

# 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 type II diabetes (T2D) have been identified [?, ?]. However the combined effect sizes of all these loci account for less than 10% of the overall disease predisposition [?]. 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. The lack of large genetic effects is known as the “missing heritability” problem and does not only arise in T2D but also in almost all complex traits. Recent studies on the topic of missing heritability [?, ?] suggest that genetic interactions (a.k.a. epistasis) might be at least partly responsible for this issue.

In this chapter, we propose a novel framework that takes into account putative epistatic interactions in the context of genome wide association studies (GWAS) to uncover potentially interacting variants that might affect disease risk. Although in the computational approaches we describe is applicable to any complex trait, in this chapter we apply it to diabetes GWAS data. Type II diabetes (T2D) is a complex disease that was 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 around 8% to 9% in adults and an estimated 1.5 millions deaths were caused by diabetes in 2012 [?], which is predicted to be the 7th leading cause of death by 2030. The costs associated to treating diabetes patients in the U.S. alone are estimated around \$245 billion dollars.

The rest of the chapter is from the following paper: **P. Cingolani**, R. Sladek, M. Blanchette, “A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies”, to be submitted to PLOS Computational Biology.

## 1.2 Abstract

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

**Methods.** In this paper, we propose a novel methodology for identifying putative epistatic interactions by combining interspecies comparison and population level variation. Using crystal structures for individual proteins and for complexes as well

as genome wide multiple species alignment, we create a co-evolutionary substitution model that allows the calculation of the posterior probability of physical interaction between residues. These probabilities are then used as the interaction priors for an epistatic GWAS analysis using a Bayesian framework.

**Results.** Our 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](https://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 [?]. 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 of complex traits. For instance, all genetic variants associated to height collectively account for few centimetres in the offspring’s height [?]. Similarly the known variants related to type 2 diabetes risk collectively explain only 5% to 10% of the overall variance in disease predisposition [?, ?]. This problem is known as “missing heritability” [?] and recent theories suggest that genetic interactions (epistasis) might play an important role in it [?, ?].

The foundations for epistasis [?] 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 multi-locus genotype values from an additive linear model for the value of a phenotype” [?]. 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 [?]. Many types of situations can lead to epistatic interactions, perhaps the most common involves pairs of variants that encode amino acids whose physical interactions is required for their respective proteins to function in a pathway.

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 [?]. 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. [?]) and large databases of protein-protein interactions are now available for human [?, ?]. 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 [?].

These issues can be partially addressed using computational predictions of either pairs of interacting proteins or interacting residues [?]. 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 [?]. 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 [?]. This is a phenomenon generically known as co-evolution. 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 to analyse multiple sequence alignments of proteins from different organisms to find co-evolving sites. This types of approaches has been used to identify co-evolving sites both within a protein (e.g. N-terminal and C-terminal domains in PKG protein [?], GroES-L chaperoning system [?],  $\alpha$  and  $\beta$  haemoglobin subunits [?]), and between interacting proteins (e.g. G-protein coupled receptors and protein ligands [?]).

Many methods exist to find putative interaction loci, both within and across proteins, based on evolutionary evidence (see [?] for a review). The earliest methods for inferring co-evolution used either correlation or mutual information between two loci [?] in a multiple sequence alignment. However, these methods are known to have systematic biases due to the fact that they ignore phylogenetic relationships [?] or sequence heterogeneity problems [?]. More sophisticated methods, such as DCA [?], PSICOV [?] or mdMI [?] try to overcome these biases by using global statistical models, however they are 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 [?]), and such depth remains only available for a small number of proteins. 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) [?].

Considering epistatic interactions in GWAS is challenging for several reasons: i) interaction models are by definition non-linear [?]; ii) analysing all order  $N$  variant combinations requires great computational power and efficient algorithms or might just be infeasible because the number tests grows exponentially with  $N$  [?]; iii) multiple hypothesis testing correction can render association tests underpowered for all but very large cohorts [?, ?]; and iv) there is no consensus of what genetic interaction means, which is reflected in the difficulty to find a unified model [?, ?]. 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, the required sample sizes to detect epistatic interactions depend on phenotypic effect size and variants’ allele frequencies with some estimates assuming on the order of 10,000 to 500,000 cases [?] 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 are based on a wide array of methodologies. In Zhao et alii [?], the authors infer epistatic interaction probabilities by noting that interactions create linkage disequilibrium patterns in the disease population. A Bayesian framework in [?] takes into account several risk models and uses Dirichlet priors to solve each model analytically, then they combine them in the full model posterior distribution calculated using an MCMC sampling technique. In Ackermann et alii [?], 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 [?]. While all algorithms have relative advantages, there is no standard in epistatic GWAS analysis and we believe that better methods can be created by combining other sources of biological information, such as evolutionary evidence.

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 [?] 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. The posterior probabilities can then be used as priors to modulate the evidence of epistatic interaction derived from GWAS data. Our implementation is efficient enough to be applied to GWAS-scale datasets of tens of thousands of samples. Finally



we showcase this method using a cohort of  $\sim 13,000$  individuals in a case-control study of type II diabetes (for study details, see [?]) and identify suggestive associations of putatively epistatic interactions.

## 1.4 Methods

Our epistatic GWAS analysis pipeline involves three key steps, as shown in Figure ?? . First, we learn a co-evolutionary substitution rate matrix from pairs of amino acids that are known to be in contact within proteins. Second, we analyse a GWAS data set to identify pairs of non-synonymous variants that show (possibly weak) evidence of epistasis. Third, for each pair of variants 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 calculating a joint Bayes factor.

### 1.4.1 Substitution model for pairs of interacting amino acids

In this section, we describe how we estimate two substitution rate matrices: i) the first is the usual  $20 \times 20$  substitution rate matrix  $Q$  describing the evolution of individual amino acids; ii) the second,  $Q_2$ , is a  $400 \times 400$  substitution rate matrix for pairs of interacting residues.

We use the 100-way vertebrate multiple sequence alignment and accompanying phylogenetic tree  $T$  available from the UCSC Genome Browser [?]. This alignment includes the DNA sequences of 100 species whose genome is completely or nearly completely sequenced, with 12 primates, 44 non-primates eutherians, 5 marsupials, 14 birds, 6 reptiles, 16 ray-finned fish and 8 lobe-finned fish. The multiple alignment is performed using the multiz algorithm [?, ?].

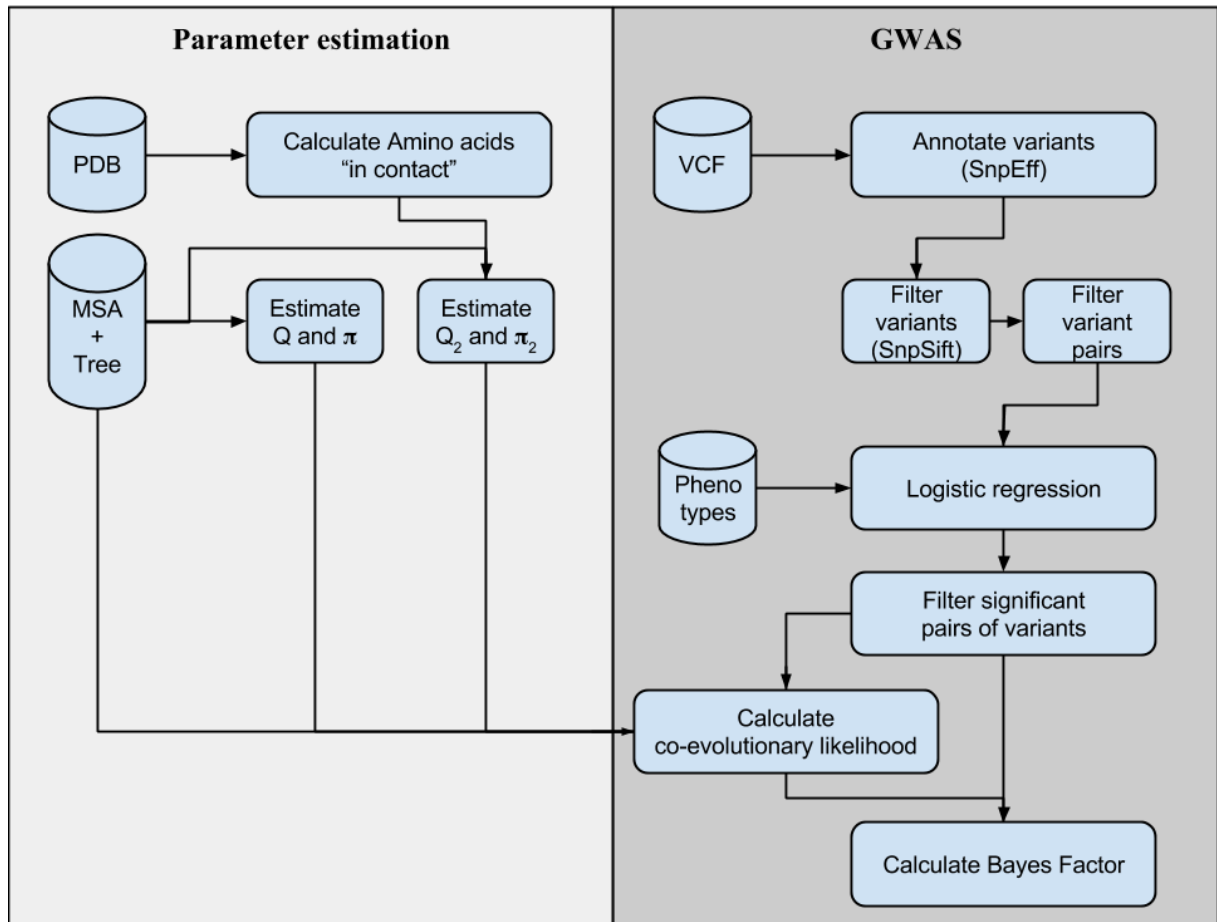


Figure 1–1: Complete pipeline example

From the  $\sim 21,000$  human protein structures (resolution less than  $3 \text{ \AA}$ ) 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  $3 \text{ \AA}$ . 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 the rate matrix  $Q$ , we consider the complete set of  $n \sim 22 \times 10^6$  protein coding sites present in the alignment, irrespective of the presence or absence of contacts.  $Q$  is obtained following classical sequence evolution theory ([?, ?]). 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 species  $s_j$  has amino acid  $b$  given that species  $s_i$  has amino acid  $a$ . Then  $P_{i,j}$  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  $Q = \log[P_{i,j}/t_{i,j}]$  [?]. Taking into account the estimation error, the equation becomes  $\hat{Q}_{i,j} = \log[P_{i,j}/t_{i,j}] + \epsilon_{i,j}$ , 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:

$$\begin{aligned}\hat{Q} &= \frac{1}{N(N-1)/2} \sum_{i < j} \hat{Q}_{i,j} \\ &= \frac{2}{N(N-1)} \sum_{i < j} \frac{1}{t_{i,j}} \log[\hat{P}_{i,j}]\end{aligned}$$

The much larger substitution matrix  $Q_2$  describes the substitution rate from any pair of amino acids  $(a, b)$  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.

#### 1.4.2 Calculating likelihood of individual and pairs of alignment columns

Given a substitution rate matrix  $Q$ , a multiple sequence alignment MSA, and a phylogenetic tree  $T$  for the sequences in MSA, the likelihood of an alignment column  $MSA(i)$  (i.e.  $L_1[MSA(i)]$ ) can be calculated using Felsenstein's algorithm [?], which has a time complexity of  $O(N \cdot |\Sigma|^2)$ , where  $|\Sigma| = 20$  is the alphabet size and  $N$  is the number of sequences in the alignment. Given matrix  $Q_2$ , the same algorithm can be used to compute the likelihood  $L_2[MSA(i), MSA(j)]$  of a pair of alignment columns  $(MSA(i), MSA(j))$ , but now in time  $O(N \cdot |\Sigma|^4)$ .

A test for co-evolution of two positions  $i, j$  from the same or different proteins is obtained from the log-likelihood ratio of these two models:

$$\ell_C[MSA(i), MSA(j)] = \log \left[ \frac{L_2[MSA(i), MSA(j)]}{L_1[MSA(i)] \cdot L_1[MSA(j)]} \right] \quad (1.1)$$

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

In a GWAS study these likelihood calculations need to be performed on a very large number of pairs of sites, thus optimizations are required to ensure manageable running time. First, pre-calculation of matrix exponentials  $P(t) = e^{Qt}$  can be performed for all values of  $t$  corresponding to individual branch lengths in the phylogenetic tree  $T$ . Second, “constant-tree caching” is used to cache likelihood values for sub-trees of the phylogenetic tree where all nodes have the same amino acid

### 1.4.3 GWAS model

Consider a GWAS with  $N_S$  samples (individuals) and  $N_V$  variants, we use the standard notation for phenotypes and code them as  $d_s = 1$  when individual  $s$  is affected by disease (cases) and  $d_s = 0$  if the individual is “healthy” (control). 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  $s$  at locus  $i$ . A logistic model of disease risk [?] is

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

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 [?]),  $\bar{g}_{s,i} = [1, g_{s,i}, c_{s,1}, c_{s,2}, \dots, c_{s,N_C}]$ , and  $\bar{\theta} = [\theta_1, \theta_2, \dots, \theta_m]$ . The parameter

estimates  $\bar{\theta}$  are obtained by solving the maximum likelihood equation

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

Using this model, we have two hypotheses: i) the null hypothesis,  $H_0$ , assumes that genotype does not influence disease probability (i.e.  $\theta_1 = 0$ ); and ii) the alternate hypothesis,  $H_1$ , assumes that the genotype does influence disease probability (i.e.  $\theta_1 \neq 0$ ). We can compare these two hypotheses using a likelihood ratio test, so we define

$$\ell_G = \log \left[ \frac{L(\bar{\theta} | H_1)}{L(\bar{\theta}' | H_0)} \right]$$

where  $\bar{\theta}'$  and  $\bar{\theta}$  are the maximum likelihood estimates for null and alternate models respectively. According to Wilk's theorem [?], the log likelihood ratio has a  $\chi_1^2$  distribution under the null hypothesis, 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

$$P(d_s | g_{s,i}, g_{s,j}, H_1) = \phi[\theta_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3 (g_{s,i} g_{s,j})] \quad (1.2)$$

$$\dots + \theta_4 c_{s,1} + \dots + \theta_m c_{s,N_{cov}}] \quad (1.3)$$

$$= \phi(\bar{\theta}^T \bar{g}_{s,i,j}) \quad (1.4)$$

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 are not located 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 sufficiently distant (say  $> 1MB$ ), we use the latter. The null hypothesis  $H_0$  assumes that variants act independently

$$P(d_s | g_{s,i}, g_{s,j}, H_0) = \phi[\theta'_0 + \theta'_1 g_{s,i} + \theta'_2 g_{s,j} + \theta'_3 c_{s,1} + \dots] \quad (1.5)$$

$$= \phi(\bar{\theta}'^T \bar{g}'_{s,i,j}) \quad (1.6)$$

where  $\bar{g}'_{s,i,j} = [1, g_{s,i}, g_{s,j}, c_{s,1}, c_{s,2}, \dots, c_{s,N_{cov}}]^T$ .

Just as in the single loci, this requires fitting the logistic regression parameters, thus we investigated several algorithms for logistic regression parameter fitting. The fastest convergence is obtained using Iterative Reweighted Least Squares (IRWLS [?]) and Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS [?]) 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 [?, ?]

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

where  $BF$ , 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 converges to Bayesian Information Criteria (BIC) [?], and ii) we can compare non-nested models. The Bayes factor for the epistatic model becomes:

$$BF_G = \frac{\int \prod_{s=1}^{N_S} \phi(\bar{\theta}^T \bar{g}_{s,i,j})^{d_s} [1 - \phi(\bar{\theta}^T \bar{g}_{s,i,j})]^{1-d_s} P(\bar{\theta}|H_1) d\bar{\theta}}{\int \prod_{s=1}^{N_S} \phi(\bar{\theta}'^T \bar{g}'_{s,i,j})^{d_s} [1 - \phi(\bar{\theta}'^T \bar{g}'_{s,i,j})]^{1-d_s} P(\bar{\theta}'|H_0) d\bar{\theta}'} \quad (1.7)$$

Calculating Bayes factors is challenging because most of the times there are no closed form equations and even if the integrals can be computed by means of numerical algorithms, it imposes a significant computational burden thus making it impractical for large datasets such as GWAS data. We can approximate the integrals using Laplace's method [?]. If  $g(x)$  has a maximum 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  $\bar{x} \in \Re^d$  is an analogous extension that uses a Hessian matrix instead of a second derivative of  $g(\cdot)$

$$\int e^{\lambda g(\bar{x})} h(\bar{x}) d\bar{x} \simeq h(\bar{x}_0) e^{\lambda g(\bar{x}_0)} \left( \frac{2\pi}{\lambda} \right)^{d/2} \left[ \frac{\partial^2 g(\bar{x})}{\partial \bar{x} \partial \bar{x}^T} \right]^{-1/2} \quad (1.8)$$

Using equation ?? we can try to approximate the improper integrals in equation ?? 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 applying Eq.?? at the point of the maximum likelihood. In order to do so, we need to calculate the Hessian matrix in Eq.??.



Fortunately in logistic models , 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.

Calculating Bayes factors involves using prior parameter distributions. In order to estimate these distributions, we run the logistic regression fitting analysis and plot the parameter distributions for different levels of significance. As expected most parameters have uni-modal distribution, except for  $\theta_3$ , which has a multi-modal distribution. For all parameters, except  $\theta_3$ , we use a normal distribution centred at the mean and variance set to 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  $\theta_3$ , which has a multi-modal distribution, we fit a mixture model parameters using an EM algorithm to fit a Gaussian bimodal.

**Computational and statistical issues.** The computational burden for the 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 analysed, thus significantly increasing the number of statistical tests. This also reduces the statistical power since the required p-value significance level would be orders of magnitude smaller. A naïve 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  $g_i$  and  $g_j$  are such that the interaction term  $(g_{s,i}g_{s,j})$  is zero in all samples, which usually happens for pairs of rare variants, then  $BF_G = 1$ . 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 equations are undetermined, so we skip such variant pairs. Fifth, if one of the variants has high allele frequency with 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 pure epistatic interactions (i.e. interactions with no marginal contribution from each independent loci) by further filtering results and keeping variant pairs whose alternative logistic model (see equation ??) has small absolute values for  $\theta_1$  and  $\theta_2$  while having large absolute values for  $\theta_3$ , specifically we keep results if  $|\theta_3| > K(|\theta_1| + |\theta_2|)$  (based on empirical data, we set  $K = 3$ ).

#### 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 analyse variants from genome sequencing data first by filtering only for non-synonymous variants, then analysing all possible pairs of variants and filtering out those that are unsuitable for further analysis. 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 analysed 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 and logistic regression models are used to calculate the total Bayes Factor ( $BF_T$ ) by means of Laplace's approximation:

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

$$\cdot \frac{L_2[MSA(i), MSA(j)]}{L_1[MSA(i)] \cdot L_1[MSA(j)]}$$

$$BF_T = BF_G \cdot L_C$$

#### 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 protein sequence alignment these evolutionary models can be used as a statistical test for co-evolution between any two amino acids without using structural information. Second, a statistical test for epistasis is developed to identify pairs of non-synonymous variants that show (often weak) evidence of interaction in the way they associate to a given trait. Finally, evidence from the co-evolution model and the epistatic GWAS model are combined in a Bayesian framework.

### 1.5.1 Co-evolutionary substitution models

The approach described in Methods was used to obtain substitution rate matrices  $Q$  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 matrices such as PAM [?] (Supplementary Figure ??).

The structure of  $Q_2$ , which describes substitution rates between one pair of interacting amino acids to another, is richer (Supplementary Figure ??). Of particular interest are pairs of amino acids having high ratio  $R(ab, cd) = Q_2(ab, cd)/(Q(a, c) \cdot Q(b, d))$  which is an indicative of co-evolution. For example the highest ratio is found in amino acid pair transition V.I  $\rightarrow$  W.W (i.e. amino acid V mutated to W in one of the sequences, and amino acid I changed to W in the other). In fact all top 10 ratios are transitions to W.W amino acid pairs. This makes sense considering that: i) individual amino acid substitution rates to tryptophan are generally very low, but that ii) tryptophan pairs are well known  $\beta$ -hairpin stabilizers and are considered as a paradigm for designing stable  $\beta$ -hairpins [?].

Another type of pair transitions with large  $R(ab, cd)$  ratio is the double transitions to a pair of phenylalanine amino acids from a pair 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 energetically favourable [?].

### 1.5.2 Co-evolutionary model validation

We first assessed the ability of our co-evolutionary model to detect interacting sites located within the same protein by computing the likelihood ratio of candidate pairs of sites under a 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 that interacting sites tend to have higher likelihood ratio scores than non-interacting ones (Mann-Whitney p-value  $< 2.2 \times 10^{-16}$ ). Although the likelihood ratio score by 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 a co-evolutionary model estimated 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 ( $< 3$  Å) and  $\sim 3,000$  pairs of non-interacting ( $> 30$  Å) residues from distinct proteins, obtained from co-crystal structures in PDB (see Methods). As seen on Figure ??, the

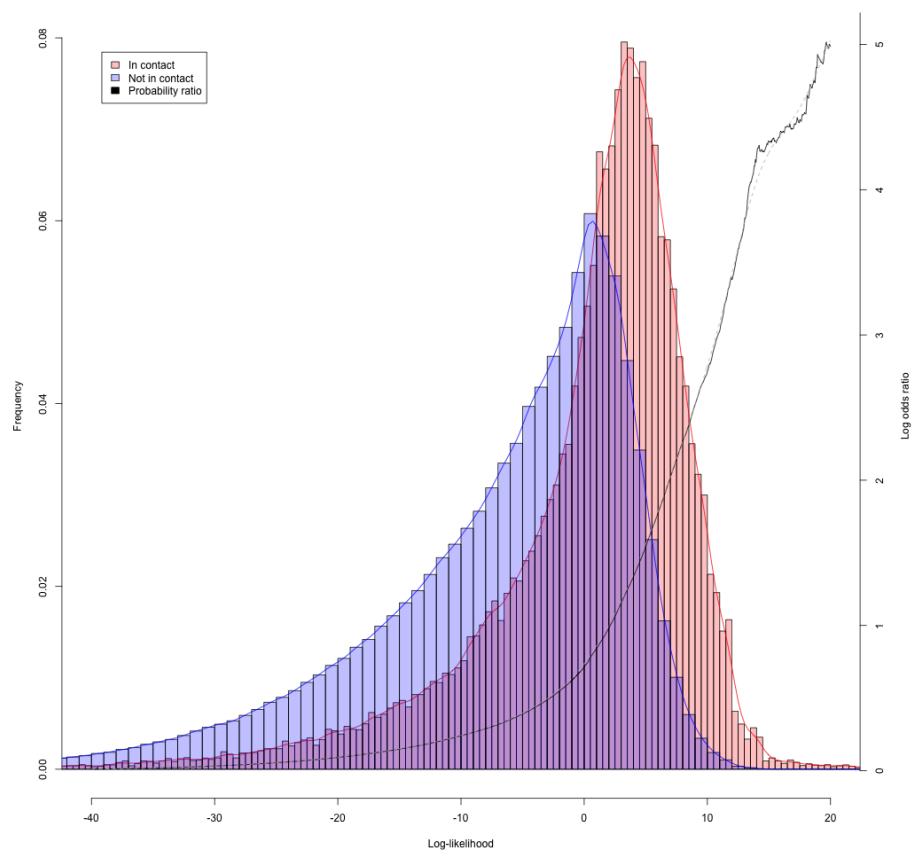


Figure 1-2: Histogram of log-likelihood values of pairs of amino acids in contact (red) and not in contact (blue) for amino acids within the protein (PDB). Log-odds of contacting vs non-contacting pairs (black) and smoothed log-odds (dotted grey).

two classes of sites have substantially different likelihood ratio distributions (Mann-Whitney one sided test:  $p - value < 2.2 \times 10^{-16}$ ), 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

$$\begin{aligned}\ell_{odds}(x) &= \log \left[ \frac{P[L_2(MSA(i), MSA(j)) \geq x]}{P[L_1(MSA(i)) \times L_1(MSA(j)) \geq x]} \right] \\ &\simeq e^{\alpha x} - \beta\end{aligned}$$

where  $\alpha = 0.195$  and  $\beta = 1.018$  (in order to avoid biases, the log odds value is capped to 4.0).

Figure ?? shows the example of a predicted contact  $\ell_C = 7.7$  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 multiple alignment columns.

Although our approach aims at identifying contacting residues from different proteins, it can also be used to predict the presence or absence of interactions between proteins as a whole. We extracted from BioGrid [?] 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 [?]), and their corresponding genes are expressed in the same tissue (GTex [?], expression of 1 FPKM or more, tissues  $\in$

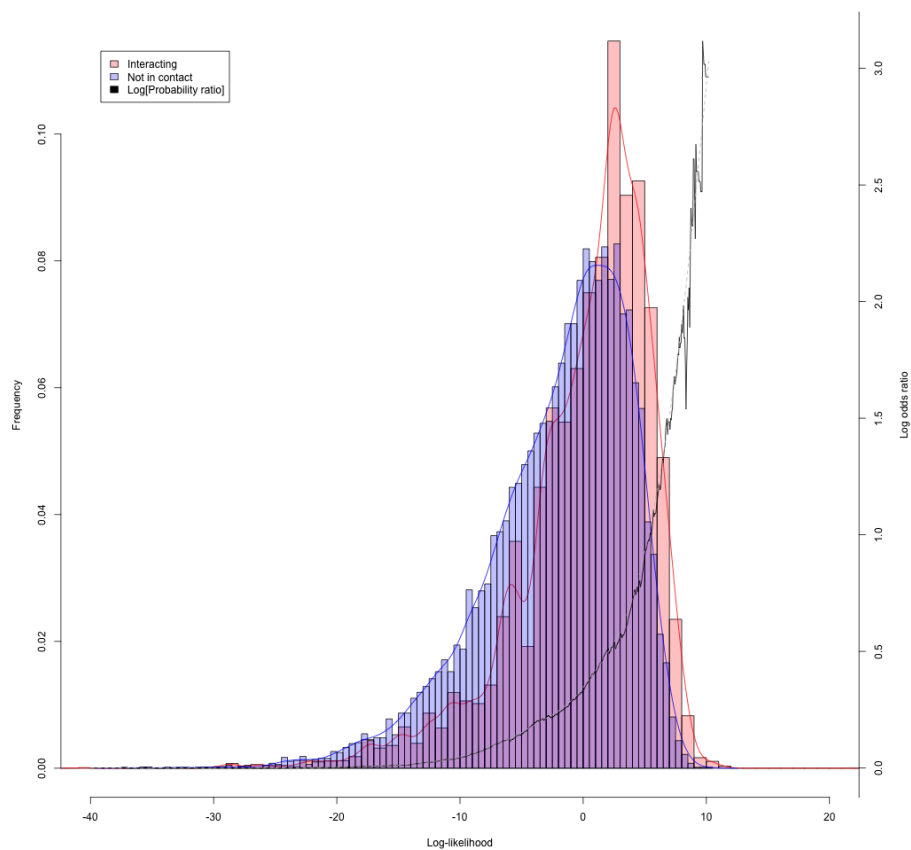


Figure 1–3: Histogram of log-likelihood values of pairs of amino acids in contact (red) and not in contact (blue) for amino acids in different proteins (co-crystallized entries from PDB). Log-odds of contacting vs non-contacting pairs (black) and smoothed log-odds (dotted grey)



{skeletal muscle, adipose tissue, pancreatic Islets}). We randomly selected as “non-interacting” pairs the same number of pairs amongst those that do not fulfil any of the three conditions.

Let the two proteins considered have amino acid sequences  $A = a_1 \dots a_m$  and  $B = b_1 \dots b_n$ . To obtain the prediction score for this pair of proteins, we identify the pair of length- $k$  sub-strings  $a_i, a_{i+1}, \dots, a_{i+k-1}$  and  $b_j, b_{j+1}, \dots, b_{j+k-1}$  that exhibit the strongest support for parallel or anti-parallel interactions

$$\ell_C(k) = \max \left[ \sum_{l=0}^{k-1} \ell_C[MSA(a_{i+l}), MSA(b_{j+l})], \sum_{l=0}^{k-1} \ell_C[MSA(a_{i+l}), MSA(b_{j+k-1-l})] \right]$$

where  $k = 3$  was determined empirically to provide the best predictive power. As shown in Figure ??), prediction accuracy is quite good (p-value  $< 2 \cdot 10^{-42}$ ), taking into account the modest amount of information considered by the model.

### 1.5.3 Epistatic GWAS analysis

We applied our methods to a cohort of  $\sim 13,000$  individuals in a case-control study of type II diabetes [?]. 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 1.7 million coding variants. The filters described in Methods section resulted in a number of variant pairs being analysed of less than 50 million. By means of the z-score relationship between Bayes Factor and p-values shown in [?], we can set the GWAS significance threshold for 50 million pairs at  $\log_{10}[BF_T] = 8.0$ .

**Results.** Variants annotated and filtered according to the previous paragraphs resulted in the identification of variant pairs having log likelihood  $\ell_G$  above 6 (see

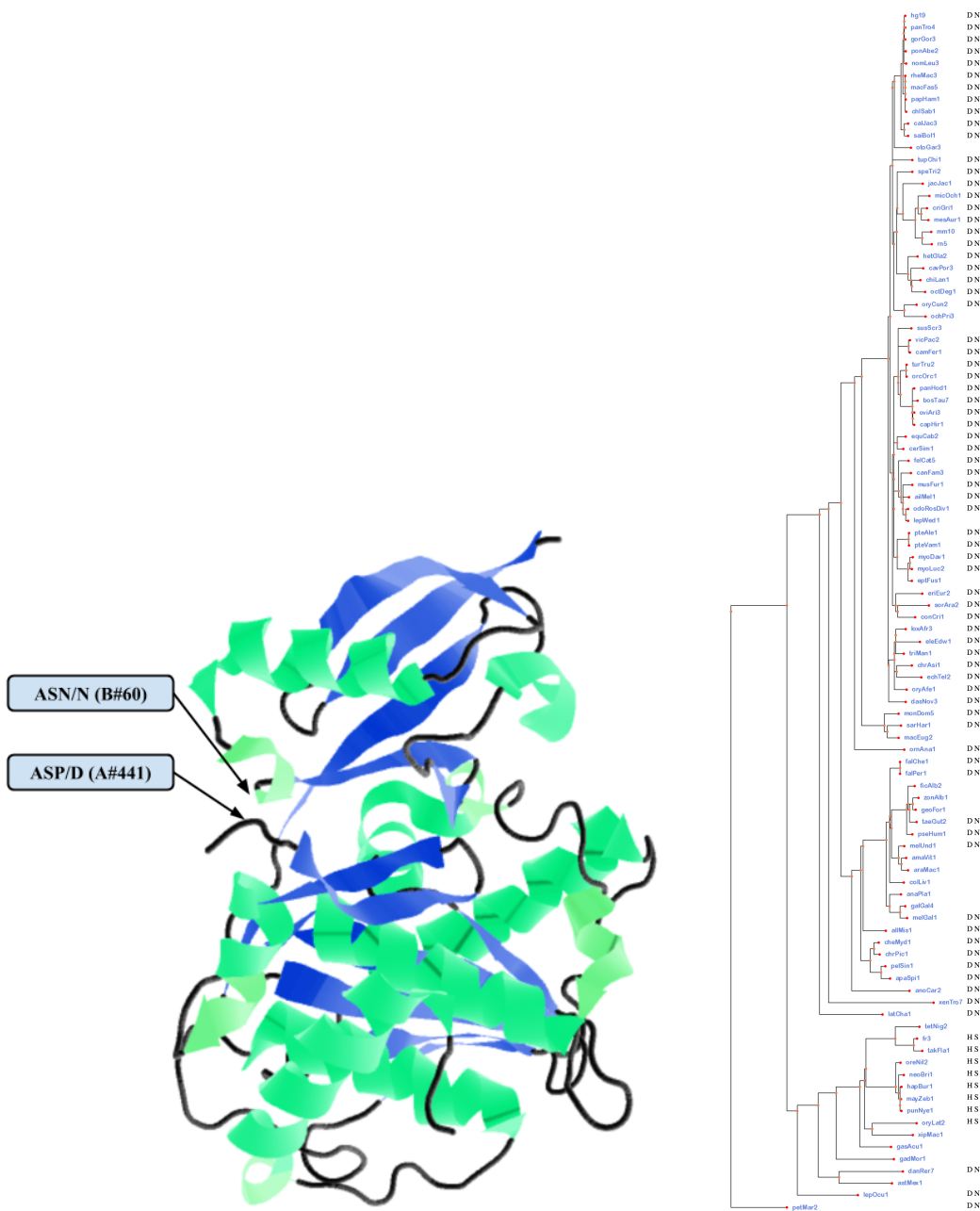


Figure 1–4: Example of interaction between amino acid #441 of *Senp1* and #60 of *Sumo1* proteins detected by our method with  $\ell_C = 7.7$ . Left: PDB structure 2G4D, shows that the amino acids are in close proximity. Right: Multiple sequence alignment and phylogenetic tree showing the putative compensatory amino acid substitution pair D.N replaced by H.S.

equation ??) that were further analysed under co-evolutionary and Bayesian models. The complete analysis took less than 2 days using a 1,000 CPU-cluster, thus showing that an epistatic GWAS analysis is feasible using current computational resources. Table ?? shows the main results from our GWAS epistatic analysis. Genes highlighted in red are associated with diabetes or known to be in diabetes-related pathway. It should be noted that some of the top results include amino acid modification sites such as Phosphoserine (or Glycosylation, not shown), which are likely to be interaction loci.

**Associations are purely epistatic.** In order to understand the nature of the putative epistatic associations, we compared our top epistatic GWAS results against a large single marker GWAS study of type II diabetes from the DIAGRAM consortia [?, ?] as well as a meta analysis from the same consortia [?]. For each of our top 1,000 results we looked up the closest entry from the DIAGRAM dataset (i.e. the entry having smallest distance, as measured in number of bases) and analysed the corresponding p-values from the three publicly released DIAGRAM datasets.

Only 14 of the top 1,000 results show suggestive association ( $p\text{-value} \leq 10^{-4}$ ) in a nearby loci (average distance 977 bases). Of these, only one of them is in our top 10 results (8:30982425-T/G variant in *WRN* gene) has DIAGRAM p-value of  $4.6 \times 10^{-4}$  and odds ratio 1.06. Neither of the other 13 suggestive single marker variants are our top 50 results. This indicates that our top results are populated by pure epistatic putative effects having little marginal association with disease risk. In light of this evidence, it should be noticed that *conditional search* methods (based on selecting the top single-marker results and performing epistatic analysis only on this subset)

Variant 1			Variant 2			Logistic regression		Co-evolutionary	Combined model
Coordinate	Gene	Functional annotation	Coordinate	Gene	Functional annotation	log10(BF)	p-value	Log-Likelihood	log10(BF)
20:48129705_G/T_A	PTGIS	STOP_GAINED	16:81348733_G/T	GAN	NON_SYNONYMOUS_CODING	6.94	7.80E-07	2.94	8.21
4:90743415_T/C	SNCA	NON_SYNONYMOUS_CODING	2:179659911_G/A	TTN	NON_SYNONYMOUS_CODING	6.47	2.73E-06	1.50	7.12
3:53213690_G/C	PRKCD	NON_SYNONYMOUS_CODING	14:75746689_C/T	FOS	NON_SYNONYMOUS_CODING	6.43	2.26E-06	2.11	7.35
2:242795125_G/A	PDCD1	NON_SYNONYMOUS_CODING	8:13356801_G/A	DLC1	NON_SYNONYMOUS_CODING	6.36	2.82E-06	1.01	6.80
11:57582923_G/A	CTNND1	AA_modification:Phosphoserine	2:220420784_G/A	OBSL1	NON_SYNONYMOUS_CODING	6.22	3.85E-06	0.29	6.35
1:112298763_T/C	DDX20	NON_SYNONYMOUS_CODING	3:125826058_T/C	ALDH1L1	NON_SYNONYMOUS_CODING	6.07	6.60E-06	4.62	8.08
2:179457146_G/A	TTN	NON_SYNONYMOUS_CODING	22:35695930_C/A	TOM1	NON_SYNONYMOUS_CODING	6.04	4.77E-06	2.72	7.22
1:45797504_C/G	MUTYH	NON_SYNONYMOUS_CODING	8:30982425_T/G	WRN	NON_SYNONYMOUS_CODING	6.00	9.62E-06	2.74	7.20
16:4855278_A/G	GLYR1	AA_modification:Phosphoserine	8:13259100_G/A	DLC1	NON_SYNONYMOUS_CODING	5.90	1.58E-05	3.57	7.45
11:45975129_C/T	PHF21A	NON_SYNONYMOUS_CODING	19:49458190_C/A	BAX	NON_SYNONYMOUS_CODING	5.81	1.55E-05	2.31	6.81
7:128490102_G/A	FLNC	NON_SYNONYMOUS_CODING	8:13357339_G/C	DLC1	NON_SYNONYMOUS_CODING	5.81	1.02E-05	6.60	8.67
11:236090_G/A	SIRT3	NON_SYNONYMOUS_CODING	4:110615838_C/T	CASP6	NON_SYNONYMOUS_CODING	5.79	9.91E-06	1.44	6.42
8:144993930_C/G	PLEC	NON_SYNONYMOUS_CODING	14:73422258_C/G	DCAF4	NON_SYNONYMOUS_CODING	5.79	1.08E-05	4.79	7.87
1:45224997_A/C	KIF2C	NON_SYNONYMOUS_CODING	2:108921032_C/T	SULT1C2	NON_SYNONYMOUS_CODING	5.70	2.30E-05	4.98	7.86
11:134252895_C/T	B3GAT1	NON_SYNONYMOUS_CODING	15:75012984_T/C	CYP11A1	NON_SYNONYMOUS_CODING	5.68	1.56E-05	1.81	6.47
6:116441645_C/G	COL10A1	NON_SYNONYMOUS_CODING	20:30072135_G/A	REM1	NON_SYNONYMOUS_CODING	5.63	1.24E-05	0.08	5.67
1:201016295_G/A	CACNA1S	NON_SYNONYMOUS_CODING	19:17000695_G/A	F2RL3	NON_SYNONYMOUS_CODING	5.61	2.89E-05	4.12	7.40
7:43351409_T/G	HECW1	NON_SYNONYMOUS_CODING	2:219294200_A/G	VIL1	NON_SYNONYMOUS_CODING	5.59	8.39E-07	0.62	5.87
15:67457334_A/G	SMAD3	NON_SYNONYMOUS_CODING	1:201052381_A/G	CACNA1S	NON_SYNONYMOUS_CODING	5.58	2.34E-05	0.60	5.84
10:53822300_A/G	PRKG1	NON_SYNONYMOUS_CODING	9:140007465_G/A	DPP7	NON_SYNONYMOUS_CODING	5.56	1.46E-05	3.30	7.00
2:225362477_C/T	CUL3	NON_SYNONYMOUS_CODING	9:120476787_C/G	TLR4	STOP_GAINED	5.56	2.96E-05	3.16	6.93

Table 1–5: Results from epistatic GWAS analysis of type II diabetes sequencing data. Genes form diabetes related pathways marked in red.

would have missed most of our top results. Likewise, an extension of that same limitation applies to our own method since we might be missing any pure epistatic interaction of order three above, a limitation that doesn't seem to be plausible to overcome without introducing prior biological knowledge of interacting loci (which is obviously not available at the moment).

#### 1.5.4 Power analysis

In order to assess our disease association power we performed extensive simulations. As it is often the case in this kind of analyses some simplifying but realistic assumptions were required to make simulations computationally tractable. We assumed that: i) the disease has a logistic risk model; ii) risk model has purely epistatic terms, i.e. that individual loci are assumed to have no marginal effect in disease risk; iii) disease prevalence is 8% according the well accepted prevalence for type II diabetes; iv) cofactors influencing disease such as population admixture, age and sex were perfectly reduced by the model, meaning that no residual effects remains after

correction; v) genome wide significance was established as  $0.05/(10^6)^2 = 5.10^{-14}$ ; and vi) that the study has an equal number of cases and controls.

Under these assumptions we calculated power by running 100 iterations for every model having a combination of: i) sample sizes ranging from 10,000 to 1,000,000, ii) log odds disease risk ranging from 0.1 to 5, and iii) allele frequencies of each of the two loci ranging from 1% to 20%. Results for some selected models are shown in Figure ??.

As expected these simulations show that very large sample sizes are required to find epistatic effects on rare variant loci (i.e. allele frequencies below 1%). Even if the risk factor is assumed to be over 2, 3, or 5, sample sizes requirements are expected to be over 400,000, 200,000 and 100,000 respectively (see Figure ??, top-left plot). For common variants (i.e. allele frequency of 5%) the sample sizes look more attainable in the near future, ranging from 10,000 to 100,000 samples when risk factors decrease from 5 to 0.6 (Figure ??, middle row, left plot). Finally, for relatively common variants (allele frequency of 10%) the sample sizes up to 100,000 would be required for risk factors of 0.3, which is still considered a high risk factor (Figure ??, bottom right plot).

Unsurprisingly, our power calculation results show that large sample sizes are required in order to find epistatic interactions even under assumptions of relatively common variants and relatively high risk factors. This highlights how elusive finding epistatic genome wide associations can be.

**Power increase.** Finally it is worth mentioning that our epistatic GWAS method is able to increase power by using additional information from a co-evolutionary

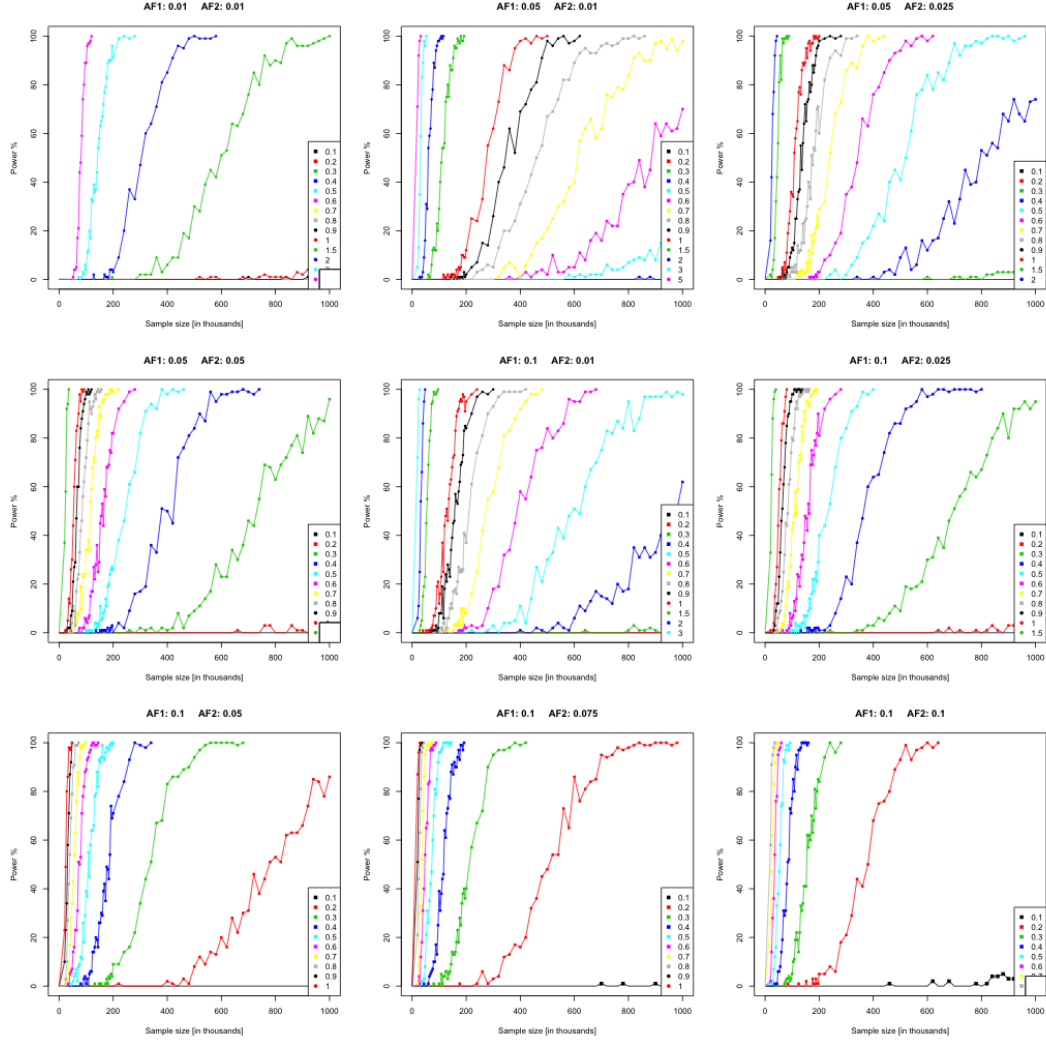


Figure 1–6: Power calculation using a case/control pure two-loci epistatic model for disease prevalences of 8% (type II diabetes). X-axis shows the number of samples in thousands (ranging from 10,000 to 1,000,000) and several joint risk values (from 0.1 to 5) are shown as different colors lines within each plot. Each plot represents a different allele frequencies combination (for each of the two loci), from top to bottom and left to right, the allele frequency value pairs are:  $[AF_1 = 1\%, AF_2 = 1\%]$ ,  $[AF_1 = 5\%, AF_2 = 1\%]$ ,  $[AF_1 = 5\%, AF_2 = 2.5\%]$ ,  $[AF_1 = 5\%, AF_2 = 5\%]$ ,  $[AF_1 = 10\%, AF_2 = 1\%]$ ,  $[AF_1 = 10\%, AF_2 = 2.5\%]$ ,  $[AF_1 = 10\%, AF_2 = 5\%]$ ,  $[AF_1 = 10\%, AF_2 = 7.5\%]$ ,  $[AF_1 = 10\%, AF_2 = 10\%]$

model. Using distribution estimates (see figure ??) we calculate that co-evolutionary information can increase Bayes Factor between  $10^2$  and  $10^4$ , which is roughly equivalent to reducing genome wide association p-value threshold by two orders of magnitude [?], thus shifting the power curves in Figure ?? to the left. Thus the net effect of adding information from a co-evolutionary model is to reduce sample size requirements roughly by half, which is quite remarkable given the little additional information being used by the model (only a multiple sequence alignment).

## 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 an epistatic GWAS 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. Our highly optimized algorithms are applicable to such large scale sequencing genomic studies and 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 indeed feasible to apply our methods to GWAS-scale datasets. Second, although larger cohorts are needed in order to find risk alleles that have lower frequencies and are not captured by this study, we show several suggestive associations of pairs of putatively interacting variants with type II diabetes.

The co-evolutionary model we propose in section ?? requires multiple sequence alignment and the corresponding phylogenetic tree. Intuitively, using a MSA with larger number of sequences should improve co-evolutionary model detection and other

co-evolutionary approaches indeed require very large MSA. Unfortunately, at this moment, MSA consisting of a large number of sequences are available only for a small fraction of all human proteins and they often consist of a mixture of ortholog and paralog sequences which may lead to biases in co-evolutionary models. Furthermore, both the phylogenetic 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 pre-calculation and “constant tree caching”) that allow the algorithm to be applied at genome-wide scale. Many multiple sequence alignments (such as Pfam) 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 ; and ii) we would add the problem of reconciling different phylogenetic trees from two proteins, which may lead to inconsistencies. For all these reasons we selected UCSC’s multi-100way [?], a genome wide multiple sequence alignment of 100 vertebrates which has a single genome wide phylogenetic tree. This MSA is expected to grow with the advent of projects like G10K [?] thus enabling more precise co-evolutionary predictions.

In order to further validate our co-evolutionary model in the context of human disease, we tested whether  $\ell_C$  scores can separate clinically relevant variants from ClinVar database [?] according to their clinical significance attribute (CLNSIG). Interestingly, variants categorized as “benign” (i.e. non-pathogenic) or “druggable” (i.e. affecting drug response) have higher scores (mean  $\ell_C$  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 mutations in other proteins.

**Comparison to other Co-Evolutionary methods.** There are several methodologies that can be used to predict putative interactions based on co-evolutionary theory. Nevertheless most methods are limited with respect to their applicability to GWAS-scale analyses:

*Phylogenetic tree similarity* can be used as a proxy for the co-evolution of interacting proteins. Computational methods use matrix alignment [?] which has demonstrated some degree of success. Unfortunately, such methods have two limiting factors: i) they require large (distinct) phylogenetic trees for each protein which are not be available for all proteins in the genome; and ii) solve an optimization problem requiring long computational times to match matrices (e.g. simulated annealing) for each putative pair or proteins.

*Correlation and Mutual information* based methods aim to detect changes on one of the interacting proteins that are compensated by mutations in the other [?, ?]. Although these methods are fast enough to be applicable to GWAS scale studies, they still have at least two limitations: i) they require large number of sequences in the multiple sequence alignment to overcome noise (as we mentioned large MSAs are not be available throughout the whole genome); and ii) there are affected by biases mainly caused by phylogenetic relationships and indirect correlations. There are methods based on mutual information than could perform some phylogenetic correction [?], nevertheless they still require a large number of sequences in the

MSA and are known to be biased by allele frequencies [?] which might limit their applicability for GWAS studies (particularly on non-common variants).

*Global models* are designed to disentangle direct interactions from indirect ones. Several methods have been proposed which rely on: i) estimating parameters of Boltzmann distributions [?, ?], ii) mean field approximations of Boltzmann distributions [?], iii) constrained optimizations for finding approximations of the inverses of large singular matrices [?], iv) marginalizing multidimensional extensions to mutual information [?]; or v) solving a Bayesian network model [?]. All these models are so computationally heavy that can only be applied to very small sets of proteins. Furthermore, in some of the respective papers the authors mention that it is computationally infeasible to apply them to a single pair of proteins if the length is over a some low number of amino acids (e.g. 60 [?] or 500 [?]). It is therefore not possible to apply any of these method to GWAS-scale studies at the moment.

Our method attempts to solve two of the main problems common to some of the aforementioned methods: i) the requirement of large number of sequences, ii) the phylogenetic bias, and iii) running time. These goals are achieved, at least partially, by using a well known Markov model of evolution.

**Comparison to other Epistatic GWAS methods.** Many reasons have been given in this manuscript and elsewhere to indicate why detecting epistasis this is a very difficult task, with the most commonly cited one being the enormous number of statistical tests required, consequently having a significant reduction of statistical power and an increase in computational resources required. Thus methods based on exhaustive search can be computationally infeasible for all but very low order

interactions analysis [?]. Their counterpart are *conditional search* methods [?] which are usually based on selecting the top single value associated variants and then performing epistatic analysis on a small subset. Unfortunately these methods will ignore pure epistatic interactions and only detect marginal ones [?, ?]. Since there is no biological indication on whether complex traits have marginal or pure epistatic effects [?, ?, ?], it might not be safe to rely exclusively on these type of methods when performing an association study.

Stochastic search based methods [?] show great potential, but as far as we known they have not yet produced any significant results most likely due to small sample sizes used in the respective publications. Some authors pointed out that some of most well known methods in this family may have difficulties with large number of samples [?].

Finally methods based machine learning have been proposed and applied with different degrees of success [?, ?, ?]. One of the main limitations of machine learning methods lies in the fact that the majority of them do not result in a statistical significance metric (such as p-value or Bayesian factor), thus researchers are often weary of conducting an expensive follow up studies based in results from machine learning algorithms. Another limitation is that many machine learning approaches do not for allow appropriate correction from population admixture and other cofactors.

The methodology we proposed in this paper is based on a well established statistical procedures (Logistic Regression) using gold standard corrections for known population cofactors (eigen-analysis) as well as other disease cofactors (such as age

and sex are known to affect risk of type II diabetes). Our method performs exhaustive search of second order interactions thus is capable of finding pure epistatic interactions pairs as well as marginal ones. Finally, we address the power limitations by using co-evolutionary results from a well established Markov model of evolution and combining them with our association analysis by means of a Bayesian model. This has the advantage that not only is being based on solid and well accepted theoretical grounds, but also it can increase statistical power and is capable of analysing large GWAS datasets that are becoming available now.

As a limitation, our methodology currently cannot analyse variants that are in complete LD. There is evidence suggesting that genes relatively close (in genomic coordinates) tend to have correlated in expression levels and likely to be in the same pathway or even to interact [?]. Nevertheless it seems that complete LD would be a result of “selective sweeps”, in which an allele giving significant fitness advantage becomes more frequent so rapidly that there is little recombination. These selective sweeps seem to be rare [?], thus this limitation might not be encountered often in practice.

**Future work.** We plan to extend our method to include context specific information by creating  $Q_2$  estimates for different protein domains. This would allow to obtain better estimates for well characterized protein interaction regions. Another line of work is to perform GWAS using kernel based statistics of multiple variants [?] 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. It has also been suggested that positive

selection might be used as an additional prior of epistatic interactions. Nevertheless a study comparing positive selection maps from 9 different methods [?] shows that only 14% of the regions have been identified in more than one study. Although there seems to be a lack of concordance at the moment, using positive selection estimates would be an interesting venue to explore in the future.