

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

Pablo Cingolani 2015

CHAPTER 1

Epistatic GWAS analysis

1.1 Preface

In recent years, over 80 genetic loci related to T2D have been identified [113, 37]. Nevertheless, the overall effect sizes of these loci account for less than 10% of the overall disease predisposition [107]. 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, known as the “missing heritability” problem, does not only arise in T2D but also in almost all complex traits. In recent studies about missing heritability [174, 175] it was proposed that this effect might be partly explained by taking into account epistasis (i.e. gene interactions).

In this chapter, we propose a novel framework that takes into account putative epistatic interactions into genome wide association studies (GWAS).

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 [68], 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.

The rest of the chapter is published in: **P. Cingolani**, R. Sladek, M. Blanchette, “A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies”

1.2 Abstract

Motivation. Epistasis, broadly defined as genetic interactions, is one of the likely causes why genome-wide association studies (GWAS) account for a small portion of heritable disease risk. Due to their high complexity, reduced statistical power and sometimes prohibitive computational requirements, epistatic GWAS have rarely been performed.

Results. In this paper, we propose a novel methodology for analysing putative epistatic interactions by combining multiple genome alignments and sequencing information. Using protein structures for individual and co-crystallized complexes information and genome wide multiple species alignment we create a co-evolutionary model that allows the calculation of the posterior probability of physical interaction between residues given evolutionary data. These probabilities are then used as the interaction priors for an epistatic GWAS analysis as basis for genome wide Bayesian framework.

Results. Our optimized algorithms can be applied to genome wide scale sequencing studies for tens of thousands of samples, that typically yield millions of variants. We applied our approach to a large type II diabetes (T2D)

case-control cohort and inferred a number of putative interactions associated with increased risk of developing T2D.

Availability. Our code is publicly available at github.com/pcingola/Epistasis

1.3 Introduction

Genetic studies aim to discover how a phenotype of interest, such as disease risk or height, is affected by an individual’s genetic background. Genome wide association studies (GWAS) are powerful techniques aimed at finding statistical associations between a phenotype and genetic variants [33]. 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 heritability. For instance, all genetic variants associated to height collectively account for few centimetres in the offspring’s height [165]. Similarly the known variants related to type 2 diabetes risk collectively explain only 5% to 10% of the overall variance in disease predisposition [113, 37]. This problem is known as “missing heritability” [107] and recent theories suggest that genetic interactions (epistasis) might play an important role in it [174, 175].

The foundations for epistasis [59], 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” [59]. 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 [35]. Many types of situations can lead to epistatic interactions. Among them, perhaps the most common involved pairs of variants that encode amino acids whose physical interactions is regulated for their function.

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 [154]. 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. [141]) and large databases of protein-protein interactions are now available for human [145, 141]. In almost all cases, these methods identify the presence of an interaction between proteins but do not discern the exact residues mediating such interactions. Furthermore, it is estimated that up to 80% of the human protein-protein interactions remains unknown [154].

These issues can be partially addressed using computational predictions of either pairs of interacting proteins or interacting residues [142]. 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 pairs of residues to identify interacting sites and interfaces [108]. Interacting residues and their neighbours 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 [121]. For example 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, one can use statistical methods on multiple sequence alignments 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 [65], GroES-L chaperoning system [132],

α and β haemoglobin subunits [121]), and between interacting proteins (e.g. G-protein coupled receptors and protein ligands [65]).

Many methods exist to find putative interaction loci, both within and across proteins, based on evolutionary evidence (see [49] for a review). One of the simplest methods for inferring co-evolution uses mutual information between two loci [108] in a multiple sequence alignment. However, methods based on correlation or mutual information are known to have systematic biases due to the fact that they ignore phylogenetic relationships [49], or sequence heterogeneity problems [161]. More sophisticated methods, such as DCA [112], PSICOV [77] or mdMI [32] try to overcome these biases, however they are usually not suitable for GWAS-scale analysis for two main reasons. First, they require multiple alignments of a very large number of sequences (ranging from 400 to $25L$, where L is the length of the protein [32]), and such depth is not usually available at whole genome scale. Second, 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% or less) and predictive power is quite low (only 6% of the “top scoring pairs” are real interactions) [32].

Applying epistatic interaction models to GWAS studies is a challenging problem for several reasons: i) interaction models are by definition non-linear [59]; ii) analyzing all order N variant combinations requires great computational power and efficient algorithms because the number tests grows exponentially with N [122]; iii) multiple hypothesis testing correction can render association tests underpowered for all but very large cohorts [59, 122]; and iv) there is no consensus of what genetic interaction means, which is reflected in

the difficulty to find a unified model [122, 106]. 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 [80] 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. In [172], the authors infer epistatic probabilities by noting that interactions create linkage disequilibrium patterns in the disease population. A “Bayesian partitioning model” is applied in [171]. In [2], 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 attribute selection problem, thus many machine learning methodologies have been proposed [111].

In this work we propose an approach to prioritize pairs of variants identified in case/control cohorts by combining genome wide association with epistatic interaction models. In a nutshell, our method uses recently computed 100-way vertebrate genome alignments [17] 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 factors in 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. Our implementation is sufficiently efficient to be applied to GWAS-scale datasets of tens of thousands of samples. Finally we apply this

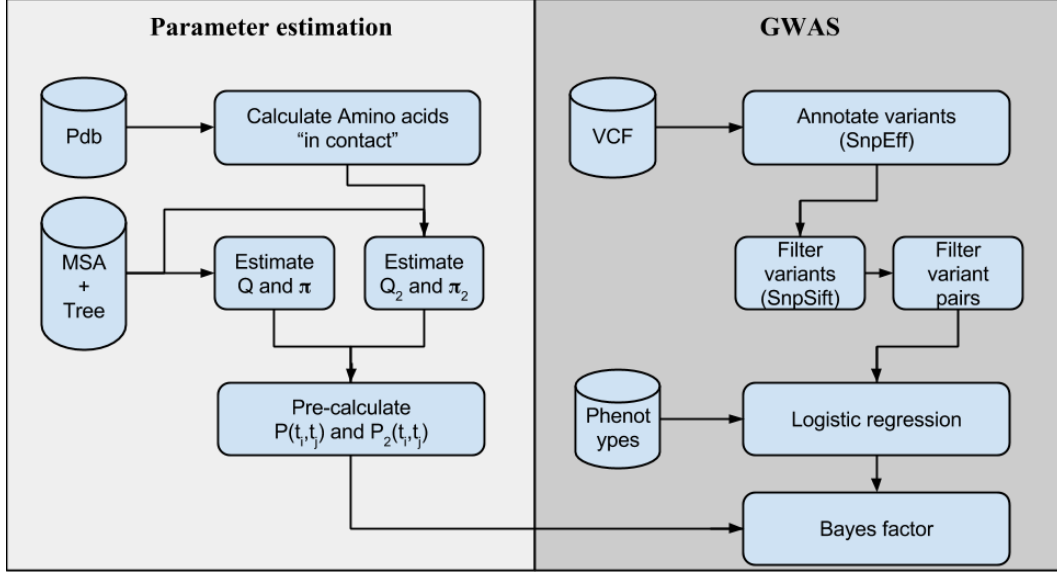


Figure 1–1: Complete pipeline example

methods to a cohort of $\sim 13,000$ individuals in a case-control study of type II diabetes (for study details, see [109]) and identify suggestive associations of putatively epistatic interactions.

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 [81]. 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 Sarcophagii. The multiple alignment is performed using “multiz” algorithm [17, 84].

From the $\sim 21,000$ human protein structures (resolution less than 3 Å) available in Protein Data Bank, we extracted a set of $\sim 770,000$ pairs of “within protein interactions” residues, defined as pairs of residues from the same protein where at least one pair of atoms is within 3 Å 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 ([169] [55]). 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 $Q = \log[P_{i,j}/t_{i,j}]$ [169]. 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:

$$\hat{Q} = \frac{1}{N(N-1)/2} \sum_{i < j} \hat{Q}(t_i + t_j) \quad (1.1)$$

$$= \frac{2}{N(N-1)} \sum_{i < j} \frac{1}{t_i + t_j} \log[\hat{P}(t_i + t_j)] \quad (1.2)$$

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 [55]. 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:

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

where the denominator assumes that the amino acids i and j evolve independently. Similarly, the log-likelihood is defined as

$$LL[M_{SA}(i, j)] = \log \left[\frac{P(i, j|M_{SA}, Q_2)}{P(a_i|M_{SA}, Q) \cdot P(a_j|M_{SA}, Q)} \right] \quad (1.3)$$

Given a pair of columns in the multiple sequence alignment M_{sa} , the corresponding phylogenetic tree \mathcal{T} , we calculate the likelihood ratio by propagating the probabilities from the leaves up the root of the phylogenetic tree using Felsenstein's algorithm [55] 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 $\bar{d} = [d_1, \dots, d_{N_S}]$ 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 [12] is

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

where $\phi(\cdot)$ is the sigmoid function, $c_{s,1}, c_{s,2}, \dots$ are covariates for each individual s (these covariates usually include sex, age and eigenvalues from population structure analysis [125]), $\bar{g}_{s,i} = [1, g_{s,i}, c_{s,1}, c_{s,2}, \dots, c_{s,N_C}]$, and $\bar{\theta} = [\theta_1, \theta_2, \dots, \theta_m]$. 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 influence disease probability. We can compare these two hypothesis using a likelihood ratio test, we define

$$LR = \frac{L(\bar{\theta}_{alt}|H_1)}{L(\bar{\theta}_{null}|H_0)} \quad (1.4)$$

$$LL = \log [LR] = \log \left[\frac{L(\bar{\theta}_{alt}|H_1)}{L(\bar{\theta}_{null}|H_0)} \right] \quad (1.5)$$

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^2_1 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

$$P(d_s|g_{s,i}, g_{s,j}, H_1) = \phi[\theta_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3 (g_{s,i} g_{s,j})] \quad (1.6)$$

$$\dots + \theta_4 c_{s,1} + \dots + \theta_m c_{s,N_{cov}}] \quad (1.7)$$

$$= \phi(\bar{\theta}^T \bar{g}_{s,i,j}) \quad (1.8)$$

where $\bar{g}_{s,i,j} = [1, g_{s,i}, g_{s,j}, (g_{s,i} g_{s,j}), c_{s,1}, c_{s,2}, \dots, c_{s,N_{cov}}]^T$. 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

$$P(d_s|g_{s,i}, g_{s,j}, H_0) = \phi[\theta'_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3 c_{s,1} + \dots] \quad (1.9)$$

$$= \phi(\bar{\theta}^T \bar{g}'_{s,i,j}) \quad (1.10)$$

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

Logistic regression parameter fitting was implemented using several different algorithms to reach the desired performance. The fastest convergence is obtained using Iterative Reweighted Least Squares (IRWLS [47]) and BroydenFletcherGoldfarbShanno algorithm (BFGS [21]) 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 [82, 156]

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

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

$$B_F = \frac{\int \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_s)^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_s)]^{1-d_s} P(\bar{\theta}|H_1) d\theta}{\iint \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_{s,i}) \phi(\bar{\theta}'^T \bar{g}_{s,j})^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_{s,i}) \phi(\bar{\theta}'^T \bar{g}_{s,j})]^{1-d_s} P(\bar{\theta}|H_0) P(\bar{\theta}'|H_0) d\theta d\theta'} \quad (1.11)$$

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 [82]. If $g(x)$ has a minimum at x_0 , it can be shown that

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

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

$$\int e^{\lambda g(\bar{x})} h(\bar{x}) d\bar{x} \simeq h(\bar{x}_0) e^{\lambda g(\bar{x}_0)} \left(\frac{2\pi}{\lambda} \right)^{d/2} \left[\frac{\partial^2 g(\bar{x})}{\partial \bar{x} \partial \bar{x}^T} \right]^{-1/2} \quad (1.12)$$

Using equation 1.12 we can try to approximate the complex integrals in equation 1.11 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.1.12, at the point of the maximum likelihood. In order to do so, we need to calculate the Hessian matrix in Eq.1.12. 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 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 [1] (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 Å or less [23].

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 [135].

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 [75].

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% of non-interacting pairs have a likelihood ratio above the median likelihood ratio of interacting pairs.

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 $\sim 3,000$ pairs of interacting ($< 3 \text{ \AA}$) and $\sim 3,000$ pairs of non-interacting ($> 30 \text{ \AA}$) 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

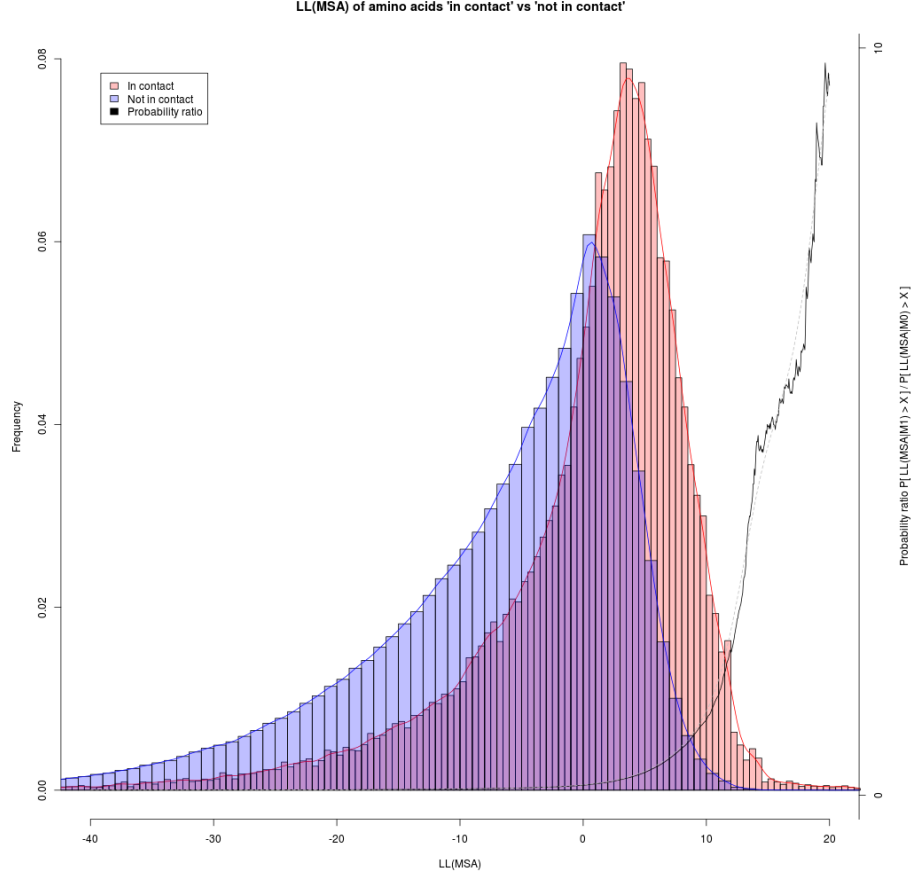


Figure 1–2: Red & Blue: Histogram of $LL(M_{SA})$ “in contact” vs ‘null’ within amino acids within the protein (PDB). Black: Cumulative probability ratios $R_{M_{SA}} = P[LL(M_{SA}|M1) > X] / P[LL(M_{SA}|M0) > X]$. Dotted grey: Smoothed $R_{M_{SA}}$

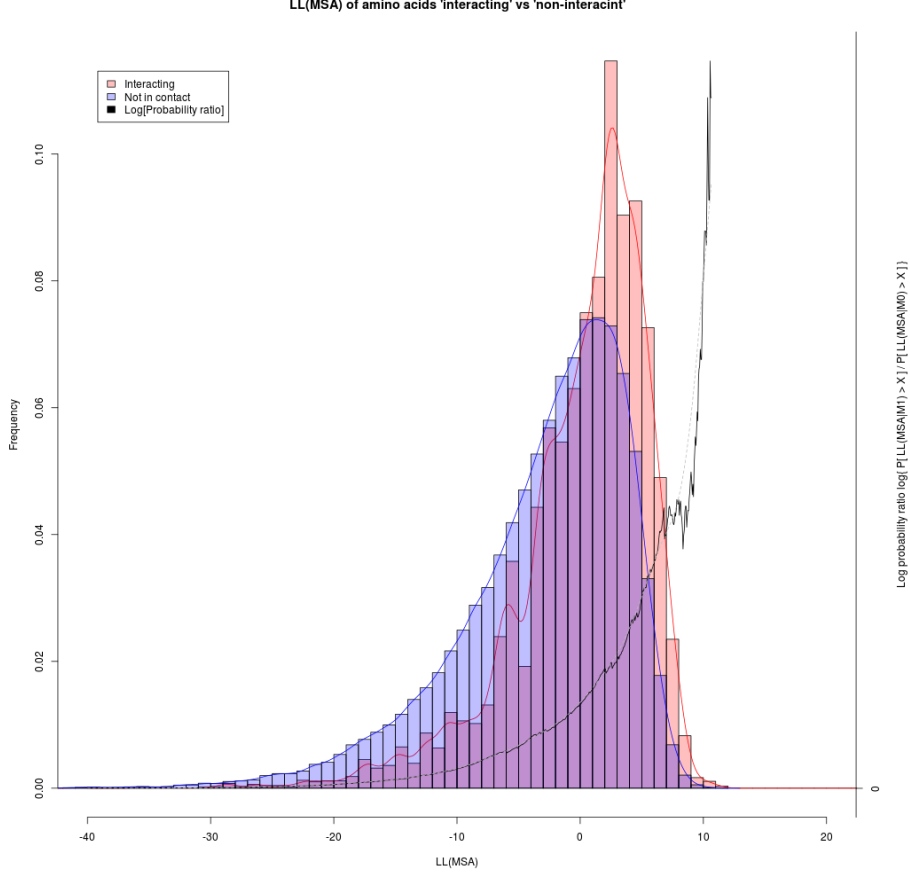


Figure 1–3: A) $LL(M_{SA})$ in contact vs 'null' interacting amino acids in co-crystallized proteins (from PDB), roughly 40% of interacting records have $LL(M_{SA}) > 1$. B) Log odds ratio of cumulative $LL(M_{SA})$ probability (interacting / non-interacting)

distributions as $\log_{odds}(x) = \log[P(LL(M_{SA}|Q2) \geq x)/P(LL(M_{SA}|Q) \geq x)]$, which fits well an exponential function $e^{0.195x} - 1.018$ where $\alpha = 0.195$ and $\beta = 1.018$. The log odds value is capped to 4.0 to avoid artificially increasing Bayes Factors.

Figure 4 shows the example of a predicted contact $LL(M_{SA}) = 7.7$ between SENP1 and SUMO1 proteins detected by our method. The co-crystallized structure from PDB highlights the interacting amino acids (less than 3 Å 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 [145] 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 [147]), and their corresponding genes are expressed in the same tissue (GTex [103], expression of 1 FPKM or more, tissues $\in \{\text{skeletal muscle, adipose tissue, pancreatic Islets}\}$). 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 $a_i, a_{i+1}, \dots, a_{i+k-1}$ and $b_j, b_{j+1}, \dots, b_{j+k-1}$ 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 is $k = 3$. We calculate the average likelihood ratio of three consecutive pairs of amino acids ($avg_3[LL(M_{SA})]$), either in the forward or reverse directions. For a given pair of genes, we calculate the highest $avg_3[LL(M_{SA})]$ and pick the highest number as representative for that pair. As shown in Figure ??), prediction accuracy is quite good (p-value $< 2 \cdot 10^{-42}$), considering the modest amount of information considered.

1.5.3 Epistatic GWAS analysis

We apply our methods to a cohort of $\sim 13,000$ individuals in a case-control study of type II diabetes [109]. This multi-ethnic study covers exons of unrelated individuals from five major ancestral groups (European descent, South Asian, East Asian, Hispanic and African American descent) using an average sequencing coverage over $80\times$ yielding $1.7M$ coding variants.

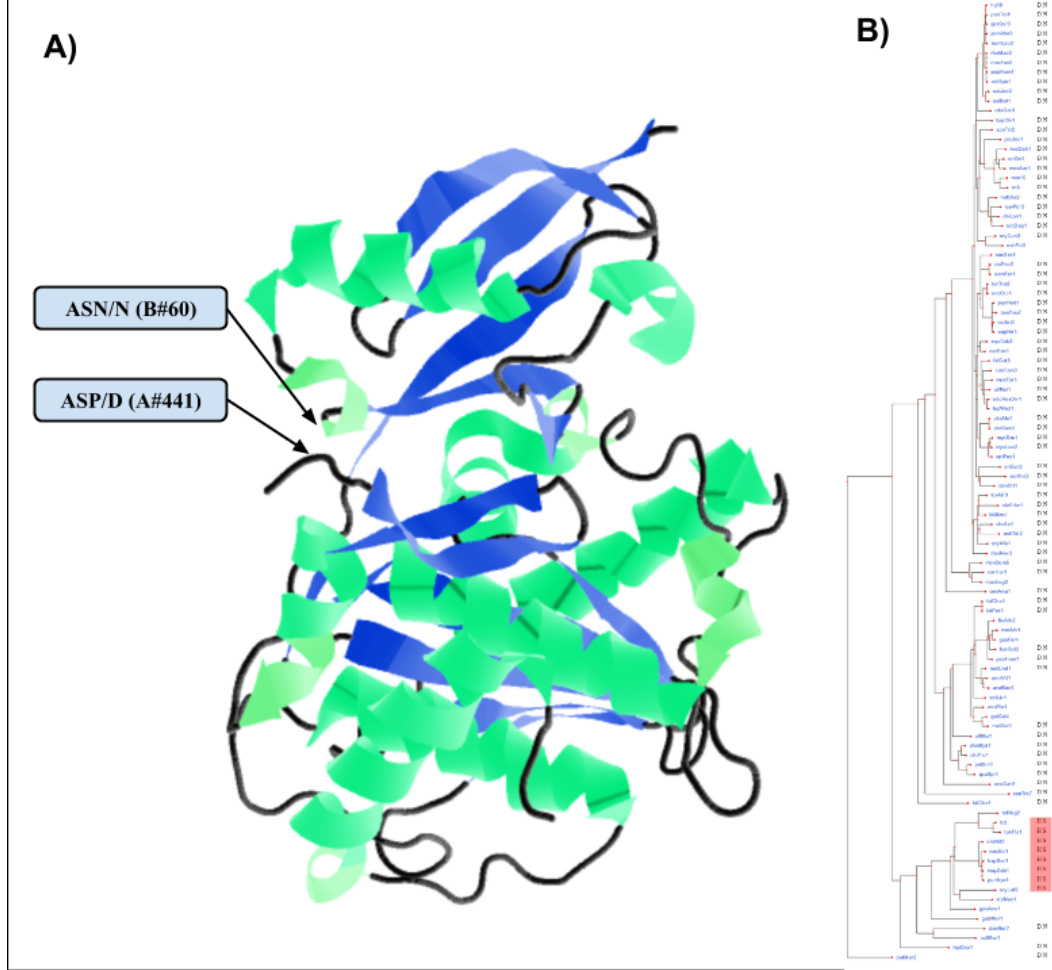


Figure 1–4: Example of amino acid interaction between SENP1 and SUMO1 proteins detected by our method $LL(M_{SA}) = 7.7$. A) PDB structure 2G4D, shows that the amino acids are in close proximity: amino acid #441 in SENP1 (ASP/D) interacts with amino acid #60 in SUMO1 (ASN/N). B) Multiple sequence alignment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

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 be 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, as required by our co-evolutionary model, only variants overlapping a multiple sequence alignment are taken into account (when several multiple sequence alignments overlapped a region, the alignment with the longest number of proteins was selected). Third, 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. Fourth, 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. Once the results are obtained, we can focus on interactions by further filtering results and keeping variant pairs whose alternative logistic model (see equation 1.6) has small absolute values for θ_1 and θ_2 while having large absolute values for θ_3 , specifically we keep results if $|\theta_3| > K(|\theta_1| + |\theta_2|)$ (based on empirical data, we set $K = 3$).

After all filters have been applied, using a model GWAS data which involves $\sim 13,000$ individuals and over $1.7M$ variants, the number of variant pairs to be analyzed is less than 10^8 variants, as opposed to the naive estimation of $\sim 10^{12}$ variant pairs. By means of the z-score relationship [66], 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 ??.

Results. Variant annotated and filtered according to the previous paragraphs lead to $\sim 32M$ pairs of variants having high log likelihood in our logistic regression model ($LL_{LogReg} > 6$, in equation 1.4) that were further analysed under co-evolutionary and Bayesian models. The complete analysis takes less

Combined model log10(BF)	Logistic regression		Variant 1			Variant 2		
	p-value	log10(BF)	Coordinate	Gene	Functional annotation	Coordinate	Gene	Functional annotation
8.21	7.80E-07	6.94	20:48129705_G/T_A	PTGIS	STOP_GAINED	16:81348733_G/T	GAN	NON_SYNONYMOUS_CODING
7.12	2.73E-06	6.47	4:90743415_T/C	SNCA	NON_SYNONYMOUS_CODING	2:179659911_G/A	TTN	NON_SYNONYMOUS_CODING
7.35	2.26E-06	6.43	3:53213690_G/C	PRKCD	NON_SYNONYMOUS_CODING	14:75746689_C/T	FOS	NON_SYNONYMOUS_CODING
6.80	2.82E-06	6.36	2:242795125_G/A	PDCD1	NON_SYNONYMOUS_CODING	8:13356801_G/A	DLC1	NON_SYNONYMOUS_CODING
6.35	3.85E-06	6.22	11:57582923_G/A	CTNND1	AA_modification:Phosphoserine	2:220420784_G/A	OBSL1	NON_SYNONYMOUS_CODING
8.08	6.60E-06	6.07	1:112298763_T/C	DDX20	NON_SYNONYMOUS_CODING	3:125826058_T/C	ALDH1L1	NON_SYNONYMOUS_CODING
7.22	4.77E-06	6.04	2:179457146_G/A	TTN	NON_SYNONYMOUS_CODING	22:35695930_C/A	TOM1	NON_SYNONYMOUS_CODING
7.20	9.62E-06	6.00	1:45797504_C/G	MUTYH	NON_SYNONYMOUS_CODING	8:30982425_T/G	WRN	NON_SYNONYMOUS_CODING
7.45	1.58E-05	5.90	16:4855278_A/G	GLYR1	AA_modification:Phosphoserine	8:13259100_G/A	DLC1	NON_SYNONYMOUS_CODING
6.81	1.55E-05	5.81	11:45975129_C/T	PHF21A	NON_SYNONYMOUS_CODING	19:49458190_C/A	BAX	NON_SYNONYMOUS_CODING
8.67	1.02E-05	5.81	7:128490102_G/A	FLNC	NON_SYNONYMOUS_CODING	8:13357339_G/C	DLC1	NON_SYNONYMOUS_CODING
6.42	9.91E-06	5.79	11:236090_G/A	SIRT3	NON_SYNONYMOUS_CODING	4:110615838_C/T	CASP6	NON_SYNONYMOUS_CODING
7.87	1.08E-05	5.79	8:144993930_C/G	PLEC	NON_SYNONYMOUS_CODING	14:73422258_C/G	DCAF4	NON_SYNONYMOUS_CODING
7.86	2.30E-05	5.70	1:45224997_A/C	KIF2C	NON_SYNONYMOUS_CODING	2:108921032_C/T	SULT1C2	NON_SYNONYMOUS_CODING
6.47	1.56E-05	5.68	11:134252895_C/T	B3GAT1	NON_SYNONYMOUS_CODING	15:75012984_T/C	CYP1A1	NON_SYNONYMOUS_CODING
5.67	1.24E-05	5.63	6:116441645_C/G	COL10A1	NON_SYNONYMOUS_CODING	20:30072135_G/A	REM1	NON_SYNONYMOUS_CODING
7.40	2.89E-05	5.61	1:201016295_G/A	CACNA1S	NON_SYNONYMOUS_CODING	19:17000695_G/A	F2RL3	NON_SYNONYMOUS_CODING
5.87	8.39E-07	5.59	7:43351409_T/G	HECW1	NON_SYNONYMOUS_CODING	2:219294200_A/G	VIL1	NON_SYNONYMOUS_CODING
5.84	2.34E-05	5.58	15:67457334_A/G	SMAD3	NON_SYNONYMOUS_CODING	1:201052381_A/G	CACNA1S	NON_SYNONYMOUS_CODING
7.00	1.46E-05	5.56	10:53822300_A/G	PRKG1	NON_SYNONYMOUS_CODING	9:140007465_G/A	DPP7	NON_SYNONYMOUS_CODING
6.93	2.96E-05	5.56	2:225362477_C/T	CUL3	NON_SYNONYMOUS_CODING	9:120476787_C/G	TLR4	STOP_GAINED

Table 1–5: Results from epistatic GWAS analysis of type II diabetes sequencing data

than 2 days using a 1,000 CPU-cluster, thus that an epistatic GWAS analysis of large cohort sequencing data is feasible using current computational resources.

Table 1–5 shows the main results from our GWAS epistatic analysis. Selected genes are shown in red in table 1–5 belong to a hand curated set of genes either associated with diabetes or known to be in diabetes related pathway. does this look like something interesting?

It should be noted that some of the top results include known amino acid modification sites such as Phosphoserine (or Glycosylation, not shown), these sites any biological meaning? it looks like these site should indeed be interacting....

From the top results what story can we tell?

1.6 Discussion

In this paper, we propose a novel methodology for genome wide association studies of pairs of variants under putative epistatic interaction. 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, but our highly optimized algorithms are suitable for large scale sequencing genomic studies.

The co-evolutionary model we propose in section 1.4.1 requires multiple sequence alignment and the corresponding phylogenetic tree. Both the tree and the number of sequences in the M_{SA} 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 [81], a genome wide multiple sequence alignment of 100 organisms, which has single genome wide phylogenetic tree.

We tested whether our co-evolutionary model can separate clinically relevant variants from ClinVar database [88] according to their clinical significance attribute (CLNSIG). Interestingly, variants categorized as “benign” or “druggable” have higher scores (mean $LL(M_{SA})$ within protein) than variants categorized as pathogenic (Supplementary Tables ??, ?? and Figure ??). We speculate that this might be because amino acids that can be compensated would be characterized as “benign” whereas deleterious amino acids changes cannot be compensated by mutation.

Finally, we show the application of our methods to a large scale exome sequencing study for type II diabetes consisting of $\sim 13,000$ samples and $\sim 1,7M$ variants. First, this shows that it is feasible to apply our methods GWAS-scale datasets. Second we show that some association of pairs of putatively interacting variants with type II diabetes have Bayes Factors showing suggestive association. Although larger cohorts are needed in order to find risk alleles that may be hidden by lower frequencies, thus not captured by this study.

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 [166] 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] Dario Acampora, Virginia Avantaggiato, Francesca Tuorto, Paolo Barone, Heinrich Reichert, Robert Finkelstein, and Antonio Simeone. Murine *otx1* and drosophila *otd* genes share conserved genetic functions required in invertebrate and vertebrate brain development. *Development*, 125(9):1691–1702, 1998.
- [2] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.
- [3] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, and S.R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.
- [4] Orly Agamy, Bruria Ben Zeev, Dorit Lev, Barak Marcus, Dina Fine, Dan Su, Ginat Narkis, Rivka Ofir, Chen Hoffmann, Esther Leshinsky-Silver, et al. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *The American Journal of Human Genetics*, 87(4):538–544, 2010.
- [5] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 1995.
- [6] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [7] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
- [8] Gerd Anders, Sebastian D Mackowiak, Marvin Jens, Jonas Maaskola, Andreas Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, and Christoph Dieterich. dorina: a database of rna interactions in post-transcriptional regulation. *Nucleic acids research*, page gkr1007, 2011.

- [9] Jennifer A Anderson, William D Gilliland, and Charles H Langley. Molecular population genetics and evolution of drosophila meiosis genes. *Genetics*, 181(1):177–185, 2009.
- [10] Manuel Ascano, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. Identification of rna–protein interaction networks using par-clip. *Wiley Interdisciplinary Reviews: RNA*, 3(2):159–177, 2012.
- [11] Geraldine A Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, pages 11–10, 2013.
- [12] D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.
- [13] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.
- [14] David J Begun and Charles F Aquadro. Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster. 1992.
- [15] Callum J Bell, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4–65ra4, 2011.
- [16] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.
- [17] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arrian 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.
- [18] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, et al. Annotation of functional variation in personal genomes using regulomedb. *Genome research*, 22(9):1790–1797, 2012.

- [19] Sydney Brenner, AOW Stretton, and S Kaplan. Genetic code: the non-sensetriplets for chain termination and their suppression. 1965.
- [20] Saverio Brogna and Jikai Wen. Nonsense-mediated mrna decay (nmd) mechanisms. *Nature structural & molecular biology*, 16(2):107–113, 2009.
- [21] 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.
- [22] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl 1):D102–D106, 2008.
- [23] Lukas Burger and Erik van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, 2010.
- [24] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.
- [25] Judy C Chang and Yuet Wai Kan. beta 0 thalassemia, a nonsense mutation in man. *Proceedings of the National Academy of Sciences*, 76(6):2886–2889, 1979.
- [26] B Charlesworth, JA Coyne, and NH Barton. The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist*, pages 113–146, 1987.
- [27] Venkateswara R Chintapalli, Jing Wang, and Julian AT Dow. Using flyatlas to identify better drosophila melanogaster models of human disease. *Nature genetics*, 39(6):715–720, 2007.
- [28] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688, 2012.
- [29] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyererson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

- [30] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):0–1, 2012.
- [31] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Toxicogenomics in non-mammalian species*, page 92, 2012.
- [32] 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.
- [33] 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.
- [34] FS Collins, ES Lander, J. Rogers, RH Waterston, and I. Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.
- [35] 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.
- [36] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.
- [37] 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.
- [38] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [39] UniProt Consortium et al. Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic acids research*, 41(D1):D43–D47, 2013.
- [40] Heather J Cordell. Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

- [41] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.
- [42] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.
- [43] Jasmin Coulombe-Huntington, Kevin CL Lam, Christel Dias, and Jacek Majewski. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS genetics*, 5(12):e1000766, 2009.
- [44] BS Cox, MF Tuite, and CS McLaughlin. The ψ factor of yeast: a problem in inheritance. *Yeast*, 4(3):159–178, 1988.
- [45] Robert Culverhouse, Brian K Suarez, Jennifer Lin, and Theodore Reich. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.
- [46] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.
- [47] 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.
- [48] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12):e1001025, 2010.
- [49] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.
- [50] Glynn Dennis Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, Richard A Lempicki, et al. David: database for annotation, visualization, and integrated discovery. *Genome biol*, 4(5):P3, 2003.
- [51] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

- [52] TD Dreesen, DH Johnson, and S Henikoff. The brown protein of drosophila melanogaster is similar to the white protein and to components of active transport complexes. *Molecular and cellular biology*, 8(12):5206–5215, 1988.
- [53] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.
- [54] William G Fairbrother, Ru-Fang Yeh, Phillip A Sharp, and Christopher B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002.
- [55] Joseph Felsenstein and Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.
- [56] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.
- [57] Kevin M Flanigan, Diane M Dunn, Andrew von Niederhausern, Michael T Howard, Jerry Mendell, Anne Connolly, Carol Saunders, Ann Modrcin, Majed Dasouki, Giacomo P Comi, et al. Dmd trp3x nonsense mutation associated with a founder effect in north american families with mild becker muscular dystrophy. *Neuromuscular Disorders*, 19(11):743–748, 2009.
- [58] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, page gks1236, 2012.
- [59] Hong Gao, Julie M Granka, and Marcus W Feldman. On the classification of epistatic interactions. *Genetics*, 184(3):827–837, 2010.
- [60] M. Garber, M. Guttman, M. Clamp, M.C. Zody, N. Friedman, and X. Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.
- [61] David Gatfield, Leonie Unterholzner, Francesca D Ciccarelli, Peer Bork, and Elisa Izaurralde. Nonsense-mediated mrna decay in drosophila: at the intersection of the yeast and mammalian pathways. *The EMBO journal*, 22(15):3960–3970, 2003.
- [62] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455, 2005.

- [63] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.
- [64] Jeremy Goecks, Anton Nekrutenko, James Taylor, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.
- [65] 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.
- [66] Steven N Goodman. Toward evidence-based medical statistics. 2: The bayes factor. *Annals of internal medicine*, 130(12):1005–1013, 1999.
- [67] L Guarente, YS Lin, M Carey, L Breeden, MA Osley, J Gould, S Kim, M Kane, and L Hereford. Generating yeast transcriptional activators containing no yeast protein sequences. *Nature*, 350, 1991.
- [68] 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.
- [69] Markus Hafner, Steve Lianoglou, Thomas Tuschl, and Doron Betel. Genome-wide identification of mirna targets by par-clip. *Methods*, 58(2):94–105, 2012.
- [70] D.L. Hartl and A.G. Clark. *Principles of population genetics*. Sinauer associates Sunderland, Massachusetts, 2007.
- [71] David Haussler, Stephen J O’Brien, Oliver A Ryder, F Keith Barker, Michele Clamp, Andrew J Crawford, Robert Hanner, Olivier Hanotte, Warren E Johnson, Jimmy A McGuire, et al. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.
- [72] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.
- [73] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

- [74] Douglas A Hosack, Glynn Dennis Jr, Brad T Sherman, H Clifford Lane, Richard A Lempicki, et al. Identifying biological themes within lists of genes with ease. *Genome Biol*, 4(10):R70, 2003.
- [75] 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.
- [76] Ivaylo P Ivanov, Andrew E Firth, Audrey M Michel, John F Atkins, and Pavel V Baranov. Identification of evolutionarily conserved non-aug-initiated n-terminal extensions in human coding sequences. *Nucleic acids research*, 39(10):4220–4234, 2011.
- [77] 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.
- [78] Kenneth A Jones, Beth Borowsky, Joe A Tamm, Douglas A Craig, Margaret M Durkin, Meng Dai, Wen-Jeng Yao, Mary Johnson, Caryn Gunwaldsen, Ling-Yan Huang, et al. Gabab receptors function as a heteromeric assembly of the subunits gababr1 and gababr2. *Nature*, 396(6712):674–679, 1998.
- [79] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl 1):D411–D419, 2011.
- [80] 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.
- [81] 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.
- [82] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.
- [83] Martin Alexander Kennedy. Mendelian genetic disorders. *Encyclopedia of Life Sciences*, 2001.

- [84] Szymon M Kielbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.
- [85] Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.
- [86] Marilyn Kozak. An analysis of 5'-noncoding sequences from 699 vertebrate messenger rnas. *Nucleic acids research*, 15(20):8125–8148, 1987.
- [87] P. Kumar, S. Henikoff, and P.C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.
- [88] 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.
- [89] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, et al. nextprot: a knowledge platform for human proteins. *Nucleic acids research*, 40(D1):D76–D83, 2012.
- [90] B. Langmead and S.L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 2012.
- [91] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.
- [92] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.
- [93] C Lazure, WT Hum, and DM Gibson. Sequence diversity within a subgroup of mouse immunoglobulin kappa chains controlled by the igk-ef2 locus. *The Journal of experimental medicine*, 154(1):146–155, 1981.
- [94] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

- [95] H. Li. Improving snp discovery by base alignment quality. *Bioinformatics*, 27(8):1157–1158, 2011.
- [96] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.
- [97] H. Li and R. Durbin. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(5), 2009.
- [98] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589, 2010.
- [99] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008.
- [100] Chenxing Liu, Fuquan Zhang, Tingting Li, Ming Lu, Lifang Wang, Weihua Yue, and Dai Zhang. Mirsnp, a database of polymorphisms altering mirna target sites, identifies mirna-related snps in gwas snps and eqtls. *BMC genomics*, 13(1):661, 2012.
- [101] Dongmei Liu and Russell L Finley. Cyclin y is a novel conserved cyclin essential for development in drosophila. *Genetics*, 184(4):1025–1035, 2010.
- [102] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. *Human mutation*, 32(8):894–899, 2011.
- [103] 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.
- [104] D.G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J.K. Pickrell, S.B. Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.
- [105] Baijayanta Maiti, Sandrine Arbogast, Valérie Allamand, Mark W Moyle, Christine B Anderson, Pascale Richard, Pascale Guicheney, Ana Ferreira, Kevin M Flanigan, and Michael T Howard. A mutation in the sepn1 selenocysteine redefinition element (sre) reduces selenocysteine incorporation and leads to sepn1-related myopathy. *Human mutation*, 30(3):411–416, 2009.

- [106] 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.
- [107] 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.
- [108] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.
- [109] et. al. McCarthy, Mark. Variation in protein-coding sequence and predisposition to type 2 diabetes. *To be published*, 2015.
- [110] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.
- [111] 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.
- [112] 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.
- [113] 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.
- [114] Eszter Nagy and Lynne E Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects rna abundance. *Trends in biochemical sciences*, 23(6):198–199, 1998.
- [115] Francesco Napolitano, Renato Mariani-Costantini, and Roberto Tagliaferri. Bioinformatic pipelines in python with leaf. *BMC bioinformatics*, 14(1):201, 2013.

- [116] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.
- [117] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.
- [118] Jason O’Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*, 5(3):28, 2013.
- [119] Annette L Parks, Kevin R Cook, Marcia Belvin, Nicholas A Dompe, Robert Fawcett, Kari Huppert, Lory R Tan, Christopher G Winter, Kevin P Bogart, Jennifer E Deal, et al. Systematic generation of high-resolution deletion coverage of the drosophila melanogaster genome. *Nature genetics*, 36(3):288–292, 2004.
- [120] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.
- [121] 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.
- [122] Patrick C Phillips. Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.
- [123] Adrian Platts, Susan J Land, Lang Chen, Grier Page, Parsa Rasouli, Luan Wang, Xiangyi Lu, and Douglas M Ruden. Massively parallel resequencing of the isogenic drosophila melanogaster strain w1118; iso-2; iso-3 identifies hotspots for mutations in sensory perception genes. *Fly*, 3(3):192–204, 2009.
- [124] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.
- [125] 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.
- [126] Towfique Raj, Manik Kuchroo, Joseph M Replogle, Soumya Raychaudhuri, Barbara E Stranger, and Philip L De Jager. Common risk alleles

- for inflammatory diseases are targets of recent positive selection. *The American Journal of Human Genetics*, 92(4):517–529, 2013.
- [127] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.
 - [128] Ho Sung Rhee and B Franklin Pugh. Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, pages 21–24, 2012.
 - [129] Erin Rooney Riggs, Karen E Wain, Darlene Riethmaier, Melissa Savage, Bethanny Smith-Packard, Erin B Kaminsky, Heidi L Rehm, Christa Lese Martin, David H Ledbetter, and W Andrew Faucett. Towards a universal clinical genomics database: the 2012 international standards for cytogenomic arrays consortium meeting. *Human mutation*, 34(6):915–919, 2013.
 - [130] A.F. Rope, K. Wang, R. Evjenth, J. Xing, J.J. Johnston, J.J. Swensen, B. Moore, C.D. Huff, L.M. Bird, J.C. Carey, et al. Using vaast to identify an x-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency. *The American Journal of Human Genetics*, 2011.
 - [131] Douglas M Ruden, D Curtis Jamison, Barry R Zeeberg, Mark D Garfinkel, John N Weinstein, Parsa Rasouli, and Xiangyi Lu. The edge hypothesis: Epigenetically directed genetic errors in repeat-containing proteins (rcps) involved in evolution, neuroendocrine signaling, and cancer. *Frontiers in neuroendocrinology*, 29(3):428–444, 2008.
 - [132] 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.
 - [133] Radhakrishnan Sabarinathan, Hakim Tafer, Stefan E Seemann, Ivo L Hofacker, Peter F Stadler, and Jan Gorodkin. The rnasnp web server: predicting snp effects on local rna secondary structure. *Nucleic acids research*, 41(W1):W475–W479, 2013.
 - [134] Simon P Sadedin, Bernard Pope, and Alicia Oshlack. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, 28(11):1525–1526, 2012.
 - [135] Clara M Santiveri and M Jiménez. Tryptophan residues: Scarce in proteins but strong stabilizers of β -hairpin peptides. *Peptide Science*, 94(6):779–790, 2010.

- [136] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683–691, 2011.
- [137] William Saurin, Maurice Hofnung, and Elie Dassa. Getting in or out: early segregation between importers and exporters in the evolution of atp-binding cassette (abc) transporters. *Journal of molecular evolution*, 48(1):22–41, 1999.
- [138] Eric Schadt. First steps on a long road. *Science*, 331(6018):691–691, 2011.
- [139] Gert C Scheper, Marjo S van der Knaap, and Christopher G Proud. Translation matters: protein synthesis defects in inherited disease. *Nature Reviews Genetics*, 8(9):711–723, 2007.
- [140] Valerie Schneider and Deanna Church. Genome reference consortium. 2013.
- [141] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLoS computational biology*, 3(3):e42, 2007.
- [142] 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.
- [143] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.
- [144] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
- [145] Chris Stark, Bobby-Joe Breitkreutz, Teresa Regul, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.
- [146] Nina Stoletzki and Adam Eyre-Walker. The positive correlation between dn/ds and ds in mammals is due to runs of adjacent substitutions. *Molecular biology and evolution*, 28(4):1371–1380, 2011.

- [147] 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.
- [148] H Sugihara, V Andrisani, and PM Salvaterra. Drosophila choline acetyltransferase uses a non-aug initiation codon and full length rna is inefficiently translated. *Journal of Biological Chemistry*, 265(35):21714–21719, 1990.
- [149] GATK team. The genome analysis toolkit. Accessed: 2015.
- [150] Stephen T Thibault, Matthew A Singer, Wesley Y Miyazaki, Brett Milash, Nicholas A Dompe, Carol M Singh, Ross Buchholz, Madelyn Demsky, Robert Fawcett, Helen L Francis-Lang, et al. A complementary transposon tool kit for drosophila melanogaster using p and piggybac. *Nature genetics*, 36(3):283–287, 2004.
- [151] Van Khanh Tran, Yasuhiro Takeshima, Zhujun Zhang, Yasuaki Habara, Kazuhiro Haginoya, Atsushi Nishiyama, Mariko Yagi, and Masafumi Matsuo. A nonsense mutation-created intraexonic splice site is active in the lymphocytes, but not in the skeletal muscle of a dmd patient. *Human genetics*, 120(5):737–742, 2007.
- [152] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.
- [153] Jens Tyedmers, Maria Lucia Madariaga, and Susan Lindquist. Prion switching in response to environmental stress. *PLoS biology*, 6(11):e294, 2008.
- [154] 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.
- [155] GEORG HM WAALER. The location of a new second chromosome eye colour gene in drosophila melanogaster. *Hereditas*, 2(3):391–394, 1921.
- [156] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, 33(1):79–86, 2009.
- [157] K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

- [158] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934, 2012.
- [159] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.
- [160] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.
- [161] 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.
- [162] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhi-jani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872–876, 2008.
- [163] David C Whitcomb, Michael C Gorry, Robert A Preston, William Furey, Michael J Sossenheimer, Charles D Ulrich, Stephen P Martin, Lawrence K Gates, Stephen T Amann, Phillip P Toskes, et al. Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature genetics*, 14(2):141–145, 1996.
- [164] Reed B Wickner. [ure3] as an altered ure2 protein: evidence for a prion analog in *saccharomyces cerevisiae*. *Science*, 264(5158):566–569, 1994.
- [165] 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.
- [166] 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.
- [167] Yumi Yamaguchi-Kabata, Makoto K Shimada, Yosuke Hayakawa, Shinsei Minoshima, Ranajit Chakraborty, Takashi Gojobori, and Tadashi Imanishi. Distribution and effects of nonsense polymorphisms in human genes. *PloS one*, 3(10):e3393, 2008.

- [168] Jian-Hua Yang, Jun-Hao Li, Peng Shao, Hui Zhou, Yue-Qin Chen, and Liang-Hu Qu. starbase: a database for exploring microRNA-mRNA interaction maps from argonaute clip-seq and degradome-seq data. *Nucleic acids research*, 39(suppl 1):D202–D209, 2011.
- [169] Ziheng Yang. *Computational molecular evolution*, volume 284. Oxford University Press Oxford, 2006.
- [170] Amelia Younossi-Hartenstein, Patricia Green, Gwo-Jen Liaw, Karen Rudolph, Judith Lengyel, and Volker Hartenstein. Control of early neurogenesis of the drosophila brain by the head gap genes *nan*, *otd*, *ems*, and *mbd*. *Developmental biology*, 182(2):270–283, 1997.
- [171] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.
- [172] 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.
- [173] Jesse D Ziebarth, Anindya Bhattacharya, Anlong Chen, and Yan Cui. Polymir database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic acids research*, page gkr1026, 2011.
- [174] 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.
- [175] 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.