Significance and Background

Most common diseases (those with prevalence > 1%, including 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 genome-wide significant associations for GWAS number at most in the hundreds, statistical analyses of GWAS results indicate that many thousands of variants affect their susceptibility.[10,11]

However, we still remain unclear about why complex diseases are as common as they are, of how their genetic architecture is shaped, i.e. the number of disease predisposing mutations present in the population and the relationships between their effect sizes and frequencies. In contrast to Mendelian diseases, where we have a fairly good quantitative understanding of how the mutational target size, dominance/recessivity, selection and demography combine to determine disease prevalence and the distribution of allele frequencies,[12] 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, natural selection, and genetic drift.[13] This suggests that a given genetic disease is common if it has a large mutational target size; that variation in prevalence among diseases with similar fitness costs reflects differences in target sizes; and that variation among disease in their genetic architecture arises from differences in the distribution of mutational effect sizes. It is also often suggested that 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,[14,15] which should 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.[16,17,18] 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 relationships to other diseases and quantitative traits.[19,20,21,22,23,24] 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 likely affected the prevalence and architecture of complex diseases. The “thrifty gene” hypothesis,[25,26] for example, 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. Such effects were documented in a recent study of rapid changes in the prevalence of type 2 diabetes and obesity in response to food shortage and economic crisis in Cuba in the 1990s.[27] It has been further suggested that large changes to the environment may disrupt developmental canalization, leading to wholesale changes to the distribution of genetic effects on disease risk.[28,29] While these ideas likely apply to more to some diseases that to others (and are also not mutually exclusive), they generally lack quantitative predictions that allow 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 regard, two studies been especially influential. Pritchard[30] 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,[31] many complex diseases are known to affect fitness (refs) and the magnitude of variants’ effects on different traits are often correlated,[16,17,18] 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[32] 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 0 the model converges to Pritchard's pleiotropic limit.[30] Eyre-Walker's model has also been used as a basis for inferences based on GWAS results for type 2 diabetes[33,34] and prostate cancer[35] (see background to Aim 2). However, the relationship between effect size and fitness in this model has little justification beyond mathematical convenience. Moreover, more recent work illustrates that assuming a different relationship profoundly affects the predictions about genetic architecture.[36] 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 by which 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, an individual's risk (R) of developing disease is a monotonic function (ℓ) of an underlying (and generally unobserved) disease liability trait (Z), such that

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

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 that captures the environmental contribution to variation in liability. The canonical models used to relate genotype to disease risk in human genetics, including Wright's threshold model,[37,38,39,40] the logistic model typically employed in GWAS,[1] and Risch's multiplicative model[41] are all included in this description and correspond to different choices for the function ℓ. Previous work[42,43] suggests that the results of the model are relatively insensitive to the choice of ℓ, and I will therefore take ℓ to be the probit link function of Wright's threshold model for the remainder of this proposal; I will nonetheless explore other choices to understand what impact they might have on my 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 and infinite sites model of 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.[44] ϕ and Φ are the Gaussian pdf and cdf respectively, and arise from our choice of the theshold model for ℓ. P is the disease prevalence. In this model, an individual allele 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 be additive (co-dominant) 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), equations (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 intriguing insights. First, while disease prevalence depends on the fitness cost

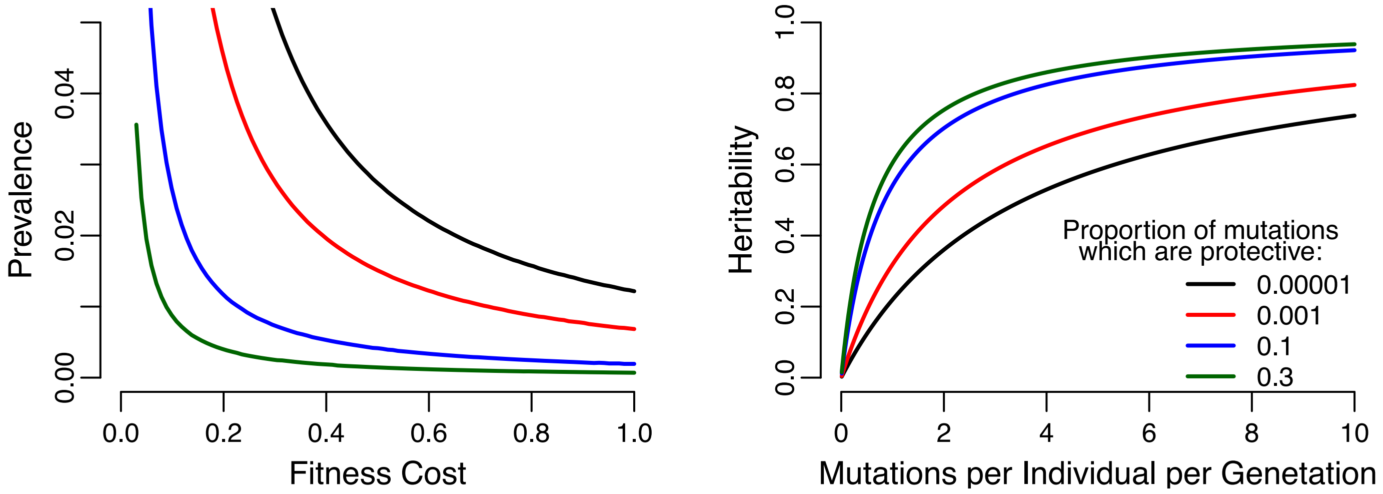

Figure 1: Solutions to the model specified by equations (2) and (3). The left panel shows the diseaes prevalence at equilibrium as a function of the fitness cost for various different choices of the strength of mutation bias. The right panel shows the heritability of disease liability as a function of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

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. It follows that bulk properties like the heritability of disease liablitiy are similarly insensitive to changes in fitness cost. While preliminary, these results already contradict common intuitions in the field, such as that alleles underlying more harmful diseases should necessarily 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: Modeling the Relationship between Population Genetic Processes and the Architecture of Complex Diseases

Building on my preliminary studies, I will extend the model to consider 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 numerical methods to obtain solutions. 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 assess their effects on human diseases (aim 2). It seems likely that one finding of my modeling work will be that certain qualitatively distinct models yield similar patterns in GWAS data, and are therefore unidentifiable. A major product of this theoretical aim will therefore be a thorough investigation of the identifiability of the evolutionary parameters which govern the genetic architecture, 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 We are particularly interested in the scenario where the change has just occurred (e.g. the effect of diet and lifestyle change on type 2 diabetes), implying negligible change in allele frequencies. The most straightforward way to model environmental change may also be the most relevant one: namely, to assume an instantaneous change to the mean or the variance of the environmental component of liability. 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 $h^2_{new} = \frac{h^2_{old}}{1+\psi}$, 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 allele frequencies and effect sizes of mutations on the liability scale would be unchanged, the effect on risk, 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, because higher prevalence implies 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 (observed in GWAS) together with estimates of the current prevalence and fitness cost might allow us to infer the size or type of environmental shift, and/or the ancestral disease prevalence.

In addition to these simple scenarios, I will also study how more complex forms of environmental change affect 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 have plausibly occurred when ancient humans migrating out of African encountered their new Eurasian environment.

Pleiotropy Mounting evidence indicates that alleles that affect the risk of a given disease often also affects other traits.[16,17,18] This suggests 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 those on the risk of developing other diseases. An allele that increases the liability for a given disease may also increase or decrease the liability for several other diseases. The net selection on that allele would therefore be a sum over the effect sizes on different 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). Under this assumption, the marginal distributions of effects on liability for each disease are identical conditional on a mutation's ultimate effects on fitness. Given a specification of parameters, this marginal distribution $(p\left(\alpha \mid s, \Theta_A\right))$ 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

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

where the distribution of frequencies, $p\left(x \mid s\right)$, can be computed from diffusion theory,[45] and the distribution of selection coefficients, $p\left(s\right)$, 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 widespread.[46,47,48] 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.[13,49] The result is that alleles which impact quantitative traits experience symmetric 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.[49]

Inclusion of pleiotropic effects on quantitative traits will therefore involve specifying the relative proportion of mutational effects on fitness which derive from disease vs those which derive 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. $p\left(h \mid s, \Theta_A\right)$). Mutations with large effects on disease liability and small effects on continuous traits would exhibit directional selection but with some dominance effect due to the reduction in fitness of heterozygotes caused by stabilizing selection. Those with small effects on diease and large effects on continuous traits would show asymmetric 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 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,[17,50] 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.[51,52,53,54]

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.[15,55,56,57,58,59] Specifically, changes in population size, such as the Out of Africa bottleneck and explosive recent growth, affect allele frequencies[55,56] 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.[15] 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,[55,60] which will later be used 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 (e.g. fitness cost, prevalence), which will allow me to simulate a single allele at a time over a dense grid of selection parameters and consider the effects on disease risk as a whole as a weighted sum over alleles with different selection parameters. My investigation of the preliminary model also suggests that disease prevalence at equilibrium is fairly sensitive to population size. This is a potentially important consideration, 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, in which case this effect will be studied and incorporated into my simulations.

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 linkage disequilibrium among variants.

Specific Aim 2: Inference of Evolutionary Parameters based on 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.[33,34,35] They employ an Approximate Bayesian Computation approach based on Eyre-Walker's model of architecture[32] (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[33,34] and prostate cancer.[35] 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[33] or the proportion of variance explained by SNPs with minor allele frequency < 5%[34,35]) between simulations and GWAS to infer 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 important and arbitrary assumption, e.g. that the distribution of selection coefficients for GWAS alleles resembles that for substitutions in proteins, despite mulitple lines of evidence that it does not.[61,62,63] 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 that rely on it challenging. I will address each of these concerns in the development of my inference approach.

Approach I will develop a composite likelihood framework for inferring the parameters that govern the evolution of complex disease. Using a likelihood approach to this problem will allow for both rigorous model comparison and parameter estimation.[64,65] The data I will rely on include paired estimates of effect sizes and allele frequencies from the loci discovered in GWAS, $\{(\alpha_i, x_i)\}$, and an estimate of the remaining genetic variance not explained by the genome-wide significant loci (V_A^*). The composite-likelihood of the model parameters, $(\Theta_A$, given these data then takes the following general form

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

where: $Pr\left(K \mid \Theta_A, \Theta_S\right)$ is the probability of observing K genome-wide significant loci, which is Poisson with mean that depends on the evolutionary parameters (Θ_A) and the GWAS parameters (Θ_S); $Pr\left(\alpha_i, x_i \mid \Theta_A\right)$ is the probability density of variants with given effect size and frequency, given by Eq (4)); $H\left(\alpha_i, x_i \mid \Theta_S\right)$ is the power to discover such a variant, which is already known;[66] and $p\left(V_A^* \mid \Theta_A, \Theta_S\right)$ gives the distribution of the remaining additive genetic variance not explained by genome-wide significant SNPs. The parameters Θ_A whose composite likelihood are being estimated depend on the specific model being considered, where, 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.

Intuitively, the number of variants discovered (K) in a GWAS gives information about the 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). The genetic variance explained by loci which do not meet genome-wide signficance (V_A^*) carries information about all of the parameters. The proposed composite likelihood approach overcomes all the aforementioned shortcomings of earlier inferences attempts. Most notably: 1) rather than assuming a poorly justified distribution of selection coefficients, I will infer it directly using a non-parametric approach, and 2) by relying on models that are more principled rather than ad-hoc, my inferences will have much clearer biological interpretation.

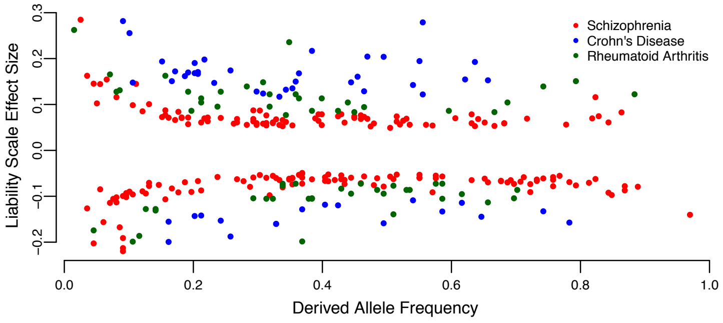

Figure 2: The relationship between allele frequency and effect size at known GWAS loci for three different complex diseases. Differences among diseases reflect differences in the underlying parameters which govern their evolution (see text for further detail).

The approach is also easily ammenable to extension. The first extension I will pursue involves modifying the term describing the genetic variance accounted for by loci not reaching genome-wide significant SNPs ($p\left(V_A^* \mid \Theta_A, \Theta_S\right)$), to account for how this genetic variance is distributed among loci with differing minor allele frequencies (i.e. $p\left(V_A^*\left(x\right) \mid \Theta_A, \Theta_S\right)$). While not as informative as directly observed genome wide significant loci, it provides consierably richer information than the unexplained variance alone, (as, for example, diseases with more variance contributed by low frequency alleles will have architectures dominated by larger selection coefficients). There are multiple methods available which have the capability to estimate how genetic variance is distributed among allele frequency bins,[4,67,68,69] all of which rely in some way on comparing the pattern of phenotypic resemblance among individuals with pattern of genetic relatedness for alleles in different frequency classes. As part of the proposed work, I will explore which available method is best suited to incorporate into my approach.

This aim will produce the first statistical framework for inferring the parameters underlying complex genetic disease architecture and prevalence that is grounded in quantitative and population genetic principles. I will apply this framework to GWAS data from at least 10 complex disease, where I will compare models to ascertain which factors play a role in which disease and estimate the salient parameters for that disease under the best model. An open access software package will be made freely available allowing other researchers to apply the inference to their datasets, and to improve the method by incorporating future theoretical advances.