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

In this section we introduced the basic concepts and methodologies used in GWAS. Although fairly mature, there is still heavy research and continuous improvement on GWAS statistical methods. Not only it is well known that traditional (i.e. single marker) GWAS methods fail under non-additive models [8], but also variants so far discovered using these methods do not account for all the expected phenotypic variance attributed to genetic causes (i.e. missing heritability). As other authors pointed out [7, 25, 26], this might be because we need to look for epistatic variants which are not taken into account using these methods. In the next section, and in Chapter **??**, we cover the topic of epistatic GWAS analysis.

1.1.1 Historical perspective

William Bateson first described epistasis in 1907.(2) Like pleiotropy, this concept was developed to explain deviations from Mendelian inheritance [20] The term literally means "standing upon", and Bateson used it to describe characters that were layered on top of other characters thereby masking their expression. [20] The commonly used definition of epistasis–an allele at one locus masks the expression of an allele at another locus–reflects this original definition. [20]

The term 'epistasis' was initially used in the context of Mendelian inheritance; environmental effects are relatively unimportant for Mendelian traits, so Ii individuals can be clearly assigned to one of a limited number of classes according to their phenotype. Here, epistasis was used to describe the situation in which the actions of one locus mask the allelic effects of another locus, in the same way that completely dominant alleles mask the effects of the recessive allele at the same locus. [2]

The term 'epistatic' was first used in 1909 by Bateson (1) to describe a masking effect whereby a variant or allele at one locus (denoted at that time as an 'allelomorphic pair') prevents the variant at another locus from manifesting its effect. [6] This was seen as an extension of the concept of dominance. There are, however, some problems with this definition, particularly when applied to binary traits. In human genetics, the phenotype of interest is often qualitative and usually dichotomous, indicating presence or absence of disease. [6] Mathematical models for the joint action of two or more loci usually focus on the penetrance, the probability of developing disease given genotype. [6] Suppose that a predisposing allele is required at both loci in order to exhibit the trait, i.e. one or more copies of both allele A and allele B are required. Then, when the effects of both loci are considered, we obtain the penetrance table shown in Table 2. In this table, the effect of allele A can only be observed when allele B is also present: without the presence of B, the effect of A is not observable. The effect at locus A would appear to be 'masked' by that at locus B. [6] This leads to a situation that is not precisely analogous to that described by Bateson (1). In Bateson's (1) definition, it is clear that if factor B is epistatic to factor A, we do not expect factor A to also be epistatic to factor B. [6] Table 3 is usually assumed to correspond to a situation in which the biological pathways involved in disease influenced by the two loci are at some level separate or independent (5). [6]

Epistasis, or interactions between genes, has long been recognized as fundamentally important to understanding the structure and function of genetic pathways and the evolutionary dynamics of complex genetic systems. [18] It has been approximately 100 years since William Bateson invented the term 'epistasis' to describe the discrepancy between the prediction of segregation ratios based on the action of individual genes and the actual outcome of a

dihybrid cross1 [18] The use of the term epistasis has since expanded to describe nearly any set of complex interactions among genetic loci [18] Over the years geneticists have used epistasis to describe three distinct things: the functional relationship between genes, the genetic ordering of regulatory pathways and the quantitative differences of allele-specific effects [18] Over the years the disparate needs of geneticists have led to a plethora of differently nuanced meanings for the term epistasis, all of which involve gene interactions at various levels [18] 'Functional epistasis' addresses the molecular interactions that proteins (and other genetic elements) have with one another, whether these interactions consist of proteins that operate within the same pathway or of proteins that directly complex with one another18 [18] 'Compositional epistasis' is a new term that is intended to describe the traditional usage of epistasis as the blocking of one allelic effect by an allele at another locus. [18] 'statistical epistasis' is the usage of epistasis that is attributed to Fisher (BOX 1), in which the average deviation of combinations of alleles at different loci is estimated over all other genotypes present within a population. [18]

It should be apparent that the global analysis of geneinteraction patterns bears a striking resemblance to what is now called systems biology [18]

often been defined as a deviance from genetic additive effects, which is essentially treated as a residual term in genetic analysis and leads to low power in detecting the presence of interacting effects [24]

Following the identification of several disease-associated polymorphisms by genome-wide association (GWA) analysis, interest is now focusing on the detection of effects that, owing to their interaction with other genetic or environmental factors, might not be identified by using standard single-locus tests [7] ...it is hoped that detecting interactions between loci will allow us to elucidate the biological and biochemical pathways that underpin disease. [7] In

recent years, the field has been revolutionized by the success of genome-wide association (GWA) studies1-5. Most of these studies have used a single-locus analysis strategy, in which each variant is tested individually for association with a specific phenotype [7] However, a reason that is often cited for the lack of success in genetic studies of complex disease6,7 is the existence of interactions between loci. [7] If a genetic factor functions primarily through a complex mechanism that involves multiple other genes and, possibly, environmental factors, the effect might be missed if the gene is examined in isolation without allowing for its potential interactions with these other unknown factors. [7]

1.1.2 Definition

In this review, we provide a historical background to the study of epistatic interaction effects and point out the differences between a number of commonly used definitions of epistasis [6] Sometimes mutations in two genes produce a phenotype that is surprising in light of each mutation's individual effects. This phenomenon, which defines genetic interaction, can reveal functional relationships between genes and pathways. [15] Recent studies have used four mathematically distinct definitions of genetic interaction (here termed Product, Additive, Log, and Min). Whether this choice holds practical consequences has not been clear, because the definitions yield identical results under some condition [15] Here, we show that the choice among alternative definitions can have profound consequences. [15]

A quantitative genetic interaction definition has two components: a quantitative phenotypic measure and a neutrality function that predicts the phenotype of an organism carrying two noninteracting mutations. Interaction is then defined by deviation of a double-mutant organism's phenotype from the expected neutral phenotype [15] A double mutant with a more extreme phenotype than expected defines a synergistic (or synthetic) interaction between

the corresponding mutations (synthetic lethality, in the extreme case). [15] Alleviating or "diminishing returns" interactions, in which the double-mutant phenotype is less severe than expected, often result when gene products operate in concert or in series within the same pathway. Alleviating interactions arise, for example, when a mutation in one gene impairs the function of a whole pathway, thereby masking the consequence of mutations in additional members of that pathway. [15] One class of phenotype, fitness, has been central to many large-scale genetic interaction studies. Although fitness was originally measured in terms of population allele frequencies (1, 22, 23), it can also be measured by using growth rates of isogenic microbial cultures. [15] Genetic interaction studies have used different measures of fitness, including: (i) the exponential growth rate of the mutant strain relative to that of wild type (4, 9, 15, 19) (the relative-growthrate measure); (ii) the increase in mutant population relative to wild type in one wild-type generation (the relative-population measure) (6); and (iii) the number of progeny per mutant organism relative to the number of progeny for wild type in one wild-type generation (the relative-progeny measure) (24) [15] Genetic interaction studies have also differed in their choice of neutrality functions, generally using either a multiplicative or a minimum mathematical function. [15] The multiplicative function, which was originally applied to fitness measures defined in terms of allele frequencies, predicts double-mutant fitness to be the product of the corresponding single-mutant fitness values. The multiplicative function can be combined with each of the three fitness measures above to yield three distinct definitions of genetic interaction (4, 6, 15, 19, 24). [15] A fourth (Min) definition of genetic interaction results from the minimum neutrality function, under which noninteracting mutations are expected to yield the fitness of the

less-fit single mutant. Each fitness measure above yields an identical set of genetic interactions under this function. A hypothetical example illustrates one rationale for the Min definition: Two single mutations each disrupt a distinct cellular pathway that limits cell growth, such that one of these mutations is substantially more limiting than the other. The double mutant might then be expected to exhibit the phenotype of the most-limiting single mutant. [15] It has not been clear whether the choice of genetic interaction definition has any practical consequences. To evaluate the impact of definition choice, we applied each of the four definitions in turn to two reference studies. [15] Here, we show that the choice of definition can dramatically alter the resulting set of genetic interactions and the extent to which they correspond to shared gene function. [15] For a gene pair (x, y), we refer to the fitness of the two single mutants and the double mutant, respectively, as Wx, Wy, and Wxy. [15] The neutrality function E(Wxy), predicting double-mutant fitness for a strain with mutations in noninteracting genes x and y, is defined differently under the Min, Product, Log, and Additive [15]

DATASET: To evaluate the impact of definition choice, we applied each of the four definitions in turn to two reference studies, St. Onge et al. (19) (Study S) and Jasnos and Korona (6) (Study J), both providing quantitative growth-rate measurements of isogenic wild-type and singleand double-mutant cell populations. [15] RESULTS: The Choice of Genetic Interaction Definition Matters: [15] Additive and Log Definitions Demonstrate Different Biases: However, we had observed that interaction strength had a significant positive bias (under all definitions) for pairs involving mutations with extreme fitness effects. [15] Product and Log Definitions Are Equivalent for Deleterious Mutations: [15] The Product Definition Reveals Functional Relationships Missed by

the Min Definition. [15] Genetic Interaction Networks from Min and Product Definitions Differ Greatly. [15]

WHICH DEFINITION TO USE?: We examined the distribution of , the deviation of the expected double-mutant phenotype from the observed double mutant phenotype, and found the Product and Log definitions to be closest to this ideal in general. Additionally, we showed that the Log and Product definitions are practically equivalent when both single mutants are deleterious. [15]

1.1.3 Epistasis in quantitative traits

In the case of QUANTITATIVE TRAITS, epistasis describes the general situation in which the phenotype of a given genotype cannot be predicted by the sum of its component single-locus effects1 [2] Epistatic QTL-mapping studies in model organisms have detected many new interactions and have therefore concluded that epistasis makes a large contribution to the genetic regulation of complex traits. [2] Complex synthetic interactions. : There is no reason to expect all forms of epistasis to be revealed simply by the absence of a gene, which is certainly an extreme approach to perturbing complex systems. For example, Kroll et al.35 devised a method for looking for interactions that are induced after systematically overexpressing genes. Using this approach, sopko et al.36 found that, when overexpressed in Saccharomyces cerevisiae, about 15% of a set of 5,280 yeast genes induced a growth defect, with most of the overexpression effects not matching the phenotypes of their corresponding deletions. [18]

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]

1.1.4 Epistasis is ubiquitous

From mutational studies we know that epistasis in the classical sense is ubiquitous because genes interact in hierarchical systems to generate biological function. [18] From a biological standpoint, there is no a priori reason to expect that traits should be additive. Biology is filled with nonlinearity: The saturation of enzymes with substrate concentration and receptors with ligand concentration yields sigmoid response curves; cooperative binding of proteins gives rise to sharp transitions; the outputs of pathways are constrained by rate-limiting inputs; and genetic networks exhibit bistable states. [25] Genetic studies in model organisms have long identified specific instances of interacting genes (17). Important examples include synthetic traits (e.g., 18), which occur only when multiple loci or pathways are all disrupted. [25] Studies have begun to reveal that epistasis is pervasive. [25] We assert that epistasis and pleiotropy are not isolated occurrences, but ubiquitous and inherent properties of biomolecular networks. [20]

1.1.5 Epistasis examples: Non-human

Extensive work on the control of qualitative genetic variation has highlighted the biological importance of epistasis at a locus-by-locus' level. On the basis of this work, several classic genotype-phenotype patterns that are caused

by epistasis such as comb type in chickens, coat colour in various animals, the BOMBAY PHENOTYPE in the ABO blood-group system in humans and kernel colour in wheat [2] In the case of quantitative genetic variation, several or many genes of largely unknown function combine with environmental influences to control trait variation. This is the case for many complex traits that are of medical relevance in humans or of economic importance in plants and livestock. [2] A clear example of this can be seen [in Fig A] which the dominant allele (I) at the KIT locus, which confers white-coat colour in the pig, is dominant over all alleles at the MC1R locus (E), which confer a darker coat colour. The effects of the various alleles at the E locus can only be determined in individuals with the recessive genotype ii at the I locus. This example was classically termed 'dominant epistasis', which gives a segregation ratio of 12:3:1 for white:black:brown, respectively [2] Table 1. Example of phenotypes (e.g. hair colour) obtained from different genotypes at two loci interacting epistatically, under Bateson's (1909) definition of epistasis [6] Coat colour variation in mammals has long been is one of the most fruitful examples in the study of the relationship between genotype and phenotype. ... epistasis arises when the effects of alleles at one locus are blocked by the presence of a specific allele at another locus. For example, a cross between agouti and extension (now called the melanocortin 1 receptor or Mc1r) double heterozygotes (AaEa) yields the non-Mendelian segregation ratio of 9:4:3 (instead of 9:3:3:1) [18] In the yeast Saccharomyces cerevisiae, Brem et al. (19) analyzed as quantitative traits the levels of gene transcripts in segregants of a cross between two strains. For each transcript, they found the strongest quantitative trait locus (QTL) in the cross and then, conditional on the genotype at this locus, identified the strongest remaining QTL. In 67% of cases, these two QTLs demonstrated epistatic interactions. In bacteria, Khan et al. (20) and

Chou et al. (21) have recently demonstrated clear epistasis among collections of five mutations that increase growth rate. [25] In mouse and rat, Shao et al. (22) analyzed a panel of chromosome substitution strains, with each strain carrying a different chromosome from a donor strain on a common recipient genetic background. For dozens of quantitative traits, the sum of the effect attributable to the individual donor chromosomes far exceeds (median eight-fold) the total effect of the donor genome, indicating strong epistasis. [25] An example in insects is the abnormal-abdomen phenotype in Drosophila mercatorum (DeSalle and Templeton 1986; Hollocher et al. 1992; Hollocher and Templeton 1994). [8] The study of genetic interaction has become increasingly systematic and large-scale, especially in the yeast Saccharomyces cerevisiae (6, 8-21). [15] Eye color determination in Drosophila provides a classic example. The genes scarlet, brown, and white, play major roles in a simplified model of Drosophila eye pigmentation. Eye pigmentation in Drosophila requires the synthesis and deposition of both drosopterins, red pigments synthesized from GTP, and ommochromes, brown pigments synthesized from tryptophan. A mutation in brown prevents production of the bright red pigment resulting in a fly with brown eyes, and a mutation in scarlet prevents production of the brown pigment resulting in a fly with bright red eyes. In a fly with a mutation in the white gene, neither pigment can be produced, and the fly will have white eyes regardless of the genotype at the brown or scarlet loci. In this example the white gene is epistatic to brown and scarlet. A mutant genotype at the white locus masks the genotypes at the other loci. [20] Evidence from inbred strains of mice indicates that a quarter or more of the mammalian genome consists of chromosome regions containing clusters of functionally related genes [17] 60 genetically diverse inbred strains. [17] forming networks with scale-free architecture. Combining LD data with pathway and genome

annotation databases, we have been able to identify the biological functions underlying several domains and networks. [17] As typified by the a and b globin gene clusters, tandem duplications can give rise to gene families whose members develop divergent, but still related, functions over time. [17]

1.1.6 Epistasis examples: Human

Despite considerable efforts, few well-replicated instances of epistasis in common human disease and trait genetics have been discovered thus far. [25] The only examples to date involve interactions featuring at least one locus with a large marginal eect, such as HLA. [25] GWAS, in ankylosing spondylitis21 and psoriasis,22 discovered interactions between two dierent HLA alleles and ERAP1. (In ankylosing spondylitis, the HLA-B27 allele has an odds ratio of 40.8, and in psoriasis the HLA-C allele has an odds ratio of 4.66.) HLA also plays a role in an interaction eect described in a GWAS of Type 1 diabetes. (In Type 1 diabetes, HLA has a main eect of 5.5, but acts non-additively with the risk of all other alleles considered cumulatively.23). Finally, interaction between RET and EDNRB in Hirschsprung's disease was discovered in a genome-wide linkage study,24 in which RET was strongly associated with disease (log-odds score of 5.6). [25] D-allele of the angiotensin I converting enzyme (ACE) gene and the C-allele of the angiotensin II type 1 receptor (AGTR1) gene3. The risk of myocardial infarction is significantly increased by the ACE D-allele in patients who carry that particular AGTR1 allele. [2] There are numerous cases of epistasis appearing as a statistical feature of association studies of human disease. A few recent examples include coronary artery disease63, diabetes64, bipolar effective disorder65 and autism66. Unfortunately, in only a few cases has the functional basis of these potential interactions been revealed. [18] One of these cases involves the genetic interactions underlying the autoimmune disease multiple sclerosis. Here, Gregersen

et al.67 found evidence that natural selection might be maintaining linkage disequilibrium between the histocompatibility loci HLA-DRB5*0101 (DR2a) and HLA-DRB1*1501 (DR2b) (FIG. 3), which are known to be associated with multiple sclerosis; linkage disequilibrium can be generated by strong epistasis among adjacent loci [18] Indeed, it has been argued that epistatic interactions are a nearly universal component of the architecture of most common traits. Templeton (2000), for instance, has listed a number of phenotypes in which epistasis plays a large role. [8] In humans, variation in triglyceride levels can be explained, in part, by two sets of interactions: between ApoB and ApoE in females and between the ApoAI/CIII/AIV complex and low-density lipoprotein receptor in males (Nelson et al. 2001) [8] Even the seemingly "simple" Mendelian trait of sickle-cell anemia is revealed to be greatly modified by epistatic interactions. Individuals with sickle-cell anemia who are homozygous for two polymorphisms near the Gg locus (leading to the persistence of fetal hemoglobin) have only mild clinical symptoms [8] For example, in humans the E4 allele of apolipoprotein epsilon (ApoE) is associated with elevated blood serum cholesterol levels, but only in individuals with the A2A2 genotype at the low density lipoprotein receptor (LDLR) locus.(3) In other words, the contribution of the ApoE allele to cholesterol levels depends on the genotype at the LDLR locus. [20]

1.1.7 Epistasis and networks

Epistasis-nonlinear genetic interactions between polymorphic loci-is the genetic basis of canalization and speciation, and epistatic interactions can be used to infer genetic networks affecting quantitative traits. [11] DATASET: Here, we compared the genetic architecture of three Drosophila life history traits in the sequenced inbred lines of the Drosophila melanogaster Genetic Reference Panel (DGRP) and a large outbred, advanced intercross population

derived from 40 DGRP lines (Flyland)[11] Surprisingly, none of the SNPs associated with the traits in Flyland replicated in the DGRP and vice versa. However, the majority of these SNPs participated in at least one epistatic interaction in the DGRP.[11] Our analysis underscores the importance of epistasis as a principal factor that determines variation for quantitative traits and provides a means to uncover genetic networks affecting these traits. [11]

1.1.8 Epistasis and evolution

epistasis can have an important influence on a number of evolutionary phenomena, including the genetic divergence between species79, ... the evolution of the structure of genetic systems8 [18] Thus far, these studies81-85 have shown that epistasis can have a strong role in limiting the possible paths that evolution can take, but not in limiting its eventual outcome. [18] linkage can facilitate the maintenance of epistatic interactions (and vice versa)86 and could help to explain how molecular complexity evolves [18] recent analysis of patterns of gene regulation suggest that there can be complex patterns of gene regulation in localized genomic regions8 [18] Gene clusters may arise as a means of promoting their coregulation through regional controls of chromatin structure and expression, and there is now considerable evidence, well summarized by Hurst et al. [1], that for variety of eukaryotes, including yeast, Caenorhabditis, Drosophila, higher plants, and mammals, genes sharing expression patterns are more likely to be in proximity than would be expected by chance. [17] ...And finally, Fisher [2] and later Nei [3,4] have argued on theoretical grounds that when genes interact epistatically, evolutionary selection will promote their genetic linkage as a means of enhancing the coinheritance of favorable allelic combinations. [17]

1.1.9 Missing heritability

IN 2002: Thus, for fixed K, p , and p , maximizing the broad AB heritability (h 2 p V /V) under the constraint repreIT sented by formula (2) is equivalent to the maximizing of VI. [8]. TABLE 2 and 3: Maxima of heritability using epistasis. [8]. Three-locus models can also give rise to higher relative risks than are possible in corresponding two-locus models. Three-locus penetrance models maximizing heritability at the low end of disease prevalence [8]

missing heritability: overestimation of the denominator happens when epistasis is ignored (phantom) [25] phantom heritability could be 62.8% in Cohn's disease, thus accounting for 80% of the current missing heritability [25] Until recently "The prevailing view among human geneticists appears to be that interactions play at most a minor part in explaining missing heritability." [25] But "[they] show that simple and plausible models can give rise to substantial phantom heritability." [25] ...although the pervasiveness of epistasis in experimental organisms suggests that the true heritability h2 of traits may be much lower than current estimates [25]

Researchers of many complex diseases (including non-insulin-dependent diabetes mellitus, prostate cancer, and schizophrenia) face the conundrum of moderately heritable diseases for which locus-by-locus analyses have not accounted for the predicted genetic variance. The models discussed in the present article provide one possible explanation for this. [8] These considerations lead us to believe that, in situations in which heritability is moderate to high but in which locus-by-locus analyses do not account for the predicted genetic variance, it is worth pursuing a hypothesis of interacting loci [near the linkage peaks] [8]

1.1.10 Detecting Epistasis / interactions

Whereas most existing epistasis screens explicitly test for a trait, it is also possible to implicitly test for fitness traits by searching for the overor under-representation of allele pairs in a given population. [1] Such analysis of imbalanced allele pair frequencies of distant loci has not been exploited yet on a genome-wide scale, mostly due to statistical difficulties such as the multiple testing problem. We propose a new approach called Imbalanced Allele Pair frequencies (ImAP) for inferring epistatic interactions that is exclusively based on DNA sequence information. [1] Most gene interaction studies explicitly measure a phenotype such as growth rate or viability [[1] However, one can also study implicit phenotypes by searching for the overor under-representation of certain allele pairs in a given population. [1] Such allele pairs are examples of Dobzhansky-Mu ller incompatibilities: they establish a fitness bias in favor of individuals inheriting the over-represented allele combination [15]. In their most extreme form such incompatibilities are embryonic lethal. [1] In this context, an implicit phenotype is a trait that is not explicitly measured in the sample but whose regulators can still be inferred from the genotype data. [1] Here, we propose to address this problem by exploiting the additional information gained from studying family trios. We show that by analyzing a sufficiently large number of individuals with known family structure it becomes possible to detect substantially more interactions than what is expected if all markers were independent. [1] Our method, called "Imbalanced Allele Pair frequencies (ImAP)" is based on inspecting 3—3 contingency tables that track the frequencies of all possible two-locus allele combinations in heterozygous individuals (assuming a diploid genome). The test that we propose is similar to a x2 test in that it compares the observed frequencies in this table to expected frequencies assuming independence. However, our version corrects

the expected frequencies for confounding factors such as family structure or allelic drift [21]. [1] In a population of 2,002 heterozygous mice with known family structure genotyped at 10,168 markers we identify 168 LD block pairs with imbalanced alleles [1]

LD: non-physical linkages between different mutations (or single nucleotide polymorphisms, SNPs) [22] These interactions can be physical protein interactions, regulatory interactions, functional compensation/antagonization or many other forms of interactions. [22] non-physical SNP linkages, coupled with knowledge of SNP-disease associations may shed more light on the role of gene interactions in human disorders. [22] exonic regions of protein-coding genes from the HapMap database to construct a database named the Linkage-Disequilibrium-based Gene Interaction database (LDGIdb). The database stores 646,203 potential human gene interactions, which are potential interactions inferred from SNP pairs that are subject to long-range strong linkage disequilibrium (LD), or non-physical linkages. To minimize the possibility of hitchhiking, SNP pairs inferred to be non-physically linked were required to be located in different chromosomes or in different LD blocks [22] Here we consider only the subpopulations that contain at least 20 individuals. [22] strong LD (r2 0.8); [22]

INBREED MICE: The process of inbreeding to homozygosity imposes intense selective pressures; all efforts among some species have failed, and with mice, only a fraction of initial attempts succeeded. [17] Accordingly, we can expect that if clustering of functionally related genes is a common feature of mammalian genomes, there is likely to be selection for coadapted allelic combinations among the genes encoding functions that influence fitness and survival during inbreeding. This would result in regions of linkage disequilibrium (LD) among inbred strain genomes; i.e., some allelic combinations should

occur more often than expected by chance. [17] Data: 1,456 SNPs, chosen for their high information content, among a set of 60 common and wildderived inbred mouse strains chosen for their genetic diversity. [17] The identity of these strains and the phylogenetic relationships among them are indicated in Figure 1, which was constructed using neighbor-joining [17] LD calculation: estimated LD using D9, the difference between the observed frequency of an allelic combination and its random expectation, relative to the maximum deviation possible given the allele frequencies of the two markers [14,15]. D9 corrects for differences in allele frequencies and describes LD equally well when there is selection for or against the combination of majority alleles. A cumulative Fisher's exact test (FET) was used to compute the probability (pFET) of obtaining an equally or more extreme distribution under the null hypothesis of random allelic association between pairs of SNPs. [17] Permutaation test: In one set, marker locations were randomized while maintaining the assignments of alleles to strains (Figure 2, red triangles), and in the other set the assignments of alleles to strains were randomized while preserving allele ratios and marker locations (Figure 2, solid circles) [17] It is difficult to escape the conclusion that the selective factors acting to generate LD domains and networks during inbreeding reflect clustering and/or interaction of functionally related elements along chromosomes [17]

Long-range linkage disequilibria (LRLD) between sites that are widely separated on chromosomes may suggest that population admixture, epistatic selection, or other evolutionary forces are at work. [12] We quantified patterns of LRLD on a chromosome-wide level in the YRI population of the HapMap dataset of single nucleotide polymorphisms (SNPs). [12] We calculated the disequilibrium between all pairs of SNPs on each chromosome (a total of .261011

values) and evaluated significance of overall disequilibrium using randomization. [12] The results show an excess of associations between pairs of distant sites (separated by .0.25 cM) on all of the 22 autosomes. [12] Disequilibria between closely-linked sites result largely from random genetic drift or (equivalently) the common ancestry of unrecombined chromosome blocks. [12] While these "long range haplotypes" can extend over a few hundred kb in unrelated humans [5], they still span only a very small fraction of an entire chromosome. [12] Considerably less attention has been paid to patterns of LD between pairs of sites that are separated by much greater genetic distances (say, 1 cM or more). [12] finding substantial long range linkage disequilibrium (LRLD) suggests that countervailing forces are at work. [12] 1) One possibility is population admixture [6], which has been proposed to explain unusual patterns of LRLD in some human populations [12] 2) A second contributing force is drift. Even in a population at demographic equilibrium, recombination between distant chromosome blocks will largely but not completely erase LD caused by drift. Recurrent bottlenecks are particularly effective at generating LD [9], and may have contributed importantly to disequilibria in nonAfrican populations of humans [12] 3) Third, epistatic selection can maintain linkage disequilibrium indefinitely [11]. Epistasis has been implicated in the LD observed between two pairs of genes in humans [12,13]. [12] 4) Fourth, the hitchhiking of linked sites with a positively-selected mutation can generate large haplotype blocks that result in disequilibria over the region that they span [3,4]. [12] 5) Fifth, structural variation in chromosomes, such as inversions, can alter patterns of recombination and consequently cause LD to extend over unusually large regions of a chromosome [14-16]. [12] to our knowledge there has been only one previous survey of associations between chromosomal regions across the entire human genome using high-density data. Sved [17] studied correlations

in heterozygosity between chromosome blocks. His analysis of the HapMap phase 3 data found evidence of associations between blocks at distances of up to 10 cM and weak correlations between blocks on different chromosomes, but he did not attempt to assess their statistical significance. Lawrence et al. [18] provided a web-based tool for exploring long distance linkage disequilibria in the HapMap data, but did not go on to study patterns in the data. [12] This paper investigates patterns of LRLD in the YRI population (the Yoruba in Ibadan, Nigeria) from the HapMap Phase 2 dataset of single nucleotide polymorphisms [23]. YRI also has weaker short-range disequilibria that might otherwise obscure the patterns of LRLD [12] We calculated the disequilibria between all pairs of SNPs on the same chromosome, then analyze these data with new statistical methods. [12] Using null distributions generated by randomization, we find significant excess of disequilibria on all 22 autosomes in the Yoruba population. [12] Data: 120 YRI haplotypes that were genotyped at over 2.86106 SNPs in HapMap Phase 2 (data build 22) [12] LD issues: Most commonly used measures of linkage disequilibria are not well suited for that purpose [8]. For example, a large value of D9 is likely to result from sampling if allele frequencies are near 0 or 1, while even a small value is unlikely to appear by chance if allele frequencies are intermediate and the sample size is large. [12] We therefore use the probability that a value of the disequilibrium D as large or larger than that in the sample would be observed if there is no association in the population from which the sample is drawn, conditioned on the sampled allele frequencies at the two loci. This probability, which we denote pD, is given by the tail of Fisher's exact test [8,28,29] [12] As the distance between a pair of sites on a chromosome grows large (specifically, the product of the recombination rate and the effective population size becomes much greater than 1), the sampling distribution for two-locus haplotypes converges

on that of Fisher's exact test [30,31]. [12] Patches: When a pair of distant sites are in disequilibrium, it is likely that other sites near to them will also be associated as a result of shortrange associations [17,32]. In effect, the underlying structure in the data is disequilibrium between pairs of chromosomal blocks rather than between pairs of individuals sites [12] To control for this we used a simple and efficient ad hoc strategy that identifies "patches" of disequilibria. [12] Results: We take two approaches to search for nonrandom patterns of LRLD. We first ask whether observed values of pD are more extreme than expected. For this purpose we determined the most extreme (that is, smallest) value of pD in each patch, then calculated the mean of these extreme values across all patches on a chromosome. We refer to this statistic as pDmax. [12] - Second, we ask whether the number of LRLD patches observed for a given chromosome is greater than expected by chance. We denote this statistic as nP. [12] To test for the statistical significance of pDmax and nP, we generate their null distributions using a randomization method [12] There are two motivations behind this method. First, it preserves the allele frequencies at each site. Second, it maintains the structure of short range disequilibria in the sample. [12] Computational time: Constructing these null distributions is the most computationally intensive part of our method. For the analyses reported below, over 4.861014 values of pD were computed, and the project consumed about 34,000 hours of CPU time. [12] All of the 22 chromosomes show significant values for pDmax at the p,0.05 level, and all remain significant after a Bonferroni correction for multiple tests. For the second test statistic, n , 19 chromosomes show significant P values, 18 of which remain significant after the Bonferroni correction. These results suggest there is long-range linkage disequilibrium in the YRI population. [12] [LRLD] have been little studied, they may be indicators of important evolutionary processes [12]

1.1.11 Epistasis & GWAS

IN 2002 OPINION: for the abandonment of linkage studies in favor of genome scans for association. However, there exists a large class of genetic models for which this approach will fail: purely epistatic models with no additive or dominance variation at any of the susceptibility loci. [8]. Is it reasonable to suppose that an approach that must succeed in identifying fully penetrant Mendelian genes will also succeed for complex diseases? [8]. The complex relationship between genotype and phenotype, however, may ultimately prove to be inadequately described by simply summing the modest effects from several contributing loci [8] The main reason that most studies of complex human phenotypes fail to find evidence for epistatic interactions may simply be that commonly used designs and analytic methods inherently minimize or exclude the possibility of epistasis (Frankel and Schork 1996) [8] The complex relationship between genotype and phenotype, however, may ultimately prove to be inadequately described by simply summing the modest effects from several contributing loci. [8] We note that the number of tests necessary to evaluate all two-, three-, and four-way interactions, for 30-60 candidate loci, has a range similar to the number of tests suggested for a single genomewide association scan using SNPs (Collins et al. 1999; Kruglyak 1999) [8] Thus, although searching for two-, three-, four-, or n-way interactions among all the markers in a genome scan would not be practicable, a candidate-locus approach based on a genome scan for linkage may be. [8]

Several approaches have been developed to detect epistasis, including the combinatorial partitioning method (CPM)7, the restricted partitioning method (RPM)8, multifactor-dimensionality reduction (MDR)2, multivariate adaptive regression spline (MARS)9, the logistic regression method10 and backward genotype-trait association (BGTA)11. Although these methods all

showed promise, they have been tested only on small data sets. [23] methods based on brute-force searches such as CPM and MDR are impractical for large data sets [23] STEPWISE LOGISTIC REGRESSION: The stepwise logistic regression approach of ref. 12 works as follows: (i) all markers are individually tested and ranked for marginal associations with the disease; (ii) the top 10% of markers are selected, among which all k-way (k 14 2 or 3) interactions are tested and ranked for associations. The authors of ref. 12 also proposed an exhaustive logistic regression testing approach, which we choose not to consider in this study because of its prohibitive computational cost. Note that even their stepwise approach can become computationally intractable for high-order interactions. [23] Recently, a simulation study12 explored the use of a stepwise logistic regression approach to identify two-way and three-way interactions. The authors demonstrated that searching for interactions in genome-wide association mapping can be more fruitful than traditional approaches that exclusively focus on marginal effects. [23]

The extent to which epistasis is involved in regulating complex traits is not known, and so we cannot assume that epistasis will be found for every trait in every population. [2] However, we argue that epistasis has been overlooked for too long and that it now needs to be routinely explored in complex trait studies. [2] For complex traits such as diabetes, asthma, hypertension and multiple sclerosis, the search for susceptibility loci has, to date, been less successful than for simple Mendelian disorders. This is probably due to complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects [6] The presence of epistasis is a particular cause for concern, since, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will

be hindered by their interaction. [6] Although genetic interactions are hard to detect in humans (see below), several cases involving variants with large marginal effects have been recently reported in Hirschsprung's disease, ankylosing spondylitis, psoriasis, and type I diabetes [25] ...geneticists have tested for pairwise epistasis between loci, but have found few significant signals. [25] ...The reason is that 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. If n loci had equivalent effects, the sample size to detect the n loci would thus scale with n^2, whereas the sample size to detect their n^2 interactions scales with n^4. [25] Suppose that we consider two variants with frequency 20% that contribute to different pathways and increase risk by 1.3-fold (which is a large effect relative to those typically seen in GWAS). The sample size required to detect the variants is 4,900 (with 50% power and genome-wide significance level of $\alpha = 5 \times 10^{-8}$ in a genome-wide association study with an equal number of cases and controls), whereas the sample size required to detect their pairwise interaction is roughly 450,000 (at 50% power and an appropriate significance level to account for multiple hypothesis testing). A researcher who studied 100,000 samples would likely discover all of the loci but would find little evidence of epistatic interactions. [25] In short, the failure to detect epistasis does not rule out the presence of genetic interactions sufficient to cause substantial phantom heritability [25]

Cases only. The most straightforward multilocus analysis of cases-only data is a χ^2 test of independent segregation for the loci. [8] Case-control. A second approach is a multilocus case-control analysis. One method for doing this would be to compare the distribution of cases among the 3L genotypes, where L is the number of biallelic loci being simultaneously examined, versus the distribution of controls. In this analysis, a sample of N cases and N

unrelated controls drawn from a population modeled by table 3 will, again, yield an expected χ^2 statistic 2N. However, the degrees of freedom under the null hypothesis are now 8. [8]

We developed a general theory for studying linkage disequilibrium (LD) patterns in disease population under two-locus disease models. [24] Our results showed that the P values of the LD-based statistic were smaller than those obtained by other approaches, including logistic regression models. [24] This was further developed by Cockerham4 and Kempthorne5 into the modern representation that treats statistical gene interactions as interaction terms in a regression model or a generalized linear model on allelic effects.2,6-11 [24] we propose to define interaction between two unlinked loci (or genes) for a qualitative trait as the deviance of the penetrance for a haplotype at two loci from the product of the marginal penetrance of the individual alleles that span the haplotype. [24] DEFINE: Deviance [24] Interaction between two unlinked loci will result in deviation of the penetrance of the two-locus haplotype from independence of the marginal penetrance of the alleles at an individual locus, which in turn will create linkage disequilibrium (LD) even if two loci are unlinked. [24] Therefore, it is possible to develop statistics for detection of interaction between two unlinked loci by use of deviations from LD [24] we assume that two disease-susceptibility loci are in Hardy-Weinberg equilibrium (HWE) and are unlinked. [24] [they show that] Under this definition, in the absence of interaction, two unlinked loci in the disease population will be in linkage equilibrium [24] Similar to linkage equilibrium, where the frequency of a haplotype is equal to the product of the frequencies of the component alleles of the haplotype, absence of interaction between two unlinked loci implies that 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] TEST STATISTIC: [24] Intuitively, we can test interaction by comparing the difference in the LD levels between two unlinked loci between cases and controls [24] We can show that test statistic TI is asymptotically distributed as a central x2 distribution under the (1) null hypothesis of no interaction between two unlinked loci [24] we compared the power of the LD-based statistic with that of the logistic model. [24] Power comparison with logistic regression analysis demonstrated that this LD-based test statistic has much higher power in detecting interaction than does the logistic regression method. [24] To further evaluate its performance for detection of interaction between two loci, the proposed LD-based statistic was applied to two published data sets. Our results showed that, in general, P values of the test statistic TI were much smaller than those of other approaches, including logistic regression analysis. [24]

Although some existing computational methods for identifying genetic interactions have been effective for small-scale studies, we here propose a method, denoted 'bayesian epistasis association mapping' (BEAM), for genome-wide case-control studies [23] BEAM treats the disease-associated markers and their interactions via a bayesian partitioning model and computes, via Markov chain Monte Carlo, the posterior probability that each marker set is associated with the disease. [23] In the past century, scientists have made great progresses in mapping genes responsible for mendelian diseases. However, genetic variants underlying most common (or 'complex') diseases are non-mendelian. [23] These variants are typically not rare in the population (42It has been speculated that epistasis ubiquitously contributes to complex traits partly because of the sophisticated regulatory mechanisms encoded in the human genome1. [23] EPI EXAMPLES: An increasing number of reports have indicated the

presence of multilocus interactions in many human complex traits, such as breast cancer2, post-PTCA stenosis3, essential hypertension4, atrial fibrillation5 and type 2 diabetes6. [23] GWAS EPISTASIS [Discussion]: We also applied BEAM to an association study of age-related macular degeneration (AMD)13, which included B100,000 SNP markers. Although BEAM did not find significant interactions in the AMD data set, it was able to discover two-way or three-way interactions among the B100,000 SNPs simulated based on the AMD data. [23] BEAM METHOD: The BEAM algorithm takes case-control genotype marker data as input and produces, via MCMC simulations, posterior probabilities that each marker is associated with the disease and involved with other markers in epistasis. [23] The input genotyped markers should be in their natural genomic order when there is linkage disequilibrium (LD) among some of them. The method can be used either in a 'pure' bayesian sense or just as a tool to discover potential 'hits'. For the former, one relies on the reported posterior probabilities to make inferential statements; as for the latter, one can take the reported hits and use another procedure to test whether these hits are statistically significant. [23] The latter approach is more robust to model selection and prior assumptions (such as Dirichlet priors with arbitrary parameters) and is less prone to the slow mixing problem in the MCMC computational procedure. We also propose the B statistic to facilitate the latter approach and show that it is more powerful than the standard w2 statistic for epistasis detections. [23] For the non-epistasis model (model 1), all three epistasis mapping methods performed similarly to the single-marker w2 test (Fig. 1), indicating that the power for detecting marginal associations was not compromised by using the more complex models. [23] Notably, results for model 4 suggest that stepwise methods can miss markers with small or no marginal effects, whereas BEAM can get these markers back through

iterations. [23] POWER ISSUES RELATED TO AF: The power of association mapping can be greatly hampered by the discrepancy of allele frequencies between unobserved disease loci and associated genotyped markers15 [23] For data sets with large MAF discrepancies and moderate LD, the power of all methods suffered. [23] At the extreme case when the MAF discrepancy was maximized (that is, MAF 14 0.5), all methods had little power in detecting interaction associations [23] The impact of LD on power seemed to be less profound than the effect of MAF discrepancy. [23] ANALISYS: DATA: The data set contains 116,204 SNPs genotyped for 96 affected individuals and 50 controls. [23] RESULTS: BEAM found no significant interactions associated with AMD from this data set. It is possible that the small sample size of 146 individuals is insufficient for detecting subtle epistasis interactions. [23]

The purpose of this Review is to provide a survey of the methods and related software packages that are currently being used to detect the interactions between the genetic loci that contribute to human genetic disease. [7] Interaction as departure from a linear model. The most common statistical definition of interaction relies on the concept of a linear model that describes the relationship between an outcome variable and a predictor variable or variables [7] Arguably the most well-known form of this type of analysis is simple linear or least squares regression26, in which we relate an observed quantitative outcome y (for example, weight) to a predictor variable x (for example, height) using a 'best fit' line or regression [7] From a statistical point of view, interaction represents departure from a linear model that describes how two or more predictors predict a phenotypic outcome [7] For a disease outcome and case-control data, rather than modelling a quantitative trait y, the usual approach is to model the expected log odds of disease as a linear function

of the relevant predictor variables [7] DEFINITION Penetrance: The probability of displaying a particular phenotype (for example, succumbing to a disease) given that one has a specific genotype. [7] DEFINITION: Marginal effects: The average effects (for example, penetrances) of a single variable, averaged over the possible values taken by other variables. These could be calculated for one locus of a two-locus system as the average of the two-locus penetrances, averaged over the three possible genotypes at the other locus. [7] For or simplicity, I have concentrated here on defining interaction in relation to two genetic factors (two-locus interactions). In practice, however, for complex diseases we might also expect three-locus, four-locus and even higher-level interactions. Mathematically, such higherlevel interactions are simple extensions to the two-locus models described earlier. [7] CASE ONLY METHODS: A case-only test of interaction can therefore be performed by testing the null hypothesis that there is no correlation between alleles or genotypes at the two loci in a sample that is restricted to cases alone. This test can easily be performed using a simple 2 test of independence between genotypes (a four degrees of freedom test) or alleles (a one degree of freedom test), or using logistic or multinomial regression in any statistical analysis package. [7] The main problem with the case-only test is its requirement that the genotype variables are not correlated in the general population. It is this assumption, rather than the design per se, that provides the increased power compared with case-control analysis [7] The caseonly test is therefore unsuitable for loci that are either closely linked or show correlation for another reason (for example, if certain genotype combinations are related to viability). [7] Tests for association allowing for interaction: From a mathematical point of view, a test for association at a given locus C while allowing for interaction with another locus B (a joint test16) corresponds to comparing the fit to the observed data

of a linear model in which the main effects of B, C and their interactions are included [7] Theoretically, if no interaction effects exist, these joint tests will be less powerful than marginal singlelocus association tests. However, if interaction effects exist, then the power of joint tests can be higher than that of single-locus approaches52. [7] CLASSIFICATION TREE: Recursive partitioning approaches are based on classification and regression trees111. Trees are constructed (see the figure) using rules that determine how wella split at a node (based on the values of a predictor variable such as a SNP) can differentiate observations with respect to the outcome variable (such as case-control status). A popular splitting rule is to use the variable that maximizes the reduction in a quantity known as the Gini impurity111,112 at each node. [7] RAMDOM FOREST: A random forest is constructed by drawing with replacement several bootstrap samples of the same size (for example, the same number of cases and controls) from the original sample. An unpruned classification tree is grown for each bootstrap sample, but with the restriction that at each node, rather than considering all possible predictor variables, only a random subset of the possible predictor variables is considered. This procedure results in a 'forest' of trees, each of which will have been trained on a particular bootstrap sample of observations. [7] BAYESIAN MODEL SELECTION: Bayesian model selection techniques92 offer an alternative approach for selecting predictor variables and the interactions between them that are the best predictors of phenotype. The key difference between Bayesian model selection and simple comparisons of nested regression models using frequentist (non-Bayesian) procedures is the specification of prior distributions for the unknown regression parameters as well as for a dimension parameter in a Bayesian approach. This dimension parameter specifies how many non-zero predictors are included [7] A posterior distribution for these parameters, given

the observed data, can then be calculated using Markov chain Monte Carlo (MCMC)93 simulation techniques, in which one traverses the space of the possible models (sets of parameter values), sampling the outputs of the simulation run at intervals. Although MCMC is a flexible approach, it can require some care with respect to the choice of prior distributions, proposal schemes (determining how one moves between models) and the number of iterations required to achieve convergence. [7] BEAM: Bayesian Epistasis Association Mapping. A recently proposed MCMC approach that is specifically designed to detect interacting, as well as non-interacting, loci is Bayesian epistasis Association Mapping13, which is implemented in the software package BeAM. In BeAM, predictors in the form of genetic marker loci are divided into three groups: group 0 contains markers that are not associated with disease, group 1 contains markers that contribute to disease risk only by main effects and group 2 contains markers that interact to cause disease by a saturated model. Given prior distributions that describe the membership of each marker in each of the three groups and prior distributions for the values of the relevant regression coefficients given group membership, a posterior distribution for all relevant parameters can be generated using MCMC simulation. In addition to making inferences in a fully Bayesian inferential framework, one can use the results from BeAM in a frequentist hypothesistesting framework by calculating a 'B-statistic'13 that tests each marker or set of markers for significant association with a disease phenotype. [7] EBAM LIMITATIONS: BeAM cannot currently handle the 500,000-1,000,000 markers that are now routinely being genotyped in genome scans of 5,000 or more individuals. [7]

We extend the basic AdaBoost algorithm by incorporating an intuitive importance score based on Gini impurity to select candidate SNPs. [14] Permutation tests are used to control the statistical significance. [14] We have

performed extensive simulation studies using three interaction models to evaluate the efficacy of our approach at realistic GWAS sizes, and have compared it with existing epistatic detection algorithms. [14] CURRENT METHODS: Generally speaking, existing approaches for searching gene- gene or SNP-SNP interactions can be grouped into four broad categories. [14] 1) Methods in the first category rely on exhaustive search. Classical statistics such as the Pearson's 2 test or the logistic regression that are commonly used as single-locus tests for GWAS can potentially be used in searching for pairwise interactions. Marchini et al. (2005) have shown that explicitly modeling of interactions between loci for GWAS with hundreds of thousands of markers is computationally feasible. They also showed that these 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. [14] 2) The second category consists of methods relying on stochastic search, with BEAM (Zhang and Liu, 2007) as one representative of such algorithms. Later algorithms in this category [e.g. epiMODE (Tang et al., 2009)] largely adopted and extended BEAM. BEAM uses Markov chain Monte Carlo (MCMC) sampling to infer whether each locus is a disease locus, a jointly affecting disease locus, or a background (uncorrelated) locus. The algorithm begins by assigning each locus to each group according to a prior distribution. Using the Metropolis-Hastings algorithm, it attempts to reassign the group labels to each locus. At the end, it uses a special statistic, called the B-Statistic, to infer statistical significance from the hits sampled in MCMC. This approach avoids computing all interactions, but can still theoretically find high-order interactions. The number of MCMC rounds is the primary parameter that mediates runtime, as well as power. The suggested number of

MCMC rounds is in the quadratic of the number of SNPs, which limits applicability of BEAM on large datasets. [14] 3) Methods in the third category are machine learning approaches such as tree-based methods or support vector machines (SVM). For example, a popular ensemble approach, Random Forests [14] 4) Methods in the forth category rely on conditional search. In such a case, analyses are performed in stages (Evans et al., 2006; Li, 2008). A small subset of promising loci is identified in the first stage, normally using single locus methods, and multi-locus methods are used in the later stage(s) to model interactions based on the selection in the first stage. Stepwise regression has been widely used in this case and several different strategies have been studied in the literature. 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. One should also notice that the conditional search category is more like a strategy rather than an approach. In addition to single-locusbased methods, any approaches discussed previously, especially the machine learning ones, can be used to search for candidates in the first stage. [14] THIS METHOD: We extend the basic AdaBoost algorithm by incorporating an intuitive importance score based on Gini impurity to select candidate SNP [14] Instead of trying to create a monolithic learner or model, ensemble systems attempt to create many heterogeneous versions of simpler learners, called weak learners. The opinions of these heterogeneous experts are then combined to formulate a complete picture of the data. [14] Usually, a SNP is selected to ensure largest homogeneity in the child nodes. In our implementation, we use the gain on Gini Impurity. Intuitively, when child nodes have lower impurity from a split based on an attribute (i.e. a SNP here), each child node will have purer classification. Therefore, the genotype frequencies from the two classes (case and control) are expected to be more

different. [14] Usually decision trees are built with binary splits, where individuals with one value of the feature are placed into one group, and the remainder into the other. Since genotype data is three valued, we extend this to do a ternary split. [14] Despite only using marginal effects to select SNPs, decision trees can still detect some interaction. Because of the recursive partitioning, lower nodes are effectively conditioned on the value of their parents. The core idea of AdaBoost is to draw bootstrap samples to increase the power of a weak learner. This is done by weighting the individuals when drawing the bootstrap sample. When a weak learner instance misclassifies an individual, the weight of that individual is increased (and increased more if the weak learner instance was otherwise accurate). Thus, hard to classify individuals are more likely to be included in future bootstrap samples. In the end, the ensemble votes for class labels weighting the weak learner instances by training set accuracy. [14]

1.1.12 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]

1.1.13 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.