Computational challenges in genome wide association studies: data processing, variant annotation and epistasis

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

CHAPTER 1
Epistatic GWAS analysis

1.1 Preface

[Empty: We will write the preface after the paper finalized]

1.1.1 Type II diabetes

Although this thesis focusses on the development of computational approaches that could be applied to the study of a number of complex diseases, our focus has been on type II diabetes mellitus (T2D), a complex disease first described by the Egyptians in 1500 BCE. Later the Greeks in 230 BCE used the term "diabetes" meaning "pass through" (or "siphon") denoting the constant thirst and frequent urination of the patients. In the 1700s the term "mellitus" (from honey) was added to denote that the urine was sweet and would "attracts ants".

Diabetes symptoms include frequent urination, thirst, and constant hunger, high blood sugar (hyperglycemia) and insulin resistance. Long term complication from T2D may include eyesight problems, heart disease, strokes and kidney failure. Type II diabetes, is highly correlated with obesity and disease rate has increased dramatically during the last 50 years. According to the World Health Organisation the prevalence of diabetes is 9% in adults and an estimated 1.5 millions deaths were caused by diabetes in 2012 [16], which is predicted to be the 7th leading cause of death by 2030. The costs associated to treating diabetes patients only in the U.S. are estimated around $245 billion dollars.

In recent years, over 80 genetic loci related to T2D have been identified [30, 9]. Nevertheless, the overall effect sizes of these loci account for less than 10% of the overall disease predisposition [26]. This poses the question of why, given that so much efforts has been directed at finding the genetic components of this disease, the loci found so far have such modest effects. This lack of large genetic effects do not only arise in T2D but also in almost all complex traits and could be explained by what is known as the "missing heritability" problem.

A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies P. Cingolani, R.Sladek, M. Blanchette

A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies P. Cingolani, R.Sladek, M. Blanchette

1.2 Abstract

Motivation: Since proteins perform their functions mainly by interacting with other proteins, epistasis genome wide association studies are assumed to be key in unveiling genetic risk of disease. Due to their high complexity and sometimes prohibitive computational requirements, epistatic GWAS have often been neglected. In this paper, we propose a novel methodology for analysing putative epistatic interactions by combining multiple genome alignments and sequencing information.

Results: We propose a methodology to perform genome wide association on pairs of variants whose putative epistatic interaction might have a phenotypic effect. Using Pdb information and genome wide multiple species alignment we create a statistical coevolutionary model that increases priors of putative interacting sites, these priors are then used as basis for genome wide association in a Bayesian framework. Our optimized algorithms can be applied

to genome wide scale sequencing studies for tens of thousands of samples, that typically yield millions of variants.

1.3 Introduction

Genetic studies aim to discover how a phenotype of interest, such as disease risk or height, is affected by genetic background. Genome wide association studies (GWAS) are powerful techniques aimed to find statistical association between phenotype and genetic variants (mutations or polymorphisms) [7]. Although several genetic variants related to different phenotypes have been found, variants discovered in GWAS so far can only explain a small part for the phenotypic inheritance. For instance, all genetic variants associated to height collectively account for few centimetres in the offspring's height [42], similarly all combined variants related to type 2 diabetes risk only explain 5% to 10% of the overall variance in disease predisposition [30, 9]. This problem is known as missing heritability" [26] and recent theories suggest that genetic interactions (epistasis) might play an important role in it [47, 48].

The foundations for epistasis, broadly defined as genetic interaction" [14], have been proposed almost a hundred years ago by Bateson (1909) and Fisher (1918). It was the latter who coined the term to denote a statistical deviation of multilocus genotype values from an additive linear model for the value of a phenotype" [14]. There is evidence of such interactions being involved in complex diseases. For instance an interaction between BACE1 and APOE4 having a significant association with Alzheimer's disease has consistently been replicated in different studies [8]. One of the main problems in finding association between interactions and disease is that out of the whole set of molecular interactions (the interactome) only a small part of it has been characterized [39].

Interacting proteins can be identified experimentally through several types of approaches (Yeast-Two-Hybrid, Protein fragment complementation assay, Glutathione-s-transferase, Affinity purification coupled to mass spectrometry, Tandem affinity purification, etc. [35]) and large databases of protein-protein interactions are now available for human [37, 35]. Nevertheless, these methods predict the interaction between proteins and may not discern the exact residues mediating such interactions Furthermore, it is estimated that up to 80% large of the human protein-protein interactions remains unknown [39].

These issues can be addressed using computational predictions of either pairs of interacting proteins or interacting residues [36]. A type of approaches that has been gaining popularity recently is one that makes use of the plethora of genomic sequences available for species other than human in order to discover evolutionary evidence of selective pressure on individual residues to identify interacting sites and interfaces [27]. Interacting residues and their neighbors may then be subject to compensating epistasis, where a mutation at a residue in one protein may be compensated by another mutation at a residue in the second protein [31]. Assuming that evolutionary pressure acts on both interaction sites simultaneously, co-occurring compensatory mutations can become fixed in the population with higher probability than non-compensatory ones. In light of this hypothesis, we can use statistical methods on multiple sequence alignments (MSA) of proteins from different organisms to find coevolving sites. This types of approaches has been used to identify coevolving sites both within a protein (e.g. N-terminal and C-terminal domains in PKG protein [15] or the GroES-L chaperoning system [34], α and β haemoglobin subunits [31]), and between interacting proteins (e.g. G-protein coupled receptors and protein ligands [15]).

Many methods exist to find putative interaction loci, both within and across proteins, based on evolutionary evidence (see [11] for a review). One of the simplest methods for inferring co-evolution uses mutual information (MI) between two loci [27] in a multiple sequence alignment ($\mathcal{M}_{sa}$). However, methods based on correlation or mutual information are known to have systematic biases due to phylogeny [11], or sequence heterogeneity problems [41]. More sophisticated methods, such as DCA [29], PSICOV [18] or mdMI [6] try to overcome these biases, however they are usually not suitable for GWAS-scale analysis for two main reasons: i) they require multiple alignments of a very large number of sequences (from 400 to $25L$, where L is the length of the protein [6]), such depth is not usually available at whole genome scale; and ii) they are computationally demanding (e.g. running for minutes or even days for each interacting pair of proteins being considered), making them unsuitable for analyses involving millions of variants spanning over thousands of proteins. Furthermore, a recent study shows that overall agreement between methods is not high (65

Applying epistatic interaction models to GWAS studies is a challenging problem for several reasons: i) interaction models are by definition non-linear [14]; ii) analyzing all order N variant combinations in a sequencing experiment requires great computational power and efficient algorithms because the number tests grows exponentially with N [32]; iii) multiple testing correction can render association tests underpowered for all but very large cohorts [14, 32]; and iv) there is no consensus of what genetic interaction means [25], which is reflected in the difficulty to find a unified model [32, 25]. For all these reasons and due to the lack of sequencing cohorts large enough to detect these interactions, the application of epistatic models to sequencing studies has not been widespread. Furthermore, there is no clear consensus on the required

sample size to detect epistatic interactions. Depending on phenotypic effect size and variant's allele frequency some estimates assume in the order of 10,000 to 500,000 cases [19] to be required. Such cohorts are now becoming feasible due to improvements and cost reductions in sequencing technology.

Approaches for epistatic GWAS do exist and they apply a wide array of methodologies, although we don't intend write a comprehensive review, we mention a few examples. In [46] the authors infer probabilities by noting that interactions create linkage disequilibrium patterns in the disease population. A Bayesian partitioning model" is applied in [45]. In [1] the authors look for over / under-represented allele pairs in a given population by performing an analysis of imbalanced allele pair frequencies. Finally, finding interacting variants can be viewed as an optimisation / attribute selection procedure, thus many machine learning methodologies have been proposed [28].

In this work we propose methodologies to prioritize pairs of variants from sequencing experiments by combining genome wide association with epistatic interaction models. In a nutshell, our method uses recently computed 100-way vertebrate genome alignments to calculate interaction posterior probabilities for any given pair of residues in human proteins. This is achieved by contrasting the likelihood of the observed pair of alignment columns under a joint substitution model that models dependencies between interacting sites, and a null model of independent evolution. These posterior probabilities are then used as priors to modulate the evidence of epistatic interaction derived from GWAS data. We engineer efficient algorithms that can be applied to GWAS-scale datasets of tens of thousands of samples.

Add T2D?

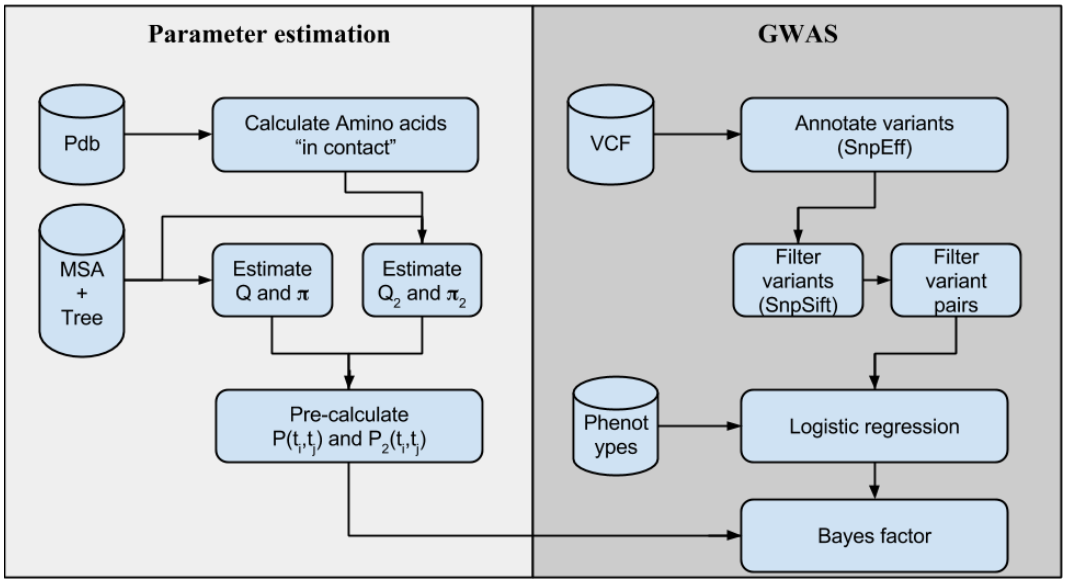

Figure 1–1: Complete pipeline example

1.4 Methods

Our epistatic GWAS analysis pipeline involves three key steps, as shown in Figure 1–1. First, we learn a co-evolutionary substitution rate matrix for pairs of amino acids that are in contact in proteins. Second, we analyze a GWAS data set to identify pairs of non-synonymous SNPs that show (possibly weak) evidence of epistasis. Third, for each pair of SNP identified in step 2, we measure the evidence of co-evolution of the pair of encoded amino acids, and combine it with the GWAS evidence by adding up the corresponding Bayes factors.

1.4.1 Substitution model for pairs of interacting amino acids

In this section, we describe how we estimate two substitution rate matrices. The first is the usual 20×20 substitution rate matrix Q describing the evolution of individual amino acids. The second, Q_2 is a 400×400 substitution rate matrix for pairs of interacting residues.

We used the 100-way vertebrate multiple sequence alignment and accompanying phylogenetic tree T available from the UCSC Genome Browser [20].

This alignment includes the DNA sequences of 100 species whose genome is completely or nearly completely sequenced, with 6 Afrotheria, 14 Avians, 14 Euarchontoglires, 16 Fish, 25 Laurasiatheria, 5 Mammalians, 12 Primates and 8 Sarcopterygii. The multiple alignment is performed using "multiz" algorithm [3, 22].

From the $\sim 21,000$ human protein structures (resolution less than 3 Angstrom) available in Protein Data Bank, we extracted a set of $\sim 770,000$ pairs of within protein amino acids interactions" residues, defined as pairs of residues from the same protein where at least one pair of atoms is within 3 Angstrom or less. Similarly, from the set of $\sim 5,700$ models of co-crystallized complexes in PDB, we extracted a set of $\sim 12,000$ pairs of protein-protein interacting" residues, defined as amino acids from different proteins that satisfy the same distance criterion.

To derive rate matrix Q, we consider the complete set of $n \sim 22M$ protein coding sites present in the alignment (that unambiguously map to the genome), irrespective of the presence or absence of contacts. Q is obtained following classical sequence evolution theory ([44] [13]). First, for each pair of species s_i and s_j, we obtain $c_i(a)$ defined as the count of amino acid a in species s_i, and $c_{i,j}(a,b)$ defined as the number of sites that have had a transition from amino acid a in s_i to b in s_j. Stationary probability of amino acid a in genome s_i is then defined as $\pi_i = c_i(a)/n$. Assuming a time reversible model, we get the frequency of change from a to b: $f_{i,j}(a,b) = f_{j,i}(a,b) = (c_{i,j}(a,b) + c_{j,i}(a,b))/(2n)$. Let $P_{i,j}$ be the amino acid transition probability matrix from s_i to s_j, i.e. $P_{i,j}(a,b)$ is the probability that s_j has amino acid b given than s_i has amino acid a. Then P is obtained through the relation $f_{i,j}(a,b) = \pi_i(a) \cdot P_{i,j}(a,b)$, or $P_{i,j}(a,b) = f_{i,j}(a,b)/\pi_i(a)$. Let $t_{i,j}$ be the total branch length between s_i and s_j (obtained from the phylogenetic tree). Assuming

time reversibility, we have $P_{i,j} = e^{Q \cdot t_{i,j}}$, and thus $Q = log[P_{i,j}/t_{i,j}]$ [44]. Taking into account the estimation error, the equation becomes $\hat{Q}(t_i + t_j) = Q = log[P_{i,j}/t_{i,j}] + \epsilon_{i,j}$, where $\epsilon_{i,j}$ is an error matrix. Under the assumption that the mean error is zero, we can approximate the rate matrix by the calculating an average of all estimates: $\hat{Q} = \frac{1}{N(N-1)/2} \sum_{i<j} \hat{Q}_(t_i + t_j) = \sum_{t_i+t_j} \frac{1}{t_i+t_j} log[\hat{P}_(t_i + t_j)]$.

The much larger substitution matrix Q_2 describes the substitution rate from any pair of amino acid (a, b) to any other pair (c, d). It is derived similarly to Q, but considering only pairs of amino acids from the set of within protein interacting pairs of amino acids. We only take into account amino acids pairs within the same chain, that are separated by 20 amino acids or more from high resolution PDB structures of human proteins.

1.4.2 Calculating likelihood of individual and pairs of alignment columns

Given a substitution rate matrix Q, the likelihood $L_1(i)$ of an alignment column i assigning an amino acid to each leaf in the tree T is calculated using the well known Felsenstein algorithm [13]. This is achieved in time $O(N \cdot |\Sigma|^2)$, where $|\Sigma| = 20$. Given matrix Q_2, the same algorithm can be used to compute the likelihood $L_2(i, j)$ of a pair of alignment columns (i, j), but now in time $O(N \cdot |\Sigma|^4)$.

A test for co-evolution of two positions i, j of the same or different proteins is obtained using the likelihood ratio under the two models:

$$Lr_{MSA}(i, j) = \frac{P(i, j|M_{SA}, Q_2)}{P(a_i|M_{SA}, Q) \cdot P(a_j|M_{SA}, Q)}$$

where the denominator assumes that the amino acids i and j evolve independently.

Given a pair of columns in the multiple sequence alignment $\mathcal{M}_{sa}$, the corresponding phylegentic tree $\mathcal{T}$, we calculate the likelihood ratio by propagating the probabilities from the leaves up the root of the phylogenetic tree using Felsestein's algorithm [13] using transition matrices Q and Q_2 for the null and alternative models respectively.

Because the calculations described in this section will need to be performed on a very large number of pairs of sites, optimizations we required to ensure manageable running time. First, pre-calculation of matrix exponentials $P(t) = e^{Qt}$ is necessary for all values of t corresponding to individual branch lengths. Another optimization (constant-tree caching") is to cache likelihood values for subtrees of the phylogenetic tree where all nodes have the same amino acid values. This optimization results in speedup only if the phylogenetic tree remains constant throughout the genome.

1.4.3 GWAS model

Given N_S samples (individuals), we use the standard notation for phenotypes and code them as $d_s = 1$ meaning that the individual s is affected by disease and $d_s = 0$ if the individual is "healthy". Let $\bar{d} = [d_1, ..., d_{N_s}]$ be a phenotype vector and $g_{s,i} \in \{0, 1, 2\}$ a genomic variant for sample s, loci i (we assume there are N_V variants). A logistic model of disease risk [2] is

$$\begin{aligned} p_{s,i} &= P(d_s|g_{s,i}) \\ &= \phi(\theta_0 + \theta_1 g_{s,i} + \theta_2 c_{s,1} + \theta_4 c_{s,2} + ...) \\ &= \frac{1}{1 + e^{\theta_0 + \theta_1 g_{s,i} + \theta_2 c_{s,1} + \theta_4 c_{s,2} + ...}} \\ &= \phi(\bar{\theta}^T \bar{g}_{s,i}) \end{aligned}$$

where $\phi(\cdot)$ is the sigmoid function, $c_{s,1}, c_{s,2}, ...$ are covariates for each individual s (these covariates usually include sex, age and eigenvalues from

population structure analysis [33]), $\bar{g}_{s,i} = [1, g_{s,i}, c_{s,1}, c_{s,2}, ..., c_{s,N_C}]$, and $\bar{\theta} = [\theta_1, \theta_2, ..., \theta_m]$. The parameter estimates $\bar{\theta}$ are obtained by solving the maximum likelihood equation

$$\begin{aligned} L(\bar{\theta}) &= \prod_{s=1}^{N_S} P(d_s|\bar{\theta}, g_{s,i}) \\ &= \prod_{s=1}^{N_S} p_{s,i}^{d_s}(1-p_{s,i})^{1-d_s} \\ &= \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_{s,i})^{d_s}(1-\phi(\bar{\theta}^T \bar{g}_{s,i}))^{1-d_s} \end{aligned}$$

where $p_{s,i} = P(d_s|\bar{\theta}, g_{s,i})$ is the probability of individual s disease outcome, given a genomic variant.

Using this model, we have two hypotheses: i) the null hypothesis, H_0, assumes that genotype does not influence disease probability (i.e. $\theta_1 = 0$). ii) the alternate hypothesis, H_1, assumes that the genotype does influences disease probability. We can compare these two hypothesis using a likelihood ratio test $\Delta = L(\bar{\theta}_{alt}|H_1)/L(\bar{\theta}_{null}|H_0)$, where $\bar{\theta}_{null}$ and $\bar{\theta}_{alt}$ are the maximum likelihood estimates for null and alt model respectively. According to Wilk's theorem, the log likelihood ratio has a χ_1^2 distribution, so we can easily calculate a p-value.

Next, we extend the logistic model to accommodate interacting loci. For an individual (sample s), we model interactions between two genetic loci i and j, having genotypes $g_{s,i}$ and $g_{s,j}$, by extending the logistic model with N_{cov} covariates $c_{s,j}$ as

$$\begin{aligned} P(d_s|g_{s,i}, g_{s,j}, H_1) &= \phi[\theta_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3(g_{s,i}g_{s,j}) + \theta_4 c_{s,1} + ... + \theta_m c_{s,N_{cov}}] \\ &= \phi(\bar{\theta}^T \bar{g}_{s,i,j})) \end{aligned}$$

where $\bar{g}_{s,i,j} = [1, g_{s,i}, g_{s,j}, (g_{s,i}g_{s,j}), c_{s,1}, c_{s,2}, ..., c_{s,N_{cov}}]^T$. An implicit assumption in this equation is that $g_{s,i}$ and $g_{s,j}$ are not correlated (e.g. they don't sit in the same LD-Block). This can be enforced either by using haplotype structure information (e.g. from HapMap) or by limiting the application of the model to variants either in different chromosomes or over 1M bases apart. The null hypothesis H_0 assumes that variants act independently

$$P(d_s|g_{s,i}, g_{s,j}, H_0) = \phi[\theta'_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3 c_{s,1} + ...] = \phi(\bar{\theta}^T \bar{g}'_{s,i,j})$$

where $\bar{g}'_{s,i,j} = [1, g_{s,i}, g_{s,j}, c_{s,1}, c_{s,2}, ..., c_{s,N_{cov}}]^T$. Since these models are not nested (i.e. H_0 is not included in H_1), we cannot directly apply Wilks theorem.

Logistic regression was implemented using several different algorithms to reach the desired performance. The fastest convergence is obtained using Iterative Reweighted Least Squares (IRWLS [10]) and BroydenFletcherGoldfarbShanno algorithm (BFGS [4]) with some code optimizations. In most cases IRWLS converges faster, so it was selected as the default implementation in our analysis.

Another way to compare the null hypothesis to the alternative hypothesis, is using a Bayesian formulation [21, 40]

$$\begin{aligned} P(H_1|\mathcal{D}) &= \frac{P(\mathcal{D}|H_1)P(H_1)}{P(\mathcal{D})} = \frac{\int P(\mathcal{D}|\theta, H_1)P(\theta|H_1)d\theta}{P(\mathcal{D})} \\ \Rightarrow \frac{P(H_1|D)}{P(H_0|D)} &= \frac{\int P(\mathcal{D}|\theta, H_1)P(\theta|H_1)d\theta}{\int P(\mathcal{D}|\theta, H_0)P(\theta|H_0)d\theta} \frac{P(H_1)}{P(H_0)} = B_F \frac{P(H_1)}{P(H_0)} \end{aligned}$$

where B_F, the ratio of the two integrals, is the Bayes factor. Using a Bayesian formulation has two main advantages: i) the hypothesis are automatically corrected for model complexity since Bayes factor asymptotically converge to Bayesian Information Criteria (BIC) is implicitly within Bayes

Factors [21], and ii) we can compare non-nested models. The Bayes factor for the epistatic model becomes:

$$B_F = \frac{\int \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_s)^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_s)]^{1-d_s} P(\bar{\theta}|H_1) d\theta}{\iint \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_{s,i}) \phi(\bar{\theta'}^T \bar{g}_{s,j}))^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_{s,i}) \phi(\bar{\theta'}^T \bar{g}_{s,j})]^{1-d_s} P(\bar{\theta}|H_0) P(\bar{\theta'}|H_0) d\theta d\theta'} \quad (1.1)$$

Unfortunately, calculating Bayes factors is not trivial and most of the times there are no closed form equations. Calculating the integrals using numerical algorithms is possible, but imposes a significant computational burden thus making it impractical for large datasets, such as GWAS data, even using large computing clusters. We can approximate the integrals using Laplace's method [21]. If $g(x)$ has a minimum at x_0, it can be shown that

$$\int e^{\lambda g(x)} h(x) dx \quad \simeq \quad h(x_0) e^{\lambda g(x_0)} \sqrt{\frac{2\pi}{\lambda g''(x_0)}}$$

The multivariate case, for $\bar{x} \in \Re^d$, is analogous, we just need a Hessian matrix instead of a second derivate of $g(\cdot)$

$$\int e^{\lambda g(\bar{x})} h(\bar{x}) d\bar{x} \quad \simeq \quad h(\bar{x}_0) e^{\lambda g(\bar{x}_0)} \left(\frac{2\pi}{\lambda}\right)^{d/2} \left[\frac{\partial^2 g(\bar{x})}{\partial \bar{x} \partial \bar{x}^T}\right]^{-1/2} \quad (1.2)$$

Using equation 1.2 we can try to approximate the complex integrals in equation 1.1 by the transformation $L(\bar{\theta}) = e^{\ell(\bar{\theta})}$, where $\ell(\cdot)$ is the log-likelihood of the data. So, we can use Laplace approximation by using Eq.1.2, at the point of the maximum likelihood. In order to do so, we need to calculate the Hessian matrix in Eq.1.2. Fortunately, for logistic models, we can make a few simplifications. Considering that $L(\bar{\theta}) = \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_s)^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_s)]^{1-d_i}$, it can be shown that for genotype terms

$$\frac{\partial^2 \ell(\bar{\theta})}{\partial \theta_i \partial \theta_j} = \sum_s g_{s,i} g_{s,j} p_s (1 - p_s)$$

Using analogous derivation for the covariates, we can find an analytic form of the Hessian, which completes the Laplace approximation formula.

1.4.4 Computational and statistical issues

It is easy to see that the computational burden for detection of pairs of interacting genetic loci affecting disease risk is significantly larger than in a standard (single variant) GWAS study. A priori all pairs of variants should are analyzed thus significantly increasing the number of statistical tests. This also reduces statistical power since the required p-value significance level would be orders of magnitude smaller. A naive approach would estimate that if a typical genetic sequencing study has 10^6 variants, a GWAS on epistatic variants would square that number of statistical tests, thus p-values required for significance would be in the order of $0.05/(10^6)^2 = 5 \cdot 10^{-14}$.

Fortunately these numbers can be reduced significantly. First, in this study, we only concentrate on non-synonymous coding variants. Second, if two variants for which the interaction term ($g_{s,i} g_{s,j}$) is zero in all samples, which usually happens for pairs of rare variants, there is no need to calculate the logistic regression since in the epistatic model has a trivial solution by setting θ_3 to zero, thus we can skip these variant pairs. Third, if the variants and the epistatic term $[g_{s,i}, g_{s,j}, g_{s,i} g_{s,j}]$ are linearly dependent, the logistic regression result will be meaningless, so we can safely skip such variant pairs. Fourth, if one of the variants has high allele frequency respect to the other, all non-zero epistatic terms may lie in the same positions as non-zero genotypes from the low frequency variant, causing logistic regression estimates to artificially inflate the coefficients of the low frequency variant and the epistatic term thus

creating an artificially high association (low p-value). So we filter out these variant pairs as well. Finally, we filter out all variants having Hardy-Weinberg p-value of less than 10^-6, since these variants also artificially inflate the logistic regression coefficients.

After all filters have been applied, using a model GWAS data which involves over 10,000 individuals and over 1 million variants, the number of variant pairs to be analyzed is less than 100 millions, as opposed to the naive estimation of 10^{12} variant pairs. By means of the z-score relationship [**?**], we set the GWAS significance threshold for 50 million pairs at $log_{10}[BF] = 8.0$.

Calculating Bayes factors involves using prior parameter distributions. In order to estimate the distributions, we run the logistic regression fitting analysis and plot the parameter distributions for different levels of significance. As expected most parameters have unimodal distribution, except for θ_3, which has a multimodal distribution (Figure **??**). For all parameters, except θ_3, we use a normal distribution centered at the mean and variance was set to at least one ($\sigma = 1$) even though most times the variance is much smaller (this is done to avoid penalizing outliers too heavily and to have smooth derivatives near the maximum likelihood estimates). For θ_3, which has a multimodal distribution, we fit mixed model parameters using an EM algorithm on high significant term ($BF_{raw} \geq 3$), as shown in Supplementary Figure / Table **??**.

1.4.5 Putting it all together

In summary, we first calculate the transitions matrices for the Markov models (Q and Q_2) based on observations from protein structures (Pdb) and multiple sequence alignments (UCSC's 100-way). We analyze variants from genome sequencing data first by filtering only for non-synonymous variants, then analyzing all possible pairs of variants and filtering out those that are unsuitable for further analysis (e.g. in linear dependence, Hardy Weinberg

p-value less than 10^-6, etc.). From the pairs of variants that pass filtering, we fit two logistic regression models (null and alternative hypothesis), then calculate a p-value using the log-likelihood ratio, and keeping pairs of variants having p-values below a predefined threshold (10^-6). These pairs of variants are then analyzed under our co-evolutionary model, we find the corresponding columns in the multiple sequence alignment and calculate the likelihoods for the null and alternative models by means of Felsenstein's algorithm (using matrices Q and Q_2 in respectively). Finally, likelihoods from co-evolutionary model and likelihoods from logistic regression models are incorporated into a Bayes Factor equation, which is calculated using Laplace's approximation.

1.5 Results

Our approach, which is summarized in Figure ??, involves three main components. First we estimate evolutionary substitution rates for individual amino acids in a protein as well as for pairs of amino acids (either from the same protein or not) that are physically interacting. Given a set of multiple sequence alignment of protein sequences, these evolutionary models can be used to calculate the likelihood of interaction between any two given amino acids, without the need for any structural information. Second, a statistical test for epistasis is developed to identify pairs of non-synonymous SNPs that show (often weak) evidence of interaction in the way they associate to a given trait. Finally, information from the co-evolution component is combined with that from the epistasis component to give more power to the epistasis test.

1.5.1 Co-evolutionary substitution models

The approach described in Methods was used to obtain substitution rate matrix Q for individual amino acids and Q_2 for pairs of physically interacting

residues within the same protein. Unsurprisingly, Q (or more precisely a transition matrix $P(t)$ obtained from Q) is very similar to well known transitions matrixes such as PAM [] (Supplementary Figure **??** and Table **??**).

Estimating Q_2 requires information about amino acids that are known to be interacting". A pair of amino acids is considered to be interacting" if any pair of atoms (one from each amino acid) has a distance of 3 Angstrom or less [5].

The structure of Q_2, which describes substitution rates between one pair of interacting amino acids to another, is richer (Supplementary Figure **??** and supplementary file **??**). Of particular interest are the pairs of pairs of amino acids for which the ratio $R(ab, cd) = Q_2(ab, cd)/Q(a, c) \cdot Q(b, d)$ is large. Those substitution pairs are the ones that are most strongly indicative of an interaction. Figure **??** shows that the number of pairs for which R deviates significantly from 1 is quite large, arguing that interacting sites have co-evolutionary rates that differ from the bulk of non-interacting sites.

For example, the case with the highest rate ratio is `[V.I -> W.W]`' (i.e. amino acid 'V' switched to 'W' in the one sequence, and amino acid 'I' changed to 'W' in the other). In fact, the top 10 pairs are transitions to 'W.W' amino acid pairs. This makes sense considering (i) individual amino acid substitution rates to tryptophan are generally very low, but that (ii) tryptophan pairs are well known β-hairpin stabilizers and are considered as a paradigm for designing stable β-hairpins [**?**].

Another type of pair transitions with large ratio is the double transitions to a pair of phenylalanine amino acids from a pairs of hydrophobic amino acids (Lysine, Asparagine, Glutamine, Arginine, Aspartic acid and Glutamic acid). Phenylalanine-Phenylalanine interaction pairs are assumed to conform $\pi - \pi$

interactions which are predicted and experimentally observed to be favorable [17].

1.5.2 Co-evolutionary model validation

We first the ability of our co-evolutionary model to detect interacting sites located within the same protein by computing the likelihood ratio of the evolutionary history of a candidate pair of sites under an co-evolutionary model (Q_2) versus under independence (Q). Although such pairs of sites are unlikely to exhibit evidence of epistasis in GWAS studies (due to linkage), accurate prediction of interacting sites in a given protein are useful for many other purposes, such as protein structure prediction and prediction of the impact of individual mutations. Figure **??** shows interacting sites tend to have higher likelihood ratio scores than non-interacting ones. Although the likelihood ratio score it itself cannot perfectly discriminate between the two classes, only 25.9

To confirm that an evolutionary model estimated based on pairs of interacting sites from the same protein is useful at predicting pairs of interacting sites between proteins, we repeated the same type of analysis on ?? pairs of interacting (¡3 Angstrom) and ?? pairs of non-interacting (¿30 Angstrom) residues from distinct proteins, obtained from co-crystal structures in PDB (see Methods). As seen on Figure **??**, the two classes of sites have substantially different likelihood ratio distributions (Mann-Whitney one sided test: $p - value < 2.2E^{-}16$), although slightly less so than for sites from the same protein. Only 29% of non-interacting sites have a likelihood ratio larger than the median for interacting sites. These empirical distributions, allow us to approximate of the log odds of the interacting" vs non-interacting" amino acids distributions as $log_{odds}(x) = log[P(LL(MSA|Q2) \geq x)/P(LL(MSA|Q) \geq x]$, which fits well an exponential function $e^{0.195x} - 1.018$ where $\alpha = 0.195$ and

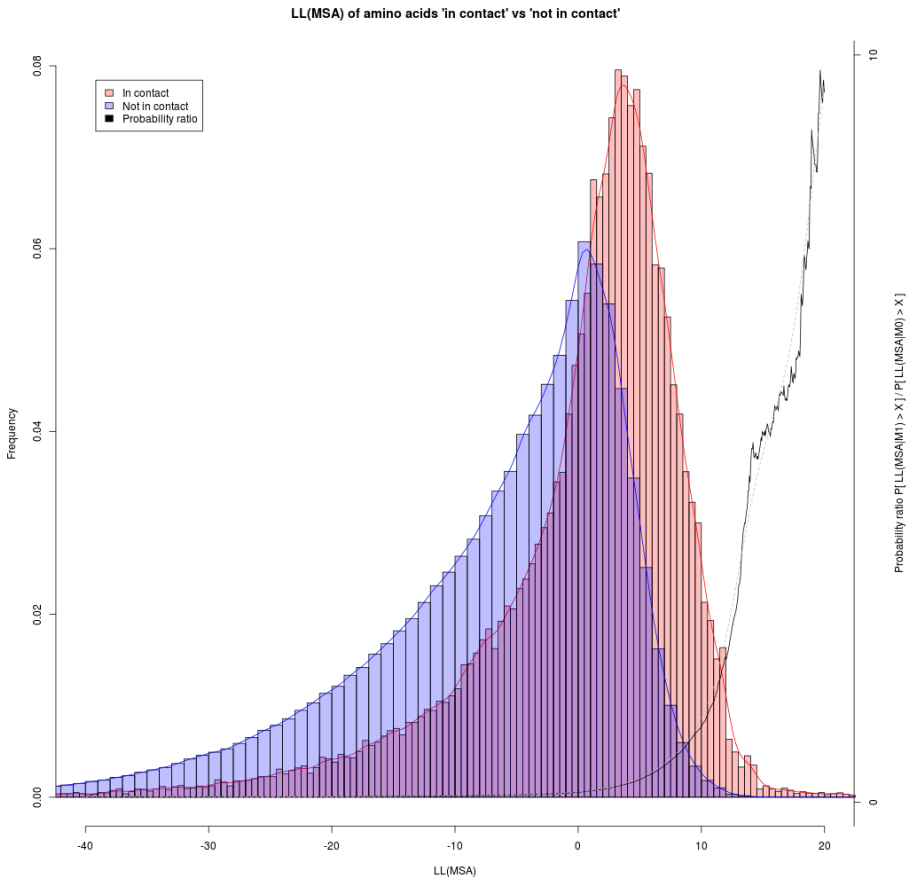

Figure 1–2: Red & Blue:Histogram of LL(MSA) "in contact" vs 'null' within amino acids within the protein (Pdb). Black: Cummulative probability ratios $R_{MSA} = P[LL(MSA|M1) > X]/P[LL(MSA|M0) > X]$. Dotted grey: Smoothed R_{MSA}

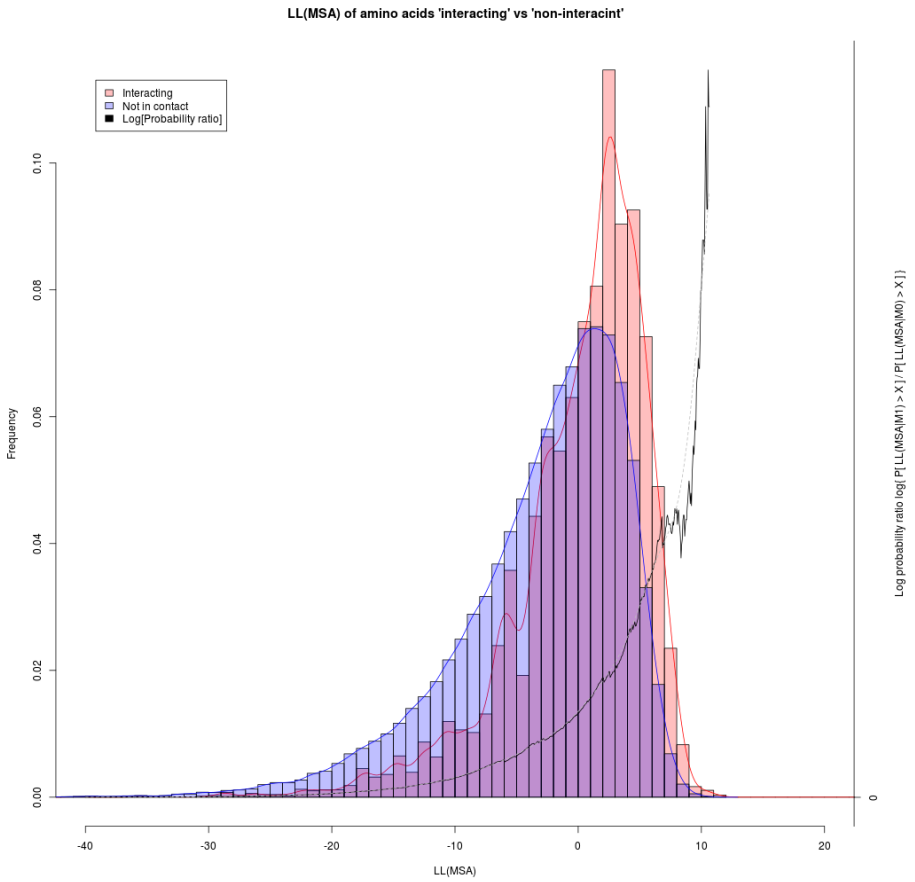

Figure 1–3: A) LL(MSA) in contact vs 'null' interacting amino acids in co-crystalized proteins (from Pdb), roughly 40% of interacting records have $LL(MSA) > 1$. B) Log odds ratio of cummulative LL(MSA) probability (interacting / non-interacting)

$\beta = 1.018$. The log odds value is capped to 4.0 to avoid artificially increasing Bayes Factors.

Figure 4 shows the example of a predicted contact $LL(MSA) = 7.7$ between SENP1 and SUMO1 proteins detected by our method. The co-crystallized structure from PDB highlights the interacting amino acids (less than 3 Angstrom apart) and the corresponding M_{SA} matched with the phylogenetic tree suggests co-evolutionary evidence.

Although our approach aims at identifying contacting residues from different proteins, it can also be used to predict interactions between proteins

as a whole. We extracted from BioGrid [37] a set of $\sim 3,000$ pairs of human proteins with evidence of interaction, and further required that both proteins belong to the same pathway (MsigDb, C2 groups [38]), and their corresponding genes are expressed in the same tissue (GTex [24], expression of 1 FPKM or more, tissues $\in$ {skeletal muscle, adipose tissue, pancreatic Islets}). We define non-interacting" pairs of proteins as those not fulfilling any of the three conditions, and chose $\sim 3,000$ such pairs randomly.

Let the two proteins considered have amino acid sequences $a_1...a_m$ and $b_1...b_n$. To obtain the prediction score for this pair of proteins, we identify the pair of length-k substrings $a_i, a_{i+1}, , a_{i+k-1}$ and $b_j, b_{j+1}, , b_{j+k-1}$ that exhibit the strongest support for parallel or antiparallel interactions, i.e. for which $\max\{\sum_{l=0}^{k-1} LL(a_{i+l}, b_{j+l}), \sum_{l=0}^{k-1} LL(a_{i+l}, b_{j+k-1-l})$ is maximized. Empirically, the value of k that seems to provide the best predictive power if $k = 3$. We calculate the average likelihood ratio of three consecutive pairs of amino acids ($avg_3[LL(MSA)]$), either in the forward or reverse directions. For a given pair of genes, we calculate the highest $avg_3[LL(MSA)]$ and pick the highest number as representative for that pair. As shown in Figure **??**), prediction accuracy is quite good ($p - value < 210^{-42}$), considering the modest amount of information considered.

1.5.3 Epistatic GWAS analysis

We run the epistatic GWAS analysis on a real dataset for Type 2 Diabetes, consisting of over 13,000 samples and 1.2 million variants (detailed results to be published in a separate paper). The complete analysis takes less than 2 days using less than 1,000 CPUs in our cluster. This shows that an epistatic GWAS analysis of large cohort sequencing data is feasible using current computational resources.

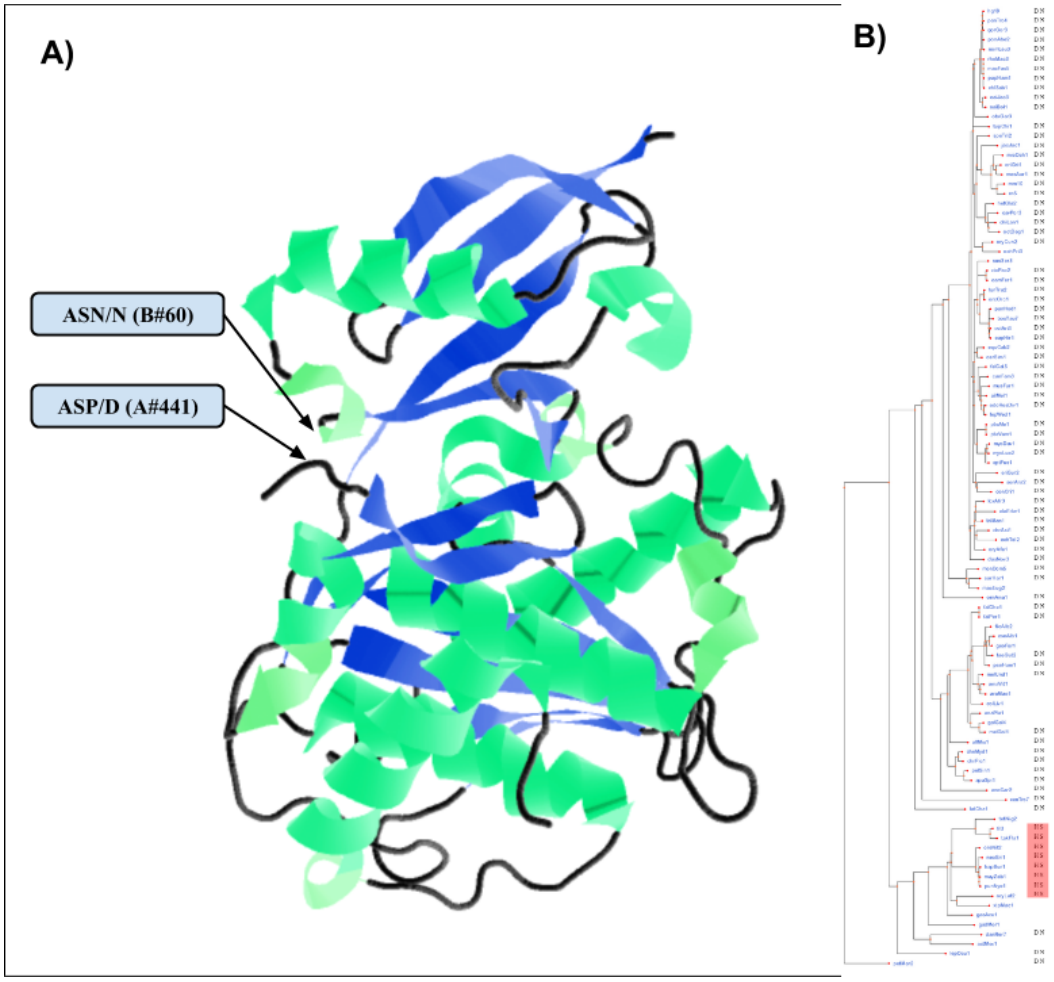

Figure 1–4: Example of amino acid interaction between SENP1 and SUMO1 proteins detected by out method $LL(MSA) = 7.7$. A) PDB structure 2G4D, shows that the amino acids are in close proximity: amino acid #441 in SENP1 (ASP/D) interacts with amino acid #60 in SUMO1 (ASN/N). B) Multiple sequence alignment and phylogenetic tree shows putative compensatory amino acid substitution pair "D-N" replaced by "H-S" (colors added for visual reference).

1.6 Discussion

In this paper, we propose a novel methodology for genome wide analysis of variants located in putative epistatic sites. Due to the large number of statistical tests required in epistatic analysis, and the corresponding reduction of statistical power, this type of analysis is meant to be applied to datasets consisting of large number of samples which can overcome the reduction in statistical power. Our analysis methods have been optimized and parallelized to be suitable for large scale sequencing genomic studies.

Our Markov epistatic model requires multiple sequence alignment and the corresponding phylogenetic tree. Both the tree and the number of sequences in the MSA should remain constant throughout the genome in order to take advantage of computational optimizations (matrix exponential precalculation and constant tree caching") that allow the algorithm to be applied at genome-wide scale. Some multiple sequence alignments (such as Pfam) usually have different number of sequences for each protein (thus different phylogenetic trees). This poses two main disadvantages for our methodology: i) we cannot benefit from the previously mentioned optimizations, since they require a constant phylogenetic tree throughout the whole genome; and ii) we would add the problem of reconciling different phylogenetic trees from two proteins, which may lead to inconsistencies. For this reasons, we selected UCSC's multi-100way [20], a genome wide multiple sequence alignment of 100 organisms, which has single genome wide phylogenetic tree.

We applied our coevolutionary model to separate clinically relevant variants from ClinVar database [23] according to their clinical significance attribute (CLNSIG). Interestingly, amino acids that categorized as "benign" or "druggable" have higher scores ($LL(MSA)$ within protein). As shown in Table **??**, benign ClinVar variants have higher mean $LL(MSA)$ distribution than

variants categorized as pathogenic (Supplementary Tables **??**, **??** and Figure **??**). We speculate that this effect might be caused because amino acids that can be compensated would be characterized as "benign" whereas deleterious amino acids changes would not be compensated by mutation.

We show the application of these methods to a large scale exome sequencing study for Type II ..

As future work, we plan to extend our method to analyze domain specific transition pairs, this would allow to obtain better estimates for known interaction domains. Another line of work is to perform GWAS using kernel based statistics of multiple variants [43] thus allowing simultaneous analysis of nearby variants in a putative interaction hotspot, in this case the epistatic information would be used as a function modifying the kernel, instead of a bayesian prior.

References

[1] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.

[2] D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[3] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004.

[4] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[5] Lukas Burger and Erik van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, 2010.

[6] Greg W Clark, Sharon H Ackerman, Elisabeth R Tillier, and Domenico L Gatti. Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. *BMC bioinformatics*, 15(1):157, 2014.

[7] G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, and K.T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[8] Onofre Combarros, Mario Cortina-Borja, A David Smith, and Donald J Lehmann. Epistasis in sporadic alzheimer's disease. *Neurobiology of aging*, 30(9):1333–1349, 2009.

[9] Diabetes SAT2D Consortium, Diabetes MAT2D Consortium, Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D Johnson, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244, 2014.

[10] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[11] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[12] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[13] Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

[14] Hong Gao, Julie M Granka, and Marcus W Feldman. On the classification of epistatic interactions. *Genetics*, 184(3):827–837, 2010.

[15] Chern-Sing Goh, Andrew A Bogan, Marcin Joachimiak, Dirk Walther, and Fred E Cohen. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2):283–293, 2000.

[16] L Guariguata, DR Whiting, I Hambleton, J Beagley, U Linnenkamp, and JE Shaw. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, 103(2):137–149, 2014.

[17] Christopher A Hunter, Juswinder Singh, and Janet M Thornton. π-π interactions: The geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *Journal of molecular biology*, 218(4):837–846, 1991.

[18] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[19] Luke Jostins, Adam P Levine, and Jeffrey C Barrett. Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments. *PloS one*, 8(10):e76328, 2013.

[20] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2014 update. *Nucleic acids research*, 42(D1):D764–D770, 2014.

[21] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[22] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[23] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, page gkt1113, 2013.

[24] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[25] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[26] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[27] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[28] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[29] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[30] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segré, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[31] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511–523, 1997.

[32] Patrick C Phillips. Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[33] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[34] Mario X Ruiz-González and Mario A Fares. Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system groes-l. *BMC evolutionary biology*, 13(1):156, 2013.

[35] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLoS computational biology*, 3(3):e42, 2007.

[36] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS computational biology*, 3(4):e43, 2007.

[37] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[38] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[39] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2009.

[40] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, 33(1):79–86, 2009.

[41] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[42] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan,

Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[43] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.

[44] Ziheng Yang. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, 2006.

[45] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[46] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[47] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[48] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.