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

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

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

CHAPTER 1
Introduction

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

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

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

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

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

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

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

1.1 Epistasis

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

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

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

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

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

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

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

identical results under some conditions, an alternative definition choice can lead to different consequences[15].

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

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

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

It has been shown choice of definition can dramatically alter the resulting set of interactions [15]. To evaluate this the authors in [15] applied all four definitions to two studies providing quantitative growth-rate measurements of cell populations. They show that: i) additive and Log definitions have different

biases; ii) Product and Log definitions are equivalent for deleterious mutations; iii) the product definition can reveal functional relationships missed by the Min definition; and iv) interaction networks from Min and Product definitions differ greatly. This leads to the question on which definition to use. By examining the deviation distribution of expected (double-mutant) phenotype from the observed phenotype they found that Product and Log definitions not only are the closest to the ideal, but also are practically equivalent when single mutants are deleterious [15].

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

1.1.1 Epistasis is ubiquitous

Epistasis is defined as departure from additive effects. Nevertheless, there is no reason to think that traits should be additive based on a purely biological perspective [25] since biology is riddled with non-linearity such as genetic networks exhibit binary states, ligand - receptors concentration having sigmoid-like response, concentration saturations of substrate - enzymes reactions, sharp transitions created by cooperative protein binding, the pathways constrained by rate-limiting inputs, etc. [25]. It has been asserted that epistatic effects are not isolated events, but ubiquitous [20] and probably inherent properties of biomolecular networks. This leads to think that epistasis in the classical sense may be ubiquitous, a thought which has been partially confirmed from mutational studies [18]. Genetic studies of synthetic traits, which occur only

when multiple loci or pathways are all disrupted, in model organisms have identified instances of interacting genes revealing that epistasis may be pervasive [25]. Researchers found [18] that when looking for interactions induced by systematically over-expressing genes in Saccharomyces cerevisiae about 15% of studied genes induced growth defects with most over-expression not matching the phenotypes of individual deletions.

1.1.2 Epistasis examples: Non-human

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

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

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

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

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

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

1.1.3 Epistasis examples: Human

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

Perhaps the best examples are interactions involving at least one locus with a large effect such as HLA [25]. Two different interaction involving HLA alleles and ERAP have been discovered in GWAS from ankylosing spondylitis

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

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

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

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

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

1.1.4 Epistasis and evolution

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

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

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

1.1.5 Missing heritability

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

heritability (under some constraints) is equivalent maximizing the interaction variance [8].

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

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

1.1.6 Detecting Epistasis / interactions

Linkage disequilibria (LD) between close sites are the result of unrecombined chromosome blocks from common ancestry [**?**], nevertheless LD between widely separated sites suggests epistatic selection forces are at work [**?**, **?**, 12]. In an analysis using Yoruba population (from Ibadan, Nigeria) of the HapMap dataset patterns of LD were quantified and significance of overall disequilibrium per chromosome was evaluated of using randomization [12], showing an excess of associations in distant on all of the 22 autosomes. Although this is suggestive of epistasis, other hypothesis should not be ruled out: i) population admixture has been proposed to explain unusual patterns of long range LD [**?**] ii) recombination between distant chromosome blocks may not completely erase LD caused by drift even in a population at demographic equilibrium, iii)

bottlenecks are particularly effective at generating LD iv) hitchhiking of linked sites with a positively-selected mutation can generate large haplotype blocks v) large inversion and other structural variation alter recombination patterns thus causing LD over unusually large regions [**?**].

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

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

The authors in [1] studying a population of 2,002 mice in family trios. They performed a χ^2 test correcting by confounding factors (such as expected frequencies, family structure and allelic drift) based on inspecting $3x3$ contingency tables of all possible two-locus allele combinations. They claim that

using their method it is possible to detect more interactions than using independent markers and as a result they were able to identify 168 LD block pairs with imbalanced alleles [1].

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

1.1.7 Epistasis & GWAS

IN 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]

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]

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]

1.1.8 Epistasisi GWAS: Power issues

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

among candidate loci may be insufficient for detection of the contributing loci. [8] The restriction on maximum heritabilities in these models is most easily seen by examining L-locus models for which no collection of L 1 loci shows marginal deviations. [8]

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

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

1.1.9 Epistatic GWAS

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

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

The definition of epistasis from a statistical perspective is a "departure from a linear model" [7]. This means that in a logistic regression model the

input for sample s includes terms with each of the genotypes at loci i and j), as well as an "interaction term" $g_{s,i} \cdot g_{s,j}$ [6].

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

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

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

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

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

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

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

References

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

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

[3] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):0–1, 2012.

[4] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Toxicogenomics in non-mammalian species*, page 92, 2012.

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

[6] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

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

[8] Robert Culverhouse, Brian K Suarez, Jennifer Lin, and Theodore Reich. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.

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

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

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

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

[13] Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.

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

[15] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[16] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

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

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

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

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

[21] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2009.

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

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

[24] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[25] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[26] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.