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

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

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

CHAPTER 1
Introduction

How does one's DNA influence their risk of getting a disease? Contrary to popular belief, your future health is not "hard wired" in your DNA. Only in a few diseases, referred as "Mendelian diseases", are there well known, almost certain, links between genetic mutations and disease susceptibility. For the majority of what are known as "complex traits", such as cancer or diabetes, genomic predisposition is subtle and, so far, not fully understood.

With the rapid decrease in the cost of DNA sequencing, the complete genome sequence of large cohorts of individuals can now be routinely obtained. This wealth of sequencing information is expected to ease the identification of genetic variations linked to complex traits. In this work, I investigate the analysis of genomic data in relation to complex diseases, which offers a number of important computational and statistical challenges. We tackle several steps necessary for the analysis of sequencing data and the identification of links to disease. Each step, which corresponds to a chapter in my thesis, is characterized by very different problems that need to be addressed.

i) The first step is to analyze large amounts of information generated by DNA sequencers to obtain a set of "genomic variants" present n each each individual. To address these big data processing problems, Chapter **??** shows how we designed a programming language (BigDataScript [5]), that simplifies the creation robust, scalable data pipelines.

ii) Once genomic variants are obtained, we need to prioritize and filter them to discern which variants should be considered "important" and which ones are likely to be less relevant. We created the SnpEff & SnpSift

[3, 4] packages that, using optimized algorithms, solve several annotation problems: a) standardizing the annotation process, b) calculating putative genetic effects, c) estimating genetic impact, d) adding several sources of genetic information, and e) facilitating variant filtering.

iii) Finally, we address the problem of finding associations between interacting genetic loci and disease. One of the main problems in GWAS, known as "missing heritability", is that most of the phenotypic variance attributed to genetic causes remains unexplained. Since interacting genetic loci (epistasis) have been pointed out as one of the possible causes of missing heritability, finding links between such interactions and disease has great significance in the field. We propose a methodology to increase the statistical power of this type of approaches by combining population-level genetic information with evolutionary information.

In a nutshell, this thesis addresses computational, analytical, algorithmic and methodological problems of transforming raw sequencing data into biological insight in the aetiology of complex disease. In the rest of this introduction we give the background that provides motivation for our research.

1.1 Epistasis

At the beginning of the 20^{th} century some deviations form classical Mendelian inheritance were characterized. William Bateson first described epistasis in 1907 [20] assessing a discrepancy between the prediction of segregation ratios assuming individual genes and the real outcome [18]. The term epistasis literally means "standing upon" was used to describe "characters" layered on top of other each other thus masking their expression. Reflecting this original definition, nowadays the term epistasis is used to describe an allele at one locus masks the expression of another allele at a different locus [6]. The way epistasis was used to describe the situation in which the actions of one locus mask the allelic effects of another locus, is an extension of dominance where a completely dominant alleles mask the effects of the recessive allele at the same locus. [2, 6].

The concept of epistasis is often interpreted as mutations in two genes producing a phenotype that is surprising considering the individual effect of each mutation and can point to functional relationships between genes and pathways [15]. Epistasis, can be used as a tool for understanding the genetic pathways' structure and function as well as evolutionary dynamics [18]. Some authors even relate the analysis of gene interaction patterns to the fundamentals of systems biology [18].

The term epistasis has expanded to describe many complex interactions among genetic loci [18]. Geneticists used epistasis to describe different things:

- Functional epistasis: The molecular interactions that proteins. Usually these interactions consist of proteins within the same pathway or of within a complex [18]

- Compositional epistasis: Describes the traditional usage of epistasis as described by Bateson (i.e. blocking of one allelic effect by an allele at another locus) [18].
- Statistical epistasis: This terminology is attributed to Fisher defined as a deviance from genetic additive effects, this essentially treats it as a residual term in genetic analysis [24].

Epistasis can be classified by the way a deviation of a double-mutant organism's phenotype differed from the expected neutral phenotype[15]. An interaction is known as “synergistic” or “synthetic” when the double mutant has a more extreme phenotype than expected When the phenotype is less severe than expected, then there is a “diminishing returns” or “alleviating” interaction, this is often attributed to gene products operating in series within the pathway. A typical example is a mutation in one gene impairing a whole pathway, thus masking the consequence of mutations in other genes of the same pathway. [15].

Often, the phenotype in human genetics is qualitative and dichotomous, indicating presence or absence of disease. [6]. Thus mathematical models calculating the joint action of more than one loci focus on the penetrance (the probability of developing disease given genotype). Assuming an allele is required at both loci in order to express the trait, the effect of allele A can only be observed when allele B is also present. This means that the effect at locus A appears masked by locus B and vice-versa [6], which is not precisely analogous to what Bateson descrived. In Bateson's definition, if factor B is epistatic to factor A, then factor A is not expected to be epistatic to factor B also. [6] Four mathematically different definitions of interaction have been used (namely Product, Additive, Log, and Min) [15], but even though some definitions yield identical results under some conditions, an alternative definition choice can lead to different consequences[15].

Defining interaction requires measuring phenotype and a neutrality. Neutrality function predicts the phenotype of an organism without interacting mutations. Fitness, is central phenotype measurement to many large-scale genetic interaction studies, it can be defined by population allele frequencies or using growth rates of microbial cultures [15]. Different measures of fitness can be used in epistasis: i) exponential growth rate of mutant strain respect to wild type ; ii) the increase in population in one wild-type generation; and iii) the relative number of progeny (in one wild-type generation) [15].

Genetic interaction studies have also differed in their choice of neutrality functions, generally using either a multiplicative or a minimum mathematical function. Multiplicative function predicts fitness to be the product of the corresponding single-mutant fitness values. This multiplicative function can be used with the three aforementioned fitness measures to obtain three different definitions of genetic interaction [15].

The "Min" definition of genetic interaction is simply the minimum neutrality of the expected results form non-interacting mutations (e..g the fitness of the less-fit mutant). All the above fitness measurement yield the same set of genetic interactions under this definition. For example if each mutation disrupts a distinct pathway limiting cell growth in a way that one mutation is substantially more limiting than the other, the double mutant might is expected have same result as the most-limiting single mutant [15].

It has been shown choice of definition can dramatically alter the resulting set of interactions [15]. To evaluate this the authors in [15] applied all four definitions to two studies providing quantitative growth-rate measurements of cell populations. They show that: i) additive and Log definitions have different biases; ii) Product and Log definitions are equivalent for deleterious mutations; iii) the product definition can reveal functional relationships missed by the Min definition; and iv) interaction networks from Min and Product definitions differ greatly. This leads to the question on which definition to use. By examining the deviation distribution of expected (double-mutant) phenotype from the observed phenotype they found that Product and Log definitions not only are the closest to the ideal, but also are practically equivalent when single mutants are deleterious [15].

The presence or knock-out of a gene are extreme aspects of "perturbation in a complex system", but there are no reasons to expect all forms of epistasis to follow this pattern [18]. When applied to quantitative traits, epistasis also describes a situation in which the phenotype cannot be predicted by the sum of its single-locus component [2]. Many epistatic QTL interactions have been detected in model organisms leading to the conclusion that epistasis makes a large contribution to the genetic regulation of complex traits [2].

1.1.1 Epistasis is ubiquitous

Epistasis is defined as departure from additive effects. Nevertheless, there is no reason to think that traits should be additive based on a purely biological perspective [25] since biology is riddled with non-linearity such as genetic networks exhibit binary states, ligand - receptors concentration having sigmoid-like response, concentration saturations of substrate - enzymes reactions, sharp transitions created by cooperative protein binding, the pathways constrained by rate-limiting inputs, etc. [25]. It has been asserted that epistatic effects are not isolated events, but ubiquitous [20] and probably inherent properties of biomolecular networks. This leads to think that epistasis in the classical sense may be ubiquitous, a thought which has been partially confirmed from mutational studies [18]. Genetic studies of synthetic traits, which occur only when multiple loci or pathways are all disrupted, in model organisms have identified instances of interacting genes revealing that epistasis may be pervasive [25]. Researchers found [18] that when looking for interactions induced by systematically over-expressing genes in Saccharomyces cerevisiae about 15% of studied genes induced growth defects with most over-expression not matching the phenotypes of individual deletions.

1.1.2 Epistasis examples: Non-human

Several genotype-phenotype patterns are known to be caused by epistasis, classic examples include [2]: coat colour in various animals, comb type in chickens, kernel colour in wheat, eye color in flies, and the h/h blood group (also known as Oh or the Bombay phenotype) in the ABO blood-group in humans.

Coat colour in mammals has been one of the most common examples. In pig, the dominant allele at the KIT locus confers white color coat and is dominant over all locus conferring darker color (melanocortin 1 receptor or MC1R). This can be determined in individuals with the recessive KIT genotype showing what was classically termed ‘dominant epistasis’, yielding a non-Mendelian segregation ratio of 9:4:3 (instead of 9:3:3:1) [2, 18].

Drosophila provides another classic example with eye color determination. Drosophila eye pigmentation scarlet, brown, and white is determined by the synthesis of two drosopterins: brown pigments (from tryptophan) and red pigments (from GTP) [**?**]. A mutation that prevents production of the brown pigment results in a fly with red eyes and a mutation preventing red pigment results in a fly with brown eyes. Flies with a mutation in the white gene, neither red nor brown pigment can be synthesized resulting in a fly with white eyes (regardless of the genotype at the brown or scarlet loci) [20].

Dozens of quantitative traits indicating strong epistasis in mouse and rat [?] by analysing a panel of chromosome substitution strains where the effects attributed to the donor chromosomes exceeds by a median eightfold the expected effect of the donor genome.

Genetic interaction have been study in a systematic and large-scale manner in Saccharomyces cerevisiae [?]. Analysis of quantitative traits loci (QTL) for transcripts levels in a two strain cross demonstrated epistatic interaction for 67% of studied pairs (first the strongest QTL was found and then the strongest remaining QTL conditional on the first genotype was selected) [?].

In a study comparing three Drosophila inbred lines (Drosophila melanogaster Genetic Reference Panel -DGRP) and a large outbred and intercross derived population [11]a set of candidate SNPs was selected by assesing allele frequency changes between the extremes of the distribution for each trait. The researchers found that the majority of these SNPs participated in at least one epistatic interaction [11]. Using this information from epistatic interacting loci they were able to infer networks affecting quantitative traits. [11].

1.1.3 Epistasis examples: Human

Few instances of epistasis in common human disease have been discovered and well-replicated so far, despite considerable efforts [25]. Although many instances of epistasis related to human disease have been published, with examples form coronary artery disease[?], diabetes[?], bipolar effective disorder[?] and autism [?]; some authors suspect there might be statistical features in the association studies because only a few have functional basis [18].

Perhaps the best examples are interactions involving at least one locus with a large effect such as HLA [25]. Two different interaction involving HLA alleles and ERAP have been discovered in GWAS from ankylosing spondylitis and psoriasis where the HLA alleles have odds ratio of 40.8 and 4.66 respectively [?]. In autoimmune disease multiple sclerosis a researchers found evidence of genetic interactions between two histocompatibility loci known to be associated with the disease (HLA-DRB5*0101 in DR2a and HLA-DRB1*1501 in DR2b) [?]. There was evidence of naturally occurring linkage disequilibrium which is suspected to be generated by strong epistasis [18]. In Type 1 diabetes HLA is assumed to act non-additively with all other risk alleles (HLA has have an effect of 5.5) [?]. in Hirschsprung's disease an interaction between RET and EDNRB was uncovered by a genome-wide linkage study (RET having a log-odds of 5.6). [?]

The ACE gene (angiotensin I converting enzyme) has an epistatic interaction with AGTR1 gene (angiotensin II type 1 receptor) gene, significantly increasing risk of myocardial infarction when the "D-allele" in ACE is present in patients carrying a particular AGTR1 allele [?].

Two different sets of interactions are assumed to be responsible for variation in triglyceride levels. Notably, the interactions depend on the patient's sex: in females the interactions involves ApoB and ApoE; and in males the interaction involves the ApoAI/CIII/AIV complex and low-density lipoprotein receptor [?]

Sickle-cell anemia is regarded as a Mendelian trait is modified by epistatic interactions evidenced by the fact that patients homozygous for two polymorphisms near the $G\gamma$ locus have only mild clinical symptoms [?].

Elevated blood serum cholesterol levels in humans is associated with an ApoE allele depending on the genotype at the LDLR (low density lipoprotein receptor) gene locus [?].

1.1.4 Epistasis and evolution

From an evolutionary perspective, some authors argue that the nonlinear epistatic interactions between polymorphic loci is the genetic basis of canalization (the robustness or ability of a population to produce the same phenotype regardless of environmental variability) and speciation [11].

It has also been pointed out that interactions have an important influence on evolutionary phenomena such as genetic divergence and affects the evolution of the structure of genetic systems [18] sine studies have shown that epistasis can have a limiting role on the possible paths that evolution can take [?]. This has been supported by analysis of complex gene regulation patterns in localized genomic regions [?]. For variety of organisms (such as yeast, Caenorhabditis, Drosophila, higher plants, and mammals) genes sharing expression patterns are more likely to be in proximity [?]. This evidence shows that regional controls of chromatin structure and expression may give rise to gene clusters by promoting their coregulation [17].

Theoretical grounds that date back to Fisher assert that when genes interact there is evolutionary pressure to promote their genetic linkage as a means of enhancing the coinheritance of favorable allelic combinations [?]. Under this assumption linkage can facilitate the maintenance of epistatic interactions and vice versa, thus explaining how some molecular evolution complexity [18].

1.1.5 Missing heritability

At the dawn of the "GWAS era" in 2002 it was hypothesised that there existed a large class of genetic models for which GWAS would fail, namely purely epistatic models conssiting of models with no additive or dominance variation at any of the susceptibility loci. Thus association case/control methods "will have no power if the loci are examined individually" [8]. Furthermore, it was mathematically shown that for such models maximizing the broad sense heritability (under some constraints) is equivalent maximizing the interaction variance [8].

In a seminal series of papers [25, 26] further mathemical prof of the link between epistasis and heritability was provided. Missing heritability arises by an overestimation of the denominator that happens when epistasis is ignored [25]. This overestimatio, called “phantom heritability”, was shown to inflate the denominator over 60% in Cohn’s disease, thus could accounting for up to 80% of the missing heritability [25]. Even though the prevailing view among geneticists is that interactions play at most a minor role in explaining missing heritability, their works showsthat simple (and plausible) models can give rise to substantial phantom heritability [25]

In moderately heritable complex diseases for which single-locus analyses have not accounted for the predicted genetic variance these epistatic models provide one possible explanation so it is worth pursuing a hypothesis of interacting loci [8].

1.1.6 Detecting Epistasis / interactions

Linkage disequilibria (LD) between close sites are the result of unrecombined chromosome blocks from common ancestry [**?**], nevertheless LD between widely separated sites suggests epistatic selection forces are at work [**?**, **?**, 12]. In an analysis using Yoruba population (from Ibadan, Nigeria) of the HapMap dataset patterns of LD were quantified and significance of overall disequilibrium per chromosome was evaluated of using randomization [12], showing an excess of associations in distant on all of the 22 autosomes. Although this is suggestive of epistasis, other hypothesis should not be ruled out: i) population admixture has been proposed to explain unusual patterns of long range LD [**?**] ii) recombination between distant chromosome blocks may not completely erase LD caused by drift even in a population at demographic equilibrium, iii) bottlenecks are particularly effective at generating LD iv) hitchhiking of linked sites with a positively-selected mutation can generate large haplotype blocks v) large inversion and other structural variation alter recombination patterns thus causing LD over unusually large regions [**?**].

Under the assumption that long range LD can hint physical protein interactions the authors of LDGIdb [22] created a catalog of over $600,000$ pairs of SNPs showing strong long-range linkage disequilibrium, i.e. pairs of SNP pairs that were either located in different chromosomes or in different LD blocks and had $r^2 \geq 0.8$ [22]. However these simple approaches may be of little utility because of technical issues that must be taken into account when performing such association, since commonly used measures of LD (such as r^2 and D') are known to give rise to large linkage when sampling minor allele frequencies (MAF) near 0 [12]. A better alternative is to measure the probability that a large value of the disequilibrium D is observed if there is no association further refined by conditioning on the sampled allele frequencies (at the two loci), which has the analytical advantage to asymptotically converge to a Fisher's exact test [12].

It is possible to implicitly test the over / under-representation of allele pairs in a given population, i.e. analysis of imbalanced allele pair frequencies [1] The underlying theory is that such allele pairs are under Dobzhansky-Muller incompatibilities which establishes a fitness bias favouring of individuals that inherit over-represented allele combination [1].

The authors in [1] studying a population of 2,002 mice in family trios. They performed a χ^2 test correcting by confounding factors (such as expected frequencies, family structure and allelic drift) based on inspecting $3x3$ contingency tables of all possible two-locus allele combinations. They claim that using their method it is possible to detect more interactions than using independent markers and as a result they were able to identify 168 LD block pairs with imbalanced alleles [1].

By exploiting the intense selective pressures imposed by the process of inbreeding mice populations it can be expected that clusters of functionally related genes are likely to be selected for coadapted allelic combinations in genes that influence fitness and survival. This hypothesis would result in regions of linkage disequilibrium (LD) among inbred strain genomes that should occur more often than expected by chance [17]. In a study using 60 inbred mouse strains [17], the authors study LD using permutation tests showing that extreme patterns of LD give rise to scale-free networks architectures. Further pathway analysis identifies biological functions underlying several of these networks, hinting that selective factors acting to generate LD networks during inbreeding are a reflect interaction of functionally [17].

1.1.7 Epistasis & GWAS

In recent years there have been a growing number of GWAs, most of them have used a single-locus analysis strategy, in which each variant is tested individually for association with a specific phenotype [7] It may be inadequate to describe complex disease the relationship between genotype and phenotype by simply summing the modest effects from several contributing loci [8]. The extent to which epistasis is involved in complex traits is not known so we cannot assume that epistasis will be found for every trait in every population. [2] However epistasis has been overlooked and that it should to be routinely explored in complex trait studies. [2]. This is particularly important for researchers of moderately heritable complex diseases for which locus-by-locus analyses have not accounted for the predicted genetic variance should pursue a hypothesis of epistatic loci [8] that owing to their interaction, might not be identified by using standard single-locus tests [7]. It is also hoped that detecting such interactions will allow to elucidate biological pathways that underpin disease [7].

Failure to detect epistasis does not rule out it's presence [25]. On the one hand the reason why complex human phenotypes fail to find evidence for epistatic interactions may simply be that analytic methods inherently exclude epistasis [8]. On the other hand, individual interaction effects are expected to be much smaller than linear effects, and the sample size required to detect an effect scales inversely with the square of the effect size. As an example provided by [25] consider two variants with frequency 20% and increasing risk by 1.3 fold, which is a large effect. In such case, assuming 50% power, significance level 5×10^{-8} and equal number of cases and controls; the sample size required for single loci analysis would be $4,900$. In comparisson, the sample size required to detect pairwise interaction between those two variants using the same power and an appropriately corrected significance level is roughly $450,000$, so a researcher studying $100,000$ samples would discover all single acting loci but would find little evidence of epistatic interactions, which may be the reason why geneticists that have tested for pairwise epistasis between loci have found few significant signals [25]. It should be noted that even though GWAs involving over $500,000$ samples are not available at the moment, studies using sample sizes in this order are expected to become available within the next couple of years.

Existing approaches for searching interactions can be grouped into five broad categories. [14]:

1. **Exhaustive search** methods use classical statistics such as the Pearson's χ^2 test or logistic regression that are natural extensions of methods commonly used for single-locus tests in GWAS. It should be noted

that the number of tests necessary to evaluate all two-way, three-way and four-way interactions for 30-60 candidate loci, has a range similar to the number of tests suggested for a single GWAS, thus searching for n-way interactions among all the markers would be impracticable. [8]. Approaches developed to detect epistasis: combinatorial partitioning method [?], restricted partitioning method [?], multifactor-dimensionality reduction [?], multivariate adaptive regression spline [?], logistic regression methods and backward genotype-trait association (BGTA)[?]. Unfortunately even though many of these look promising, many of them have only been tested on small data sets [23]. Furthermore, methods based on brute-force searches such as (combinatorial partitioning method and multifactor-dimensionality reduction) are impractical for large data sets [23]. Nevertheless it was shown [14] that it can be feasible to perform GWAS level analysis in some cases and that simple methods explicitly considering interactions can actually achieve reasonably high power with realistic sample sizes under different interaction models with some marginal effects, even after adjustments of multiple testing using the Bonferroni correction.

2. **Linkage disequilibrium** methods use patterns in disease population under two-locus disease models [24] association can be estimated assuming that deviation of the penetrance from independence at an individual locus creates linkage disequilibrium (LD) even if two loci are unlinked. [24] Under the assumptions that two disease-susceptibility loci are in Hardy-Weinberg equilibrium (HWE) and are unlinked they show that in the presence of interaction the two loci will be in linkage disequilibrium in the disease population [24]. They develop a test statistic for detection deviations from LD, intuitively they test interaction by comparing the

difference in the LD levels between two unlinked loci between cases and controls [24]. Since the frequency of a haplotype is equal to the product of the frequencies of the component alleles of the haplotype, thus in the absence of interaction the proportion of individuals carrying a haplotype in the disease population is equal to the product of the proportions of individuals carrying the component alleles of the haplotype in the disease population [24]. They further show that under the null hypothesis, this test statistic asymptotically converges to a central χ^2 distribution. In their power comparison simulations they show that in general this LD-based test statistic has much smaller p-values than those of logistic regression analysis [24] conlcuding that their test has much higher power than alternatives such as logistic regression [24].

3. **Stochastic search** methods use sampling to infer whether a locus is a risk locus, jointly affects disease, or a background locus. A Bayesian approach for genome-wide case-control studies denoted 'bayesian epistasis association mapping' (BEAM) [23] is a representative example of this type of methods. BEAM treats the disease-associated markers and their interactions via a bayesian partitioning model and computes the posterior probability that each marker (using Markov chain Monte Carlo). The method uses predictors in the form of genetic marker loci divided into three groups: i) markers not associated with disease, ii) markers individually contributing to disease risk, and iii) markers that interact [23]. Membership of each marker in each of the three groups is defined by the prior distributions. Given a prior distributions for regression coefficients values given by group membership, a posterior distribution can be generated using MCMC simulation. [7]. At the end, it uses a special statistic (B-Statistic) to infer significance from the samples in MCMC.

Although it avoids computing all interactions but theoretically could find high-order interactions. Since the method was originally designed for genotypes markers, its power can be hampered by allele frequency discrepancies between unobserved disease loci and associated genotyped marker/s[23]. This is a common problem when using indirect markers and the authors show that in an extreme case when the MAF discrepancy was maximized all tested methods had little power to detect interaction associations [23]. In the original paper, the authors apply BEAM to a data set containing $116,204$ SNPs genotyped for 96 affected individuals and 50 controls. for an association study of age-related macular degeneration (AMD). Unfortunately BEAM did not find any significant interactions [23] most likely due to the small sample size. Runtime and power are primarily determined by the number of MCMC rounds with a suggested number of MCMC iteration as the quadratic of the number of SNPs, thus limiting the applicability [14]. So it cannot be applied to large GWAS studies because computational limitations make it unsuitable to handle over $500,000$ markers with sample sizes of $5,000$ or more individuals which are now commonly sequenced or genotyped [7].

4. **Conditional search** methods usually perform analyses in stages [14]. A small subset of significant loci is identified in the first stage, typically using single locus association statistics. Then this subset is mined using multi-locus association using an exhaustive method. A well known approach in this category is “stepwise logistic regression” which works as follows: (i) all markers are individually tested for associations with disease and ranked; (ii) the top (usually 10%) are selected, 3) all two-way (or three-way) interactions are tested and ranked for associations. Even this stepwise approach can become computationally intractable

for high-order interactions [23]. Analysis of stepwise logistic regression approach to identify two-way and three-way interactions demonstrated that searching for interactions in genome-wide association mapping can be more fruitful than traditional approaches that exclusively focus on marginal effects [23]. As a counter argument for stepwise logistic regression, we should take into account that in the presence of epistasis the effect of one locus is altered or masked by another locus, thus power to detect the first locus is likely to be reduced and the joint effects will be hindered by their interaction [6]. Methods based on conditional search can greatly reduce the computational burden by a couple of orders of magnitude, but with the risk of missing markers with small marginal effect [14].

5. **Machine learning** approaches can also be used to infer epistasis. A popular approach uses Random Forests [14] or other regression trees partitioning approaches based on classification. In this context, trees are constructed using rules based on the values of a predictor variable such as a SNP to differentiate observations such as case-control status [7]. A popular rule selection mechanism is to use the variable that maximizes the reduction in Gini impurity [**?**] at each node (intuitively, when child nodes have lower impurity from a split based on an attribute each child node will have purer classification). Random forests are constructed by drawing samples with replacement from the original sample. A classification tree is created for each bootstrap sample, but only a random subset of the possible predictor variables is considered. This results in a 'forest' of trees have been trained on a particular sample of observations. [7] Instead of trying to create a monolithic learner, this type

of methods called “ensemble systems” attempt to create many heterogeneous “weak” (or simple) learners. The outcomes of these heterogeneous systems are combined to create an improved model [14].

In an extension of AdaBoost algorithm, the authors incorporates an importance score based on Gini impurity to select candidate SNP [14] in a way that genotype frequencies from the two classes (case and control) are expected to be more different. Decision trees are usually built with binary splits, but since genotype data can be $0, 1, 2$, they also extended their method to create a ternary split AdaBoost draws bootstrap samples to increase the power of a weak learner by weighting the individuals when bootstrapping. So when a weak learner misclassifies an individual, the weight of that individual is increased, as a consequence hard to classify individuals are more likely to be included in future bootstrap samples. The ensemble votes by weighting weak learner instances by training set accuracy. [14]. Using simulation, they claim that their method outperforms not only a similar ensemble approaches, but also several statistical methods, although thei mention performance degradation when the risk allele frequency is low [14].

1.1.8 Epistasisi GWAS: Power issues

We have seen that, if the true genetic model underlying a disease is purely epistatic, with no additive or dominance variation at any of the susceptibility loci, then association methods analyzing one locus at a time will have no power to detect the loci. [8] First, we expect that, with a sufficient number of contributing loci, purely epistatic interactions could account for virtually all the variation in affection status for diseases with any prevalence [8] Of course, there are subclasses of purely epistatic models (providing no marginal evidence for the involvement of any single locus) for which, in addition, no two, three, or L1 loci jointly give evidence of involvement in the disorder. This leads to the concern that even assessment of all two-, three-, and (L1)-way interactions among candidate loci may be insufficient for detection of the contributing loci. [8] The restriction on maximum heritabilities in these models is most easily seen by examining L-locus models for which no collection of L 1 loci shows marginal deviations. [8]

A small number of recent studies have explored this idea for the genome-level identification of epistatic interactions: if a large number of individuals is genotyped at a large number of genomic positions, it becomes possible to test all allele pairs for overand underrepresentation in that population [18-20]. [1] However, even though some methodological progress has been made [18], previous studies could hardly identify a significant number of interactions. The main obstacle is the humongous number of statistical hypotheses tested when comparing all markers in a genome against all markers. [1]

QTL: We present FastEpistasis, an efficient parallel solution extending the PLINK epistasis module, designed to test for epistasis effects when analyzing continuous phenotypes. [19] FastEpistasis is capable of testing the association of a continuous trait with all single nucleotide polymorphism (SNP) pairs from 500 000 SNPs, totaling 125 billion tests, in a population of 5000 individuals in 29, 4 or 0.5 days using 8, 64 or 512 processors. [19] It tests epistatic effects in the normal linear regression of a quantitative response on marginal effects of each SNP and an interaction effect of the SNP pair, where SNPs are coded as additive effects, taking values 0,1 or 2. The test for epistasis reduces to testing whether the interaction term is significantly different from zero. [19] The computations are based on applying the QR decomposition to derive least squares estimates of the interaction coefficient and its standard error. [19]

1.1.9 Epistatic GWAS

Genome wide association studies have traditionally focused on single variants or nearby groups of variants. An often cited reason for the lack of discovery of high impact risk factors in complex disease is that these models ignore loci interactions [7] which have recently been pointed out as a potential solution for the "missing heritability" problem [25, 26]. With interactions being so ubiquitous in cell function, one may wonder why they have been so neglected by GWAS. There are several reasons: i) models using interactions are much more complex [10] and by definition non-linear, ii) information on which proteins interacts with which other proteins is incomplete [21], iii) in the cases where there protein-protein interaction information is available, precise interacting sites are rarely known [21]. Taking into account the last two items, we need to explore all possible loci combinations, thus the number of N order interactions grows as $O(M^N)$ where M is the number of variants [9]. This requires exponentially more computational power than single loci models. This also severely reduces statistical power, which translates into requiring larger cohort, thus increasing sample collection and sequencing costs [9].

In Chapter **??** we develop a computationally tractable model for analysing putative interaction of pairs of variants from GWAS involving large case / control cohorts of complex disease. Our model is based on analysing cross-species multiple sequence alignments using a co-evolutionary model in order to obtain informative interaction prior probabilities that can be combined to perform GWAS analysis of pairs of non-synonymous variants that may interact.

The definition of epistasis from a statistical perspective is a "departure from a linear model" [7]. This means that in a logistic regression model the input for sample s includes terms with each of the genotypes at loci i and j), as well as an "interaction term" $g_{s,i} \cdot g_{s,j}$ [6].

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

where d_s is disease status, $\phi(\cdot)$ is the sigmoid function, $c_{s,1}, c_{s,2}, ...$ are covariates for sample s.

Models involving interactions between more than two variants can be defined similarly, but require more parameters and extremely large samples are required to accurately fit them.

Several families of approaches for epistatic GWAS exist. Here we mention a few:

- Allele frequency: In [1], an analysis of imbalanced allele pair frequencies is performed under the assumption that an implicit test for fitness can be achieved looking for over/under-represented allele pairs in a given population. In another study [24] the authors infer that interactions can create LD in disease population under two-loci model, then they show how LD-based p-values can uncover interaction and sometimes (in their simulations) outperform logistic regression tests.
- Bayesian model: In [23], a "Bayesian partitioning model" is used by providing Dirichlet prior distributions for each partition and computing posterior probabilities using Markov chain Monte Carlo (MCMC) algorithms. The methodology first test individual makers and picks only the top 10% to further investigate for epistasis, because it is prohibitive to test all loci.

- Machine learning: From a machine learning point of view, finding interacting variants is simply an *“optimisation procedure is to find a set of parameters that allows the machine-learning model to most accurately predict class membership (e.g. affected vs unaffected)”* [16]. Several approaches have emerged to tackle the “interaction problem” and used a variety of different techniques [13, 16] , such as neural networks, cellular automata, random forests, multifactor dimensionality reduction, support vector machines, etc.

Although all these models have advantages under some assumptions, none of them seems to be a “clear winner” over the rest [7]. All of these models suffer from the increase in number of tests that need to be performed, which raises two issues: i) multiple testing, which is often resolved by stringent significance threshold, and ii) computational feasibility, which is solved by efficient algorithms, parallelization, and heuristic approaches to quickly discard uninformative loci combinations. So far, no method for epistatic GWAS has been widely adopted and there is need of different approaches to be explored. In Chapter **??** we propose an approach to combine co-evolutionary models and GWAS epistasis of pairs of putatively interacting loci.

References

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

[2] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[3] 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.

[4] 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.

[5] Pablo Cingolani, Rob Sladek, and Mathieu Blanchette. Bigdatascript: a scripting language for data pipelines. *Bioinformatics*, 31(1):10–16, 2015.

[6] 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.

[7] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[8] 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.

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

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

[11] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert RH Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, et al. Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proceedings of the National Academy of Sciences*, 109(39):15553–15559, 2012.

[12] Evan Koch, Mickey Ristroph, and Mark Kirkpatrick. Long range linkage disequilibrium across the human genome. *PloS one*, 8(12):e80754, 2013.

[13] 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.

[14] Jing Li, Benjamin Horstman, and Yixuan Chen. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics*, 27(13):i222–i229, 2011.

[15] 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.

[16] 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.

[17] Petko M Petkov, Joel H Graber, Gary A Churchill, Keith DiPetrillo, Benjamin L King, and Kenneth Paigen. Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS genetics*, 1(3):e33, 2005.

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

[19] Thierry Schüpbach, Ioannis Xenarios, Sven Bergmann, and Karen Kapur. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–1469, 2010.

[20] Anna L Tyler, Folkert W Asselbergs, Scott M Williams, and Jason H Moore. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, 31(2):220–227, 2009.

[21] 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.

[22] Ming-Chih Wang, Feng-Chi Chen, Yen-Zho Chen, Yao-Ting Huang, and Trees-Juen Chuang. Ldgidb: a database of gene interactions inferred from long-range strong linkage disequilibrium between pairs of snps. *BMC research notes*, 5(1):212, 2012.

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

[24] 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.

[25] 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.

[26] 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.