics. An equilibrium is reached when

$$U^+ - U^- = V_A \underbrace{S\phi\left(\Phi^{-1}(1-P)\right)}_{\text{selection gradient}} \tag{2}$$

where $U^+ = \mu^+ \int_0^\infty \alpha f^+(\alpha)\, \mathrm{d}\alpha$ is the total per generation mutational increase in liabilty (with U^- defined similarly as mutational pressure toward decreased liability). The term on the right hand side accounts for the selection pressure toward lower liability; V_A is the additive genetic variance of liability, and the compound term it multiplies is a selection gradient in the standard quantitative genetics sense.[?] ϕ and Φ are the Gaussian pdf and cdf repsectively, and arise because of our choice of the theshold model for ℓ. P is the disease prevalence. An individual allele in this model with effect α on disease liability will experience a selection coefficient

$$s = -2\alpha S\phi\left(\Phi^{-1}(1-P)\right) \tag{3}$$

against the liability increasing homozygote, and will evolve under fitness additivity so long as α is small. Provided the model parameters (i.e., the mutation rates, distribution of effect sizes and the fitness cost of the disease), equa-

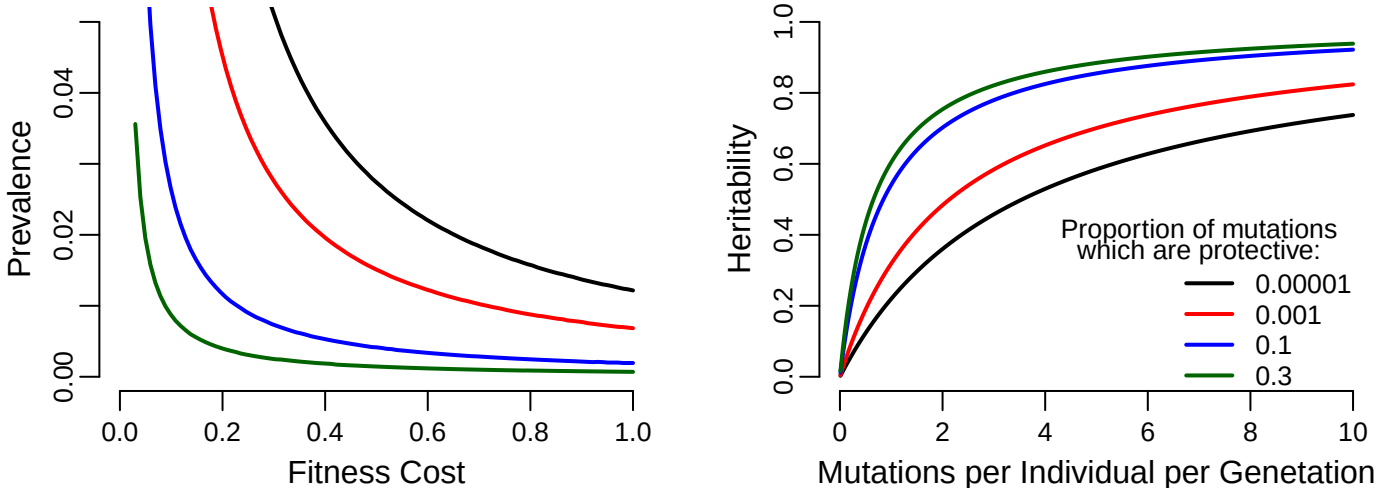

Figure 1: Solutions to the model specified by equations (2) and (3). The left panel shows the diseaes prevalence at equilibrium as a function of the fitness cost for various different choices of the strength of mutation bias. The right panel shows the heritability of disease liability as a function of the total size of the mutational target, again for different strengths of mutational bias.

tions (2) and (3), allow one to solve for the disease prevalence and summaries of diseases' genetic architecture (details not shown for brevity).

The solution provides a few curious insights. First, while disease prevalence depends on the fitness cost and the mutational bias, it is insensitive to the total mutational rate. Second, selection coefficients experienced by individual alleles (and therefore their frequency distributions) are independent of the fitness cost of the disease because the effect of changes to fitness cost are precisely canceled by changes in prevalence, and it follows that bulk properties like the heritability of liablitiy are similarly insensitive to changes in fitness cost. While preliminary, these results already contradict common intuitions in the field, such that alleles underlying more harmful diseases should be subject to stronger selection. Our preliminary results therefore already illustrate the value of a more rigorous study of population genetic models of complex diseases.

Specific Aim 1: Develop Models Relating Population Genetic Processes with the Architecture of Complex Diseases

Building on my preliminary studies, I will also model the effects of environmental change, pleiotropy and non-equilibrium demography on the genetic architecture of complex diseases. I will solve these models to obtain closed forms for summaries of genetic architure and disease prevalence. In cases where analytical treatment proves difficult, I will use simulations and other numerical methods to obtain solutions, and all analytical results will be verified in this manner as well. Modeling efforts will be targeted toward identifying how each of these factors affect architecture and prevalence, with the later goal of using GWAS results to ascertain their effects on human diseases (aim 2). With that in mind, a major product of this theoretical aim (not discussed beyond this point) will be a thorough investigation of the identifiability of parameters introduced below, and subsequently a general guidelines to ground future discussions about what aspects of complex genetic disease biology and evolutionary history can or cannot be inferred from GWAS.

Environmental Change The most straightforward way to model environmental change may also be the most relevant one: namely, to assume that the mean or the variance of the environmental component of liability variance shifts or increases suddenly. We are particularly interested in the scenario where an environmental shift has just occurred (e.g. diet and lifestyle on type 2 diabetes prevalence), but there has not been sufficient time for allele frequencies to evolve away from their previous equilibria. In the case of a shift in the environmental mean, the effect is simply an increase (or decrease) in prevalence such that $P_{new} = 1 - \Phi\left(\Phi^{-1}\left(1 - P_{old}\right) - \delta\right)$, where δ is the shift in the environmental contribution to liability measured in units of the phenotypic standard deviation (Φ, again, is the Gaussian cdf). In the case of a change in the environmental variance, both the prevalence and the heritability are affected. If the environmental variance is increased by an amount ψ (given in units of the pre-change phenotypic variance), then heritability is decreased such that $h^2_{g,new} = \frac{h^2_{g,old}}{1+\psi}$, while prevalence will

increase to $P_{new} = 1 - \Phi\left(\frac{\Phi^{-1}(1-P_{old})}{1+\psi}\right)$.

Such changes to the environment would affect the architecture observed in GWAS. While the allele frequencies and liability scale effect sizes of mutations would be unchanged, the effect size on the risk scale, estimated in GWAS in terms of the odds-ratio, would be altered. For example, for a mutation with a fixed effect on liability, an environmentally induced increase in prevalence should lead to an increased odds ratio, as a higher prevalence also means that the proportion of individuals whose liability is just shy of the threshold (and who could therefore be pushed over by an extra mutation at a single site) is higher. This suggests that the distribution of odds ratios (e.g. observed in GWAS) together with estimates of the present day prevalence and fitness cost might allow for inference of the size or type of environmental shift, and/or ancestral prevalence.

In addition to these simple scenarios, I will also study how more involved patterns of environmental change influence disease prevalence and the genetic architecture of risk. For example, we will consider the case where the environmental change occurred sufficiently long ago to have impacted the distribution of allele frequencies, as might plausibly have occurred when archaic humans migrating out of African encountered their new Eurasian environment.

Pleiotropy Mounting vidence suggests that alleles that affect the risk of a given disease often also affects other traits,[26, 11, 12] suggesting that the selection acting on such alleles depends not only on their effect on the disease under consideration but also on these pleiotropic effects. I will focus on pleiotropic effects of two kinds. The first are pleiotropic effects on the risk of developing other diseases. An allele which increases the liability for a given disease may also increase or decrease the liability for several other diseases. The selection effect on that allele would therefore by a sum over the effect sizes on difference diseases, weighted according to the fitness cost and prevalence of each disease as in Eq (3).

I will begin by considering "isotropic" models of disease only, in which all diseases have identical parameters (here and below Θ_A represents collectively all of the parameters of the model, including the number of diseases, their fitness costs and prevalence, mutational bias, etc). Specifically, this assumption ensures that conditional on a mutation's ultimate effects on fitness, the marginal distributions of effects on liability for each disease are identical. Given a specification of parameters, this marginal distribution $(p(\alpha \mid s, \Theta_A))$ will be derived from geometric considerations of the model. Ultimately, we are interested in an expression for the joint distribution of effect sizes and frequencies (i.e. "the architecture"). In general, the marginal distributions of effect sizes and frequencies are independent conditional on the selection coefficient, and therefore

$$p(\alpha, x \mid \Theta_A) = \int p(\alpha \mid s, \Theta_A)\, p(x \mid s)\, p(s)\, \mathrm{d}s \tag{4}$$

where the distribution of frequencies, $p(x \mid s)$, can be computed from diffusion theory,[?] and the distribution of selection coefficients, $p(s)$, is generally unknown (indeed, its inference is a major goal of aim 2).

The second form of pleiotropy considered will be that due to selection on continuous (i.e. non-disease) traits. I will assume that these traits are subject to stabilizing selection, as decades of work suggest is typical (refs) . Models of stabilizing selection on quantitative traits indicate that at equilibrium, the population is held near the optimum, and selection therefore acts to reduce the contribution to phenotypic variance made by any individual allele. The result is that alleles which impact quantitative traits experience symetric underdominance for fitness (i.e. the minor allele is always selected against), where the strength of selection scales with the square of the effect size on the quantitative trait.[37]

Inclusion of pleiotropic effects on quantitative traits will therefore involve specifying the relative proportion of mutational effects on fitness which derive from disease and from continuous traits. The considerations above suggest this problem can be approached by specifying a distribution of dominance coefficients given the homozygous selection coefficient and the parameters of the model (i.e. $p(h \mid s, \Theta_A)$). Mutations with large effects on disease liability and small effects on continuous traits would exhibit directional selection but with some dominance effect, while those with small effects on diease and large effects on continuous traits would show asymetric underdominance for fitness, as there will generally be selection against the liability increasing allele, but at high frequencies mean fitness might be increased by instead fixing this allele in order to reduce phenotypic variance of the quantitative trait. Specification of this distribution of dominance coefficients in terms of the model parameters will allow a new version of Eq (4) including a second integral over dominance coefficients, thereby including pleiotropy from both diseases and continuous traits.

Finally, I will eventually relax the "isometric" assumption to understand how differences among diseases/traits

in their parameters or covariance in mutational effects might alter our understanding.

I anticipate that the work described here will be useful beyond the specific goals of this proposal. Notably, beyond studying how pleiotropy affects the architecture of one disease, it provides a framework for studying the joint architecture of multiple diseases or quantitative traits together (refs) , and a basis for examining popular models in medical genetics, e.g. that some common diseases are in fact a collections of multiple distinct but biologically related disorders that present similar symptoms.[38,?,39,40]

Non-Equilibrium Demography Recent work has illustrated that the demographic history of a population can have a profound impact on the genetic architecture of complex diseases.[41,42,43,44,45,46] Specifically, changes in population size, such as the Out of Africa bottleneck and explosive recent growth, affect allele frequencies[41,42] by changing the rate of genetic drift and the frequencies at which new mutations enter the population. These effects are modulated by the selection acting on mutations, because more strongly selected mutations tend to be younger and to segregate at lower frequencies.[43] While the effects of demographic history on genetic variation at a single site under selection are fairly well understood, we lack an understanding of its effects on genetic architecture that is grounded in realistic models of complex diseases.

To develop such an understanding, I will study how non-equilibrium demography affects genetic architecture under our models of complex disease. Specifically, I will focus on our pleiotropic models and consider three main demographic scenarios: 1) a simple bottleneck, and 2) explosive growth, which will provide qualitative insights about the effects of each, and 3) a demographic model inferred for European populations,[41,47] which will be necessary for inference (aim 2). Allelic dynamics with selection and non-equilibrium demography are generally not analytically tractable, so my analysis will rely on simulations. Simulating the complete disease models directly may be computationally difficult, as a single simulation would involve running a large population with many loci forward in time. To circumvent this, I will rely on the fact that the dynamics of an individual allele are independent of all other alleles conditional on the global parameters (fitness cost, prevalence, etc), which will allow me to simulate a single allele at a time over a grid of selection parameters and consider the effects on disease risk as a whole as an appropriately weighted grid. My investigation of the preliminary model also suggests that disease prevalence at equilibrium is fairly sensitive to population size. This is significant, specifically because the selection coefficient experienced by a particular allele depends on the disease prevalence. If prevalence evolves over time in response to changes in population size, then selection coefficients will change over time as well, and I will need to incorporate this effect into my simulations as well.

Additional Complications While the factors listed in this section are obviously not exhaustive, modeling their effects will clearly advance our understanding of the way salient population genetics processes shape the architecture of complex disease. In addition, I will study the sensitivity of my results to variations on the underlying modeling assumptions, e.g. I will analytically assess the impact of large effect mutations which influence liability and use simulations to study the sensitivity of my results to the effects of LD among variants.

Specific Aim 2: Inference of Model Parameters from GWAS of Complex Diseases

Background In principle, we can use results from GWAS in order to test models of genetic architecture and infer their underlying evolutionary parameters. Previous work from the Reich and Altshuler labs has gone the farthest in this direction and is the most closely related to the research proposed in this aim.[28,29,?] They employ an Approximate Bayesian Computation approach based on Eyre-Walker's model of architecture[27] (see Background and Significance) to infer mutational target sizes and the couplings between effect size and fitness (Eyre-Walker's τ) that are consistent with GWAS results for type 2 diabetes[28,29] and prostate cancer.[?] Specifically, they simulate genetic architectures under a grid of possible values for these two parameters, and compare summaries of architecture (e.g., the number of genome-wide significant associations[28] or the proportion of variance explained by SNPs with minor allele frequency < 5%[29,?]) between simulations and GWAS to ascertain the parameter values that are consistent with observations for these diseases. While they are able to rule out the pleiotropic and direct selection extremes (τ = 0 or 1), they fail to narrow down the range of models much further. Their inferred ranges for target sizes are also unfortunately large. This severely limits the utility of the inferences, e.g., for learning about evolutionary parameters or making predictions about the performance of future mapping strategies.

These pioneering efforts suffer from several notable problems. First, they rely on relatively coarse summary statistics that discard a great deal of information, limiting their ability to constrain the parameter space. Second, they make significant and unjustified assumption, e.g. that the distribution of selection coefficients for GWAS al-

leles resembles that for substitutions in proteins, despite mulitple lines of evidence that it does not.[48,49,50] Third, as discussed previously (see Background and Significance), the Eyre-Walker model makes some rather *ad hoc* assumptions about the relationship between effect size and selection coefficient, which make biological interpretation of the inferences it produces challenging. I will address each of these concerns in the development of my inference approach.

Approach I will develop a general and extensible composite likelihood approach to infer the parameters which govern the evolution of complex genetic disease architecture. Approaching this problem in a likelihood framework will allow for both rigorous model comparisson and parameter estimation.[51,52] The data I will rely on include paired estimates of effect sizes and allele frequencies from loci discovered in GWAS, and eventually, statistical summaries of variance contributed by loci which do not meet genome-wide significance.

The major components of the likelihood include the probability of observing K genome wide significant loci, ($p\left(K \mid \Theta_A, \Theta_S\right)$; which is Poisson conditional on the parameters), the probability density of variants with a given effect size and frequency ($p\left(\alpha_i, x_i \mid \Theta_A\right)$; given by Eq (4)), and the power to discover such a variant ($H\left(\alpha_i, x_i \mid \Theta_S\right)$, which is already well studied[53]). Here, Θ_S represents the known parameters of the association study and disease (e.g. the number of cases and controls, disease prevalence), while Θ_A as above represents the parameters of the model governing the evolution of the disease. The exact parameters denoted by Θ_A will therefore depend on the specific model being considered, but for example in the case of the isotropic model of pleiotropy from other diseases, would include the mutational target size, distribution of selection coefficients, and number of diseases contributing to pleiotropy. Together

Intuitively, the number of variants discovered (K) in a GWAS of a given sample size places some bounds on the plausible size of the mutational target and on the distribution of effect sizes, and specific knowledge of their effect sizes and frequencies (α_i and x_i) provides information about their selection coefficients and the impact of pleiotropy on the genetic architecture (see Fig 2). Unfortunately, even in very large GWAS, a significant number of variants will remain undetected, restricting our inference to use only those with the largest effect sizes. We can overcome this limitation by incorporating recently developed methods related to Yang and Visscher's "GCTA"[54] that estimate the relationship between minor allele frequency (MAF) and variance explained across all SNPs regardless of whether they are genome wide significant,[55,56] as Eq (4) can be leveraged to make quantitative predictions about this relationship. Intuitively, diseases with genetic architectures dominated by large selection coefficients on individual alleles will have more of the genetic variance explained by rare alleles, though this relationship will be moderated by additional factors such as pleiotropy and demography. Putting this additional factor together with the components discussed above, we can write the likelihood of the model parameters given the data as

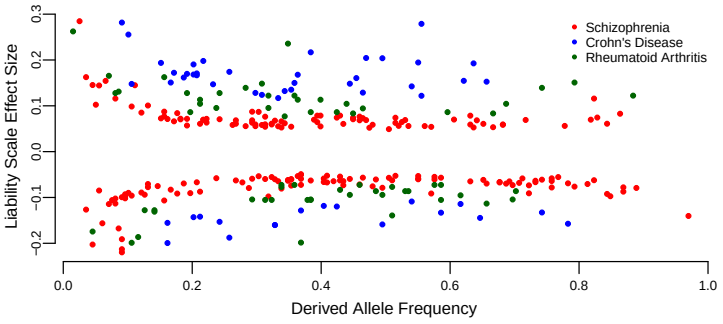

Figure 2: The relationship between allele frequency and effect size at known GWAS loci for three different complex diseases. Differences among diseases reflect differences in the underlying parameters which govern their evolution.

$$L\left(\Theta_A \mid \{(\alpha_i, x_i)\}_{i=1}^{K}, V_A^*, \Theta_S\right) = Pr\left(K \mid \Theta_A, \Theta_S\right)\left(\prod_{i=1}^{K} Pr\left(\alpha_i, x_i \mid \Theta_A\right) H\left(\alpha_i, x_i \mid \Theta_S\right)\right) p\left(V_A^*\left(x\right) \mid \Theta_A, \Theta_S\right)$$

where $p\left(V_A^*\left(x\right) \mid \Theta_A, \Theta_S\right)$ denotes this expression for how variance is distributed among allele frequencies genome-wide. There are at least two clear options for how to approach this part of the inference problem. These include 1) separate alleles into different MAF bins and estimate the proportion of variance explained by alleles in each bin,[4,56] or 2) estimating only a single variance component, but weighting the contribution of SNPs with a given frequency to the genetic relationship matrix used for this estimation according to the parameters being inferred (Θ_A).

The ultimate result of this aim will be the first statistical framework for inferring the parameters underlying complex genetic disease architecture and prevalence that is grounded in quantitative and population genetic principles. An open access software package will be made freely available so that other researchers can the inference approaches developed here to their datasets, and so that the method can be improved to incorporate

future theoretical advances.

References

[1] N Risch and K Merikangas. The future of genetic studies of complex human diseases. *Science*, 1996.

[2] P M Visscher, M A Brown, and M I McCarthy. Five years of GWAS discovery. *The American Journal of …*, 2012.

[3] The International Schizophrenia Consortium, Manuscript preparation, Data analysis, GWAS analysis subgroup, Polygene analyses subgroup, Management committee, Cardiff University, Karolinska Institutet/University of North Carolina at Chapel Hill, Trinity College Dublin, University College London, University of Aberdeen, University of Edinburgh, Queensland Institute of Medical Research, University of Southern California, Massachusetts General Hospital, and Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, June 2009.

[4] S Hong Lee, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, and Naomi R Wray. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, 44(3):247–250, February 2012.

[5] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, Michael C O'Donovan, Benjamin M Neale, Nick Patterson, and Alkes L Price. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Publishing Group*, 47(12):1385–1392, November 2015.

[6] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James T R Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, Tune H Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A Bacanu, Martin Begemann, Richard A Belliveau Jr, Judit Bene, Sarah E Bergen, Elizabeth Bevilacqua, Tim B Bigdeli, Donald W Black, Richard Bruggeman, Nancy G Buccola, Randy L Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M Cantor, Vaughan J Carr, Noa Carrera, Stanley V Catts, Kimberly D Chambert, Raymond C K Chan, Ronald Y L Chen, Eric Y H Chen, Wei Cheng, Eric F C Cheung, Siow Ann Chong, C Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J Crowley, David Curtis, Michael Davidson, Kenneth L Davis, Franziska Degenhardt, Jurgen Del Favero, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H Fanous, Martilias S Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B Freimer, Marion Friedl, Joseph I Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Ina Giegling, Paola Giusti-Rodríguez, Stephanie Godard, Jacqueline I Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Lieuwe de Haan, Christian Hammer, Marian L Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M Hartmann, Frans A Henskens, Stefan Herms, Joel N Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V Hollegaard, David M Hougaard, Masashi Ikeda, Inge Joa, Antonio Julià, René S Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C Keller, James L Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K Kähler, Claudine Laurent, Jimmy Lee Chee Keong, S Hong Lee, Sophie E Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M Loughland, Jan Lubinski, Jouko Lönnqvist, Milan Macek Jr, Patrik K E Magnusson, Brion S Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W McCarley, Colm McDonald, Andrew M McIntosh, Sandra Meier, Carin J Meijer, Bela Melegh, Ingrid Melle, Raquelle I Mesholam-Gately, Andres Metspalu, Patricia T Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W Morris, Ole Mors, Kieran C Murphy, Robin M Murray, Inez Myin-Germeys, Bertram Müller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A Nertney, Gerald Nestadt, Kristin K Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Psychosis Endophenotypes International Consortium, Christos Pantelis, George N Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T Pato, Tiina

Paunio, Milica Pejovic-Milovancevic, Diana O Perkins, Olli Pietiläinen, Jonathan Pimm, Andrew J Pocklington, John Powell, Alkes Price, Ann E Pulver, Shaun M Purcell, Digby Quested, Henrik B Rasmussen, Abraham Reichenberg, Mark A Reimers, Alexander L Richards, Joshua L Roffman, Panos Roussos, Douglas M Ruderfer, Veikko Salomaa, Alan R Sanders, Ulrich Schall, Christian R Schubert, Thomas G Schulze, Sibylle G Schwab, Edward M Scolnick, Rodney J Scott, Larry J Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M Silverman, Kang Sim, Petr Slominsky, Jordan W Smoller, Hon-Cheong So, ChrisC A Spencer, Eli A Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E Straub, Eric Strengman, Jana Strohmaier, T Scott Stroup, and ... Subramaniam. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, July 2014.

[7] A L Richards, G Leonenko, J T Walters, D H Kavanagh, E G Rees, A Evans, K D Chambert, J L Moran, J Goldstein, B M Neale, S A McCarroll, A J Pocklington, P A Holmans, M J Owen, and M C O'Donovan. Exome arrays capture polygenic rare variant contributions to schizophrenia. *Human Molecular Genetics*, 25(5):1001–1007, February 2016.

[8] Giulio Genovese, Menachem Fromer, Eli A Stahl, Douglas M Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L Moran, Shaun M Purcell, Pamela Sklar, Patrick F Sullivan, Christina M Hultman, and Steven A McCarroll. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience*, 19(11):1433–1441, October 2016.

[9] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, Anna Kähler, Laramie Duncan, Eli Stahl, Giulio Genovese, Esperanza Fernández, Mark O Collins, Noboru H Komiyama, Jyoti S Choudhary, Patrik K E Magnusson, Eric Banks, Khalid Shakir, Kiran Garimella, Tim Fennell, Mark DePristo, Seth G N Grant, Stephen J Haggarty, Stacey Gabriel, Edward M Scolnick, Eric S Lander, Christina M Hultman, Patrick F Sullivan, Steven A McCarroll, and Pamela Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190, February 2014.

[10] Toby Johnson and Nick Barton. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1411–1425, July 2005.

[11] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Publishing Group*, 48(7):709–717, May 2016.

[12] Peter M Visscher and Jian Yang. A plethora of pleiotropy across complex traits. *Nature Publishing Group*, 48(7):707–708, July 2016.

[13] Hunter B Fraser. Gene expression drives local adaptation in humans. *Genome Research*, 23(7):1089–1096, July 2013.

[14] J J Berg and G Coop. A population genetic signal of polygenic adaptation. *PLOS Genetics*, 2014.

[15] E Corona, R Chen, M Sikora, A A Morgan, and C J Patel. Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. *PLoS …*, 2013.

[16] R Chen, E Corona, M Sikora, J T Dudley, and A A Morgan. Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLoS …*, 2012.

[17] Q Ayub, L Moutsianas, and Y Chen. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *The American Journal of …*, 2014.

[18] Renato Polimanti and Joel Gelernter. Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. *PLOS Genetics*, 13(2):e1006618–14, February 2017.

[19] James V Neel. Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"? *American journal of human genetics*, 14(4):353–362, December 1962.

[20] J V Neel. The “thrifty genotype” in 1998. *Nutrition reviews*, 1999.

[21] M Franco, U Bilal, P Ordunez, M Benet, A Morejon, B Caballero, J F Kennelly, and R S Cooper. Population-wide weight loss and regain in relation to diabetes burden and cardiovascular mortality in Cuba 1980-2010: repeated cross sectional surveys and ecological comparison of secular trends. *BMJ*, 346(apr09 2):f1515–f1515, April 2013.

[22] G Gibson and G Wagner. Canalization in evolutionary genetics: a stabilizing theory? *BioEssays*, 2000.

[23] G Gibson. Decanalization and the origin of complex disease. *Nature Reviews Genetics*, 10(2):134–140, 2009.

[24] J K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 2001.

[25] J K Pritchard and N J Cox. The allelic architecture of human disease genes: common disease–common variant… or not? *Human Molecular Genetics*, 2002.

[26] B Bulik-Sullivan, H K Finucane, V Anttila, and A Gusev. An atlas of genetic correlations across human diseases and traits. *Nature*, 2015.

[27] A Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. In *Proceedings of the National Academy of …*, 2010.

[28] Vineeta Agarwala, Jason Flannick, Shamil Sunyaev, and David Altshuler. Evaluating empirical bounds on complex disease genetic architecture. *Nature Publishing Group*, 45(12):1418–1427, October 2013.

[29] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, Manuel A Rivas, John R B Perry, Xueling Sim, Thomas W Blackwell, Neil R Robertson, N William Rayner, Pablo Cingolani, Adam E Locke, Juan Fernandez Tajes, Heather M Highland, Josee Dupuis, Peter S Chines, Cecilia M Lindgren, Christopher Hartl, Anne U Jackson, Han Chen, Jeroen R Huyghe, Martijn van de Bunt, Richard D Pearson, Ashish Kumar, Martina Müller-Nurasyid, Niels Grarup, Heather M Stringham, Eric R Gamazon, Jaehoon Lee, Yuhui Chen, Robert A Scott, Jennifer E Below, Peng Chen, Jinyan Huang, Min Jin Go, Michael L Stitzel, Dorota Pasko, Stephen C J Parker, Tibor V Varga, Todd Green, Nicola L Beer, Aaron G Day-Williams, Teresa Ferreira, Tasha Fingerlin, Momoko Horikoshi, Cheng Hu, Iksoo Huh, Mohammad Kamran Ikram, Bong-Jo Kim, Yongkang Kim, Young Jin Kim, Min-Seok Kwon, Juyoung Lee, Selyeong Lee, Keng-Han Lin, Taylor J Maxwell, Yoshihiko Nagai, Xu Wang, Ryan P Welch, Joon Yoon, Weihua Zhang, Nir Barzilai, Benjamin F Voight, Bok-Ghee Han, Christopher P Jenkinson, Teemu Kuulasmaa, Johanna Kuusisto, Alisa Manning, Maggie C Y Ng, Nicholette D Palmer, Beverley Balkau, Alena Stancˇ áková, Hanna E Abboud, Heiner Boeing, Vilmantas Giedraitis, Dorairaj Prabhakaran, Omri Gottesman, James Scott, Jason Carey, Phoenix Kwan, George Grant, Joshua D Smith, Benjamin M Neale, Shaun Purcell, Adam S Butterworth, Joanna M M Howson, Heung Man Lee, Yingchang Lu, Soo-Heon Kwak, Wei Zhao, John Danesh, Vincent K L Lam, Kyong Soo Park, Danish Saleheen, Wing Yee So, Claudia H T Tam, Uzma Afzal, David Aguilar, Rector Arya, Tin Aung, Edmund Chan, Carmen Navarro, Ching-Yu Cheng, Domenico Palli, Adolfo Correa, Joanne E Curran, Denis Rybin, Vidya S Farook, Sharon P Fowler, Barry I Freedman, Michael Griswold, Daniel Esten Hale, Pamela J Hicks, Chiea-Chuen Khor, Satish Kumar, Benjamin Lehne, Dorothée Thuillier, Wei Yen Lim, Jianjun Liu, Yvonne T van der Schouw, Marie Loh, Solomon K Musani, Sobha Puppala, William R Scott, Loïc Yengo, Sian-Tsung Tan, Herman A Taylor, Farook Thameem, Gregory Wilson, Tien Yin Wong, Pål Rasmus Njølstad, Jonathan C Levy, Massimo Mangino, Lori L Bonnycastle, Thomas Schwarzmayr, João Fadista, Gabriela L Surdulescu, Christian Herder, Christopher J Groves, Thomas Wieland, Jette Bork-Jensen, Ivan Brandslund, Cramer Christensen, Heikki A Koistinen, Alex S F Doney, Leena Kinnunen, Tõnu Esko, Andrew J Farmer, Liisa Hakaste, Dylan Hodgkiss, Jasmina Kravic, Valeriya Lyssenko, Mette Hollensted, Marit E Jørgensen, Torben Jørgensen, Claes Ladenvall, Johanne Marie Justesen, Annemari Käräjämäki, Jennifer Kriebel, Wolfgang Rathmann, Lars Lannfelt, Torsten Lauritzen, Narisu Narisu, Allan Linneberg, Olle Melander, Lili Milani, Matt Neville, Marju Orho-Melander, Lu Qi, Qibin Qi, Michael Roden, Olov Rolandsson, Amy Swift, Anders H Rosengren, Kathleen Stirrups, Andrew R Wood, Evelin Mihailov, Christine Blancher, Mauricio O Carneiro, Jared Maguire, Ryan Poplin, Khalid Shakir, Timothy Fennell, Mark DePristo, Martin Hrabé de Angelis, Panos Deloukas, Anette P Gjesing, Goo Jun, Peter Nilsson,

Jacquelyn Murphy, Robert Onofrio, Barbara Thorand, Torben Hansen, Christa Meisinger, Frank B Hu, Bo Isomaa, Fredrik Karpe, Liming Liang, Annette Peters, Cornelia Huth, Stephen P O'Rahilly, Colin N A Palmer, Oluf Pedersen, Rainer Rauramaa, Jaakko Tuomilehto, Veikko Salomaa, Richard M Watanabe, Ann-Christine Syvänen, Richard N Bergman, Dwaipayan Bharadwaj, Erwin P Bottinger, Yoon Shin Cho, Giriraj R Chandak, Juliana C N Chan, Kee Seng Chia, Mark J Daly, Shah B Ebrahim, Claudia Langenberg, Paul Elliott, Kathleen A Jablonski, Donna M Lehman, and Weiping and... Jia. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, August 2016.

[30] A Caballero, A Tenesa, and P D Keightley. The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. *Genetics*, 2015.

[31] S Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 1934.

[32] J L Lush, W F Lamoreux, and L N Hazel. The heritability of resistance to death in the fowl. *Poultry Science*, 1948.

[33] D S Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 1965.

[34] N Risch. Linkage strategies for genetically complex traits. I. Multilocus models. *American journal of human genetics*, 46(2):222–228, February 1990.

[35] M Slatkin. Exchangeable Models of Complex Inherited Diseases. *Genetics*, 179(4):2253–2261, August 2008.

[36] Naomi R Wray and Michael E Goddard. Multi-locus models of genetic risk of disease. *Genome Medicine*, 2(2):10, February 2010.

[37] Alan Robertson. The effect of selection against extreme deviants based on deviation or on homozygosis. *Journal of Genetics*, 54(2):236–248, 1956.

[38] I I Gottesman and T D Gould. The endophenotype concept in psychiatry: etymology and strategic intentions. *American Journal of Psychiatry*, 2003.

[39] K S Kendler and M C Neale. Endophenotype: a conceptual analysis. *Molecular Psychiatry*, 15(8):789–797, February 2010.

[40] Kevin J Mitchell. What is complex about complex disorders? *Genome biology*, 13(1):237, January 2012.

[41] J A Tennessen, A W Bigham, T D O'Connor, and W Fu. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. …, 2012.

[42] Alon Keinan and Andrew G Clark. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, 336(6082):740–743, May 2012.

[43] Y B Simons, M C Turchin, J K Pritchard, and G Sella. The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, 2014.

[44] F Gao and A Keinan. High burden of private mutations due to explosive human population growth and purifying selection. *BMC genomics*, 2014.

[45] E Gazave, D Chang, A G Clark, and A Keinan. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics*, 2013.

[46] K E Lohmueller. The impact of population demography and selection on the genetic architecture of complex traits. *PLOS Genetics*, 2014.

[47] S Schiffels and R Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 2014.

[48] Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, August 2007.

[49] J K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 2014.

[50] Fernando Racimo and Joshua G Schraiber. Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. *PLOS Genetics*, 10(11):e1004697–14, November 2014.

[51] F Larribe and P Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, 2011.

[52] A J Coffman, P H Hsieh, and S Gravel. Computationally efficient composite likelihood statistics for demographic inference. *Molecular biology and …*, 2015.

[53] P C Sham and S M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 2014.

[54] J Yang, S H Lee, M E Goddard, and P M Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of …*, 2011.

[55] D Speed, N Cai, M Johnson, S Nejentsev, and D Balding. Re-evaluation of SNP heritability in complex human traits. *bioRxiv*, 2016.

[56] L Evans, R Tahmasbi, S Vrieze, G Abecasis, and S Das. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *bioRxiv*, 2017.