Significance and Background

A large proportion of common diseases (i.e. prevalence > (XX) %) are genetically complex. In contrast to Mendelian diseases, genetic causation of complex diseases is not straightforward, with genetic risk for any given disease spread across a large number of genes, each individually of relatively minor effect. While successful efforts to map the genetics of Mendelian diseases date well back into the 20^{th} century, only in the past decade with the maturation of the genome wide association study (GWAS) approach has some understanding of the genetic basis of complex disease begun to emerge.[1,?] It's become clear now that estimates of heritability based on the twin study design were broadly accurate (if slightly overestimated in some cases), and that the heritabilities of common complex diseases are generally high. GWAS and related approaches have also shown that a substantial proportion of the variance in risk can be attributed to a very large number of common alleles,[2,3,4,5] while sequencing and exome studies indicate a role for rare variants of large effect as well.[6,7,8] The totality of the evidence therefore suggests that thousands or perhaps even tens of thousands of individual genetic variants play a role in determining susceptibility to any given complex disease. Less clear however is our understanding of the forces which govern the prevalence and genetic architecture (i.e. the relationship between allele frequency and effect size) of complex disease. In contrast to Mendelian diseases, where nearly a century of theory explain how mutation rate, dominance, selection[9] and demography[?] combine to influence the prevalence of these diseases, there is relatively little in the way of quantitative theory on the evolution of complex disease.

Nonetheless, several qualitative mechanisms have been proposed which may potentially explain the observed patterns. Perhaps the most straightforward is that the present distribution of genetic risk reflects a simple balance between mutation and selection.[10] The prevalence of common disease may therefore be ascribed to relatively large mutational target sizes for many diseases. Variation among diseases in their prevalence and genetic architectures may simply reflect difference in features such as the mutational target size, distribution of effects, and fitness cost of the disease. The direct selection experienced by a given allele is likely only part of the story, however, as the extreme polygenicity of complex diseases indicate that pleiotropy must almost surely be extensive.[11,12] Indirect selection due to these pleiotropic effects may therefore play a major role in determining architecture and prevalence, both in steady-state scenarios, where pleiotropic effects may alter the stregth of selection for or against disease alleles or under more dynamic hypotheses, where recent positive selection on a trait genetically correlated to disease may have altered prevalence and/or architecture, as has been argued in a number of cases with conflicting evidence.[13,14,15,16,17,18] Alternative explanations invoke the recent and profound changes in the human environment (diet, lifestyle, etc.) and the possibility that the mismatch between the ancestral human environment and present conditions that have resulted in a large increase in disease prevalence.[19,20] Such was the original basis for the "thrifty genotype" hypothesis,[21,22] and subsequent observation of large scale oscilations in type 2 diabetes incidence in response to food shortage and economic crisis and subsequent recovery in Cuba in the 1990s[23] indicate that this effect can be profound indeed. While some of these ideas likely apply to some diseases (and they are also not mutually exclusive), more often than not they have been put forward as verbal models only, and so it is difficult to quantitatively check their predictions.

Each of these hypotheses are ultimately statements about population and quantitative genetic processes. While some models relating population genetic processes with the kind of data emerging from GWAS have been proposed, they generally rely on various *ad hoc* or limiting assumptions that make easily interpretable inference difficult. Nonetheless, two studies in particular have been especially influential. Pritchard, in 2001,[24] considered a model in which the effect on disease risk is completely uncoupled from their fitness cost, i.e. an extreme pleiotropy limit. While this paper was enormously influential in grounding the debate surrounding the common disease-common variant hypothesis in population genetic theory,[25] it seems unrealistic that a mutation's effect on disease risk should not have any consequences for its fitness. A second study which has been influential is that of Eyre-Walker in 2010,[26] who posited that all mutations are deleterious *a priori*, but arbitrarily assumed a relationship between effect size and selection coefficient of the form $\alpha = \delta S^{\tau}(1+\epsilon)$, where S is the selection coefficient, δ is a randomly chosen sign (i.e. 1 or -1), ϵ is a noise term, and τ is a "coupling parameter" meant to capture the effect of pleiotropy: where $\tau = 1$ means no pleiotropy and $\tau = 0$ gives Pritchard's[24] extreme pleiotropy limit. While this model has seen empirical application (see Specific Aim 2 below for more background on the Eyre-Walker model in this context), it does not have any obvious theoretical justification, and indeed possesses some odd features, such as fitenss equivalence of mutations with opposing effects on disease. While these studies have both been extremely influential, it seems worth reemphasizing the fact that neither posesses any concept of fitness surface relating an explicit disease phenotype to natural selection, and in light of our rapidly accumulating knowledge of the genetic architecture of complex disease, a fresh take on the problem seems due.

Here I propose to develope generative models for the way that genetic architecture and disease prevalence will be affected by evolutionary parameters such as mutation, natural selection, pleiotropy and demography. I will develop these models into statistical inference approaches which take advantage of the rich information present in GWAS data to infer the underlying parameters which govern the evolution of complex disease genetic architecture.

Approach

Preliminary Results Our simplest model and point of departure considers the impact of mutation and selection on a single disease in a constant environment. In this model, each individual's risk of developing disease is a non-linear transform of an underlying (and generally unobserved) disease liability trait, such that

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

where an individual's liability for disease (Z) is an additive trait, α_i and g_i are 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). An individual's probability of developing disease (i.e. their "risk": R) is a monotonic function of their liability. This form covers a range of standard models for the genetics of binary traits, including Wright's liabity threshold model,[?,?,?,?] the logistic model commonly employed in GWAS,[?] and the exponential of Risch's multiplicative model,[29] among other possibilities (the difference among these is simply in the choice of the monotonic function ℓ). Previous work[27,28] suggests that the exact choice of ℓ is unlikely to be particularly important, and our preliminary investigations (omitted due to space constraints) support this conclusion. As such I will take ℓ to be the probit link function of Wright's liability threshold model for the remainder of this proposal, though I will explore other choices to understand what impact if any they might have on our results.

The number, frequencies and effect sizes of the sites contributing to variation in liability arise from population genetics processes. In our model, liability increasing and decreasing mutations arise according to an infinite sites model with free recombination among sites at genome wide rates μ^+ and μ^-, with effect size distributions $f^+(\alpha)$ and $f^-(\alpha)$ respectively. Individuals with the disease have a reduced fitness of $1 - S$ (disease free individuals have fitness 1).

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. The steady state is reached when

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

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 final term accounts for the selection pressure toward lower liability; V_A is the additive genetic variance of liability, P is the disease prevalence, and ϕ and Φ are the Gaussian pdf and cdf repsectively. The compound term multiplying V_A is a selection gradient in the standard quantitative genetics sense, and is a simple generalization of the gradient exerted by truncation selection to cases where the "truncated" individuals have fitness greater than zero.[?] An individual allele in this model with effect size α on disease liability will experience a selection coefficient

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

against the liability increasing homozygote, and will evolve under fitness additivity so long as α is small.

Equations (2) and (3) provide a path to solving the model, as the appearance of S and P in both equations couple together the behavior of the population at the macroscopic level with the microscopic dynamics of individual alleles frequencies. While the mutation rates, effect size distributions, and fitness cost of the disease are biological inputs, disease prevalence (P) and genetic variance (V_A) are dependent variables which evolve as part of the system. The additive genetic variance, as a product of the genetic architecture, depends on how the frequencies of individual alleles evolve, which depends on the individual selection coefficients they experience.

We can then apply results from diffusion theory on the frequency spectrum of an allele conditional on its selection coefficient[30,31,?] to find the genetic architecture and disease prevalence at equilibrium.

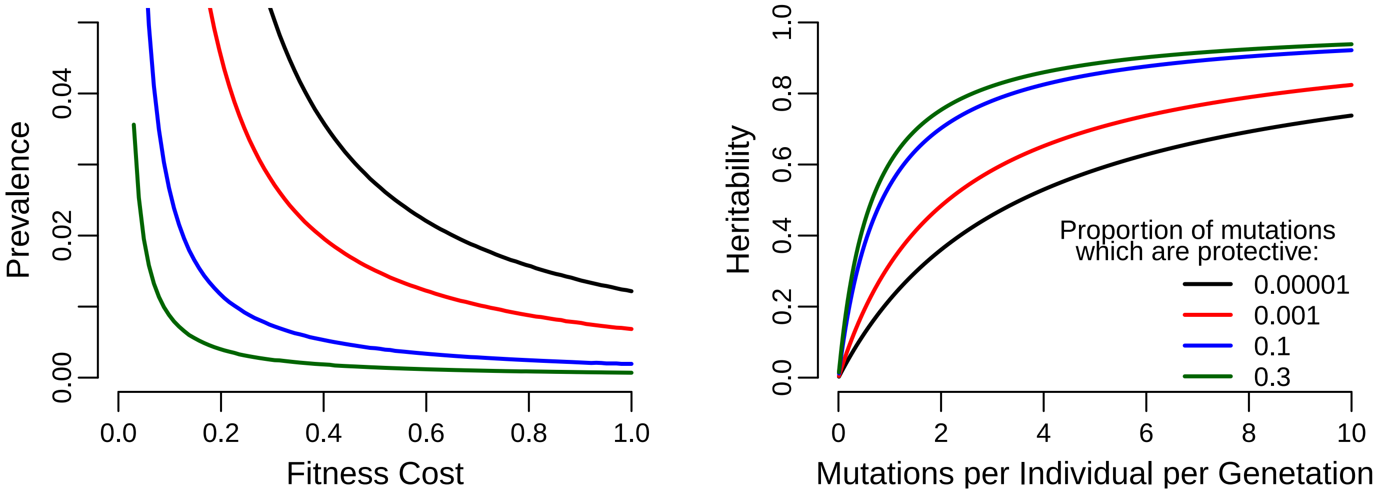

Figure 1: Caption goes here

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

I will generalize the basic model from our preliminary results to include the impacts of pleiotropy and environmental and demographic change. In each case, I will solve these generalized models to obtain analytical expressions for the genetic architecture an disease prevalence. In each case, my results will be checked by comparisson to simulations, and in cases where analytical solutions are not tractable, results will be pursued directly via numerical and simulation based methods.

Environmental Change

The most straightforward way to incorporate environmental change may also be the most relevant to human disease. The simplest environmental change models are those in which the mean or the variance of the environmental component of the phenotype shifts or increases suddenly. Given the recent and rapid changes to human environments (e.g. diet and lifestyle on type 2 diabetes prevalence), we are most interested in the scenario where an environmental shift has just occurred, 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 enironmental mean, the effect is simply an increase 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).

The result of a change in the environmental variance are slightly more complex, as it impacts both the prevalence and the heritability of the disease. The most straightforward impact is on heritability. If the environmental variance is increased by an amount ψ, then heritability is decreased such that $h^2_{g,new} = \frac{h^2_{g,old}}{1+\psi}$ (where again, ψ is given here in units of the pre-change phenotypic variance). Prevalence, on the other hand, will increase to $P_{new} = 1 - \Phi\left(\frac{\Phi^{-1}(1-P_{old})}{1+\psi}\right)$. In both of these simple environmental change scenarios, the a given mutation's effect on liability is unchanged. The effect on risk, however, is increased, as a given mutation's contribution to risk conditional on its contribution to liability also depends on the prevalence of the disease (expressions omitted due to space). The result is that the while the additive genetic variance for liability is unchanged under either scenario, we expect that on the *risk* scale it will be increased. A shift in the mean of the environmental contribution should actually cause an increase in heritability on the risk scale, while the impact of an increase in the variance of the environmental contribution on the heritability of risk are not immediately obvious, and may be architecture dependent.

In addition to these simple scenarios, I will also use simulations to study how less recent changes to the environment (e.g. following the Out of Africa event) might have impacted genetic architecture.

(large shift only part of the population: bimodality)

(large shift only part of the population: bimodality)

(large shift only part of the population: bimodality)

Pleiotropy As noted in the background above, mounting evidence suggests that genetic variation is often highly pleiotropic. We will consider pleiotropic effects of two kinds. The first includes pleiotropic effects on other disease traits. Pleiotropy of this form will generally act to modulate the additive selection coefficient felt by a given allele. Liability increasing mutations which have protective effects on other diseases will have less strongly negative selection coefficients than in the preliminary model (or may even be selected for), while those which increase liability for multiple diseases will incur additional selective cost beyond that due to direct selection on the focal disease.

The second form of pleiotropy considered will be due to stabilizing selection on continuously distributed (i.e. non-disease) quantitative traits. In models of stabilizing selection on quantitative traits at equilibrium, there is no directional component to the selection felt by individual alleles. Rather, due to variance reducing selection for individuals to cluster near the optimum, individual alleles experience symetric underdominance with respect to fitness (i.e. the minor allele is always selected against), where the strength of selection depends on the effect size of the mutation on the quantitative trait relative to the strength of stabilizing selection.[32]

This observation suggests an interesting relationship which will form the foundation for our modeling of pleiotropy. When a mutation effects only disease, selection is directional and additive, whereas when a mutation effects only quantitative traits, selection is underdominant. Mutations which have large effects on disease and smaller impacts on quantitative traits will experience directional selection with the disease causing mutation being partially dominant for fitness, while mutations which have large effects on quantitative traits and smaller effects on disease will be underdominant for fitness, but asymetrically so, with the liability increasing homozygote being less fit than the liability decreasing homozygote.

The major objective here is to derive an expression for joint distribution of effect size and allele frequency (i.e. the genetic architecture) as a function of the parameters (Θ_A) which describe the nature of pleiotropy impacting the disease architecture. These two quantities are conditionally independent of one another given the specification of the selection coefficient(s), which suggests a tractable approach for theoretical analysis and inference (aim 2). The expression for the genetic architecture can be written as an integral over the selection coefficients

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

The expression for the distribution of allele frequencies ($p\left(x \mid s, h\right)$) can be computed from standard diffusion theory,[?] and this fact can be leveraged in an inference context (aim 2) to learn the distribution of the selection coeficients ($p\left(s, h\right)$) directly from the data (alternately, plausible distribution can be specified in a theoretical context to understand how differences in the distribution of selection coeffients impact architecture). This leaves specifying the distribution of effect sizes for a given set of selection coefficients and under a given set of parameters governing the effects of pleiotropy ($p\left(\alpha \mid s, h, \Theta_A\right)$) as the primary theoretical task.

I will approach this problem first by considering simple isotropic models of pleiotropy in which mutations may affect more than one disease/trait, but there is no inherent mutational covariance among diseases or traits. In particular, I will investigate how the number of pleiotropically related diseases/traits, the strength of selection on them, and the degree of functional overlap among diseases/traits all impact the genetic architecture of disease.

Non-Equilibrium Demography It is clear from population genetic work over the last decade and a half that demographic events such as the Out of Africa bottleneck and recent explosive population growth have had a significant impact on allele frequencies and therefore potentially on genetic architecture. Recent work from Dr Sella's lab,[33] among others (cite others) suggests that both the bottleneck and recent growth may have impacted genetic architecture, though the relative importance of each depends on the (as yet unknown) selection coefficients of disease alleles. However, what most work to date on this problem has been done in the context of a single site in a vaccum with a fixed selection coefficient, with no explicit disease phenotype, and therefore no model of the relationship between effect size and selection coefficient.

I will use simulation based approaches to study how the particular course of human demographic history has impacted the evolution of complex disease architecture and prevalence. Population size changes impact architecture through their modulation of the relationship between genetic drift and natural selection. All else being equal, genetic architectures in larger populations will be composed of more rare alleles, due to the increased efficiency of selection, though the precise impact depends on the details of the model. 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.

Additional Complications (linkage-disequilibrium)
(linkage-disequilibrium)
(large deviations)
(a few more lines)
(a few more lines)

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

Background The existing work which comes closest to that proposed in this aim is recent work applying the Eyre-Walker model (discussed in intro above) to investigate the genetic architecture of complex disease. One such line of work, which includes two studies from David Altshuler's group and colleagues, simulates disease architectures under the Eyre-Walker model, and then compares the number causal type 2 diabetes variants discovered by a given study design to that which would be expected under the Eyre-Walker model simulations, with a particular focus on inferring the total mutational target size and the pleiotropy parameter τ, which captures the coupling between fitness and effect size. These studies used the number of genome wide significant variants[34, 35] as summary statistics in an approximate Bayesian computation (ABC) approach. The initial study,[34] which simply used the total number of variants discovered was able to eliminate the pleiotropic extreme (i.e. the Pritchard model) and a model in which fitness and effect are tightly coupled, but could say little beyond that. A more recent application[35] used the number of significant variants with minor allele frequency (MAF) < 5%, as a summary statistic, and was able to show that relatively small values of τ in the range of 0.1 (i.e. relatively weak coupling between effect size and fitness) are most consistent with the genetic architecture of type 2 diabetes.

Another study in this vain is that of Mancuso et al (2015),[?] who used the Eyre-Walker model to investigate the genetic architecture of prostate cancer susceptibility. Rather than using signficant hits as a summary statistic, these authors used variance partitioning methods[36] to estimate the proportion of the total genetic variance that is attributable to rare alleles ($0.01\% >$ MAF $> 1\%$), and then took this quantity as the ABC summary statistc. The result was support for moderate values of of $\tau \approx 0.5$.

The results of these studies highlight a couple of important facts. One is the difficulty of interpretting the meaning of the τ parameter in the Eyre-Walker model. While it is clear that lower values of τ represent a weaker coupling between fitness and effect size, it remains unclear how to intperpret this parameter biologically, as it was not derived from any explicit model of the impact of pleiotropic traits on the focal trait. The Eyre-Walker model is also fundamental not extensible to inference with multiple traits together, which we expect to particularly important in the future as multi-trait GWAS methods continue to mature.[?, ?]

One encouraging takeaway is the fact that inference approaches which use information about the frequencies of alleles discovered in GWAS[?, 35] obtained more precise parameter estimates than the one that did not,[34] but it is worth stating that all of these studies leave substantial amounts of the information contained in GWAS data unused, indicating that there should be substantial room for improvement. Finally, it should be noted that all of these studies rely on the assumption that the distribution of selection coefficients experienced by disease associated alleles resembles the gamma distribution often assumed for non-synonymous mutations[37] (though perhaps with little real justification[38]), rather than estimating the distribution of selection coefficients from the data directly. It is therefore clear that there is a need for an inference approach which A) uses the totality of the evidence available in GWAS data, and B) allows for the estimation of parameters which have clear biological interpretations, which is the focus of this aim.

Approach I will develop an inference approach which leverages GWAS data to infer the parameters of the models developed in Aim 1 for a number of disease GWAS datasets. The approach is based on the Poisson Random Field and related composite likelihood techniques which have been used extensively in population genetic inference. Given 1) the K genome wide significant variants discovered in a GWAS, 2) estimates of their effect sizes and allele frequencies (α_i, x_i), 3) an estimate of the heritable variation ($V^*_{G,b}$) attributable to each of B minor allele frequency bins, and 4) the parameters of the study Θ_S (i.e. the number of cases and controls and the disease prevalence assumed for the GWAS), the likelihood of the evolutionary parameters underlying the architecture (Θ_A) can be written as

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

The first term gives the probability of observing K genome wide significant variants associated with the disease (which is Poisson distributed). The second term gives the probability density of a variant with a given effect size and frequency, while the third term gives the power to detect such a variant as a function of the study parameters. The final term gives the probability that a given proportion of the heritability not captured by the K genome wide significant variants is apportioned to a given minor allele frequency bin. The first, second, and fourth terms all depend on the evolutionary parameters (Θ_A) and therefore are the major link which connects Aim 1 and Aim 2, as the the theory from Aim 1 will be used to compute these expressions.

Intuitively, the number of genome wide significant variants mostly provides information about the mutational target size, while their frequencies are informative about the strength of selection they experience, and the relationship between frequency and effect size is therefore informative about the nature of the pleiotropic effects of disease associated loci.

The proportion of the "missing heritability" attributable to different minor allele frequency bins essentially contains information about the mean strength of selection experienced by disease variants, and therefore helps constrain the range of possible models which can fit the data. It is a less rich source of information than the individual genome-wide significant variants themselves, as we do not get to observe informative features such as whether it is the derived or ancestral allele which increases disease risk, and we must fold the frequency spectrum, which in particular discards information about protective mutations.[31] Nevertheless, the theory from Aim 1 will make predictions about how variance is distributed across minor allele frequencies (hopefully add a figure here to show how two different choices of evolutionary parameters lead to different distributions of variance among bins) , which means it can be used for inference. Intuitively, the stronger seletion on disease variants is, the larger the proportion of variance we expect to be explained by low frequency bins.

The likelihood above is a composite rather than a true likelihood, meaning that it does not account for the non-indepdendence (i.e. linkage disequilibrium) among sites. Such approaches are commonplace in population genetics when full likelihood computation is infeasible. Composite likelihoods have the property that they are unbiased with respect to the maximum likelihood estimate, but understate the uncertainty about that estimate, precisely because non-independence among datapoints is ignored, leading to the appearance of stronger evidence than actually exists, and naive approaches to model comparison therefore erroneously tend to favor more complex models (Gao and Song 2010) . This limitation can be overcome by bootstrapping or potentially by adapting recently developed methods from the literature on demographic inference which show promise in circumventing these issues analytically at significantly reduced computational cost.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Anand P Patil, Rosalind E Howes, Oscar A Nyangiri, Peter W Gething, Thomas N Williams, David J Weatherall, Fr eacute d eacute ric B Piel, and Simon I Hay. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature Communications*, 1(8):1–7, October 2010.

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

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

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

[22] J V Neel. The "thrifty genotype" in 1998. *Nutrition reviews*, 1999.

[23] 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.

[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] 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.

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

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

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

[30] S A Sawyer and D L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, December 1992.

[31] C D Bustamante, J Wakeley, S Sawyer, and D L Hartl. Directional selection and the site-frequency spectrum. *Genetics*, 2001.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.