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, 13, 14], 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 [12] 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. [12] The commonly used definition of epistasis–an allele at one locus masks the expression of an allele at another locus–reflects this original definition. [12]

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. [11] 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 [11] The use of the term epistasis has since expanded to describe nearly any set of complex interactions among genetic loci [11] 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 [11] 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 [11] '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 [11] '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. [11] '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. [11]

It should be apparent that the global analysis of geneinteraction patterns bears a striking resemblance to what is now called systems biology [11]

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. [10] 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 [10] Here, we show that the choice among alternative definitions can have profound consequences. [10]

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 [10] 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). [10] 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. [10] 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. [10] 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) [10] Genetic interaction studies have also differed in their choice of neutrality functions, generally using either a multiplicative or a minimum mathematical function. [10] 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). [10] 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. [10] 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. [10] 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. [10] 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. [10] 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 [10]

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. [10] RESULTS: The Choice of Genetic Interaction Definition Matters: [10] 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. [10] Product and Log Definitions Are Equivalent for Deleterious Mutations: [10] The Product Definition Reveals Functional Relationships Missed by the Min Definition. [10] Genetic Interaction Networks from Min and Product Definitions Differ Greatly. [10]

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. [10]

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. [11]

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. [11] 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. [13] 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. [13] Studies have begun to reveal that epistasis is pervasive. [13] We assert that epistasis and pleiotropy are not isolated occurrences, but ubiquitous and inherent properties of biomolecular networks. [12]

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) [11] 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. [13] 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 eightfold) the total effect of the donor genome, indicating strong epistasis. [13] 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). [10] 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. [12]

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. [13] The only examples to date involve interactions featuring at least one locus with a large marginal effect, such as HLA. [13] GWAS, in ankylosing spondylitis21 and psoriasis,22 discovered interactions between two different 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 effect 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). [13] 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. [11] 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 [11] 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. [12]

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. [9] 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)[9] 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.[9] 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. [9]

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 [11] 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. [11] linkage can facilitate the maintenance of epistatic interactions (and vice versa)86 and could help to explain how molecular complexity evolves [11] recent analysis of patterns of gene regulation suggest that there can be complex patterns of gene regulation in localized genomic regions8 [11]

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) [13] phantom heritability could be 62.8% in Cohn's disease, thus accounting for 80% of the current missing heritability [13] Until recently "The prevailing view among human geneticists appears to be that interactions play at most a minor part in explaining missing heritability." [13] But "[they] show that simple and plausible models can give rise to substantial phantom heritability." [13] ...although the pervasiveness of epistasis in experimental organisms suggests that the true heritability h2 of traits may be much lower than current estimates [13]

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]

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]

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 [13] ...geneticists have tested for pairwise epistasis between loci, but have found few significant signals. [13] ...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. [13] 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. [13] In short, the failure to detect epistasis does not rule out the presence of genetic interactions sufficient to cause substantial phantom heritability [13]

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]

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]

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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.